REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

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

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture. The general idea that text corpora capture semantics, including cultural stereotypes and empirical associations, has long been known in corpus linguistics (*1*, *2*), but our findings add to this knowledge in three ways. First, we used word embeddings (*3*), a powerful tool to extract associations captured in text corpora; this method substantially amplifies the signal found in raw statistics. Second, our replication of documented human biases may yield tools and insights for studying prejudicial attitudes and behavior in humans. Third, since we performed our experiments on off-the-shelf machine learning components [primarily the Global Vectors for Word Representation (GloVe) word embedding], we show that cultural stereotypes propagate to artificial intelligence (AI) technologies in widespread use.

Before presenting our results, we discuss key terms and describe the tools we use. Terminology varies by discipline; these definitions are intended for clarity of the present article. In AI and machine learning, bias refers generally to prior information, a necessary prerequisite for intelligent action (*4*). Yet bias can be problematic where such information is derived from aspects of human culture known to lead to harmful behavior. Here, we will call such biases "stereotyped" and actions taken on their basis "prejudiced."

We used the Implicit Association Test (IAT) as our primary source of documented human biases (*5*). 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. We developed our first method, the Word-Embedding Association Test (WEAT), a statistical test analogous to the IAT, and applied it to a widely used semantic representation of words in AI, termed word embeddings. Word embeddings represent each word as a vector in a vector space of about 300 dimensions, based on the textual context in which the word is found. We used the distance between a pair of vectors (more precisely, their cosine similarity score, a measure of correlation) as analogous to reaction time in the IAT. The WEAT compares these vectors for the same set of words used by the IAT. We describe the WEAT in more detail below.

Most closely related to this paper is concurrent work by Bolukbasi *et al.* (*6*), who propose a method to "debias" word embeddings. Our work is complementary, as we focus instead on rigorously demonstrating human-like biases in word embeddings. Further, our methods do not require an algebraic formulation of bias, which may not be possible for all types of bias. Additionally, we studied the relationship between stereotyped associations and empirical data concerning contemporary society.

Using the measure of semantic association described above, we have been able to replicate every stereotype that we tested. We selected IATs that studied general societal attitudes, rather than those of subpopulations, and for which lists of target and attribute words (rather than images) were available. The results are summarized in Table 1.

Greenwald *et al.* introduced and validated the IAT by studying biases that they consider nearly universal in humans and about which there is no social concern (*5*). We began by replicating these inoffensive results for the same purposes. Specifically, they demonstrated that flowers are significantly more pleasant than insects, based on the reaction latencies of four pairings (flowers + pleasant, insects + unpleasant, flowers + unpleasant, and insects + pleasant). Greenwald *et al.* measured effect size in terms of Cohen's *d*, which is the difference between two means of log-transformed latencies in milliseconds, divided by the standard deviation. Conventional small, medium, and large values of *d* are 0.2, 0.5, and 0.8, respectively. With 32 participants, the IAT comparing flowers and insects resulted in an effect size of 1.35 ($P < 10^{-8}$). Applying our method, we observed the same expected association with an effect size of 1.50 ($P < 10^{-7}$). Similarly, we replicated Greenwald *et al.*'s finding (*5*) that musical instruments are significantly more pleasant than weapons (see Table 1).

Notice that the word embeddings "know" these properties of flowers, insects, musical instruments, and weapons with no direct experience of the world and no representation of semantics other than the implicit metrics of words' co-occurrence statistics with other nearby words.

We then used the same technique to demonstrate that machine learning absorbs stereotyped biases as easily as any other. Greenwald *et al.* (*5*) found extreme effects of race as indicated simply by name. A bundle of names associated with being European American was found to be significantly more easily associated with pleasant than unpleasant terms, compared with a bundle of African-American names.

In replicating this result, we were forced to slightly alter the stimuli because some of the original African-American names did not occur in the corpus with sufficient frequency to be included. We therefore also deleted the same number of European-American names, chosen at random, to balance the number of elements in the sets of two concepts. Omissions and deletions are indicated in our list of keywords (see the supplementary materials).

In another widely publicized study, Bertrand and Mullainathan (*7*) sent nearly 5000 identical résumés in response to 1300 job advertisements, varying only the names of the candidates. They found that European-American candidates were 50% more likely to be offered an opportunity to be interviewed. In follow-up work, they argued that implicit biases help account for these effects (*8*).

We provide additional evidence for this hypothesis using word embeddings. We tested the names in their study for pleasantness associations. As before, we had to delete some low-frequency names. We confirmed the association using two different sets of "pleasant/unpleasant" stimuli: those from the original IAT paper and also a shorter, revised set published later (*9*).

Turning to gender biases, we replicated a finding that female names are more associated with family than career words, compared with male names (*9*). This IAT was conducted online and thus has a vastly larger subject pool but far fewer keywords. We replicated the IAT results even with these reduced keyword sets. We also replicated an online IAT finding that female words (e.g., "woman" and "girl") are more associated than male words with the arts than with mathematics (*9*). Finally, we replicated a laboratory study showing that

[1]Center for Information Technology Policy, Princeton University, Princeton, NJ, USA. [2]Department of Computer Science, University of Bath, Bath BA2 7AY, UK.
*Corresponding author. Email: aylinc@princeton.edu (A.C.); jjb@alum.mit.edu (J.J.B.); arvindn@cs.princeton.edu (A.N.)

Fig. 1. Occupation-gender association. Pearson's correlation coefficient ρ = 0.90 with $P < 10^{-18}$.

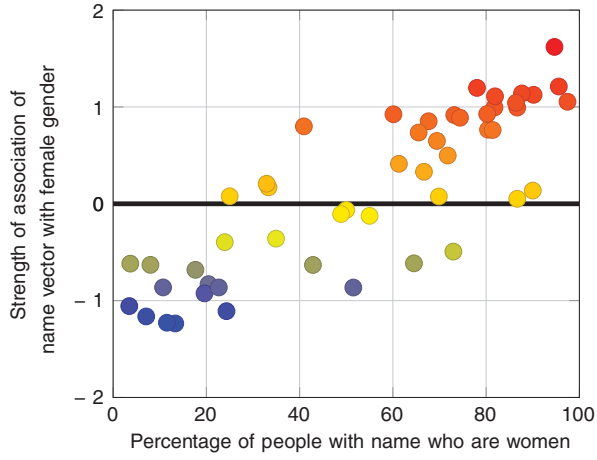

Fig. 2. Name-gender association. Pearson's correlation coefficient ρ = 0.84 with $P < 10^{-13}$.

female words are more associated with the arts than with the sciences (*10*).

Having established that word embeddings contain stereotypes matching those documented with the IAT, we turned to examine how the same embeddings related to veridical data on gender distributions. It has been suggested that implicit gender-occupation biases are linked to gender gaps in occupational participation; however, the relationship between these is complex and may be mutually reinforcing (*11*). To better understand the relationship, we examined the correlation between the gender association of occupation words and labor-force participation data. The x axis of Fig. 1 is derived from 2015 data released by the U.S. Bureau of Labor Statistics (https://www.bls.gov/cps/cpsaat11.htm), which provides information about occupational categories and the percentage of women who have certain occupations under these categories. By applying a second method that we developed, the Word-Embedding Factual Association Test (WEFAT), we found that GloVe word embeddings correlate strongly with the percentage of women in 50 occupations in the United States in 2015.

Similarly, we looked at the veridical association of gender to androgynous names—that is, names used by either gender. In this case, the most recent information that we were able to find was the 1990 census name and gender statistics. Perhaps because of the age of our name data, our correlation was weaker than for the 2015 occupation statistics, but still strikingly significant. In Fig. 2, the x axis is derived from the 1990 U.S. census data (https://www.census.gov/main/www/cen1990.html), and the y axis is as before.

A word embedding is a representation of words as points in a vector space (*12*). For all results in this paper, we used the state-of-the-art GloVe word-embedding method, in which, at a high level, the similarity between a pair of vectors is related to the probability that the words co-occur with other words similar to each other in text (*13*). Word-embedding algorithms such as GloVe exploit dimensionality reduction to substantially amplify the signal found in simple co-occurrence probabilities. In pilot experiments along the lines of those presented here (on free associations rather than implicit associations), raw co-occurrence probabilities were shown to lead to much weaker results (*14*, *15*).

Rather than train the embedding ourselves, we used pretrained GloVe embeddings distributed by its authors. This ensures impartiality, simplifies reproducing our results, and allows us to replicate the effects that may be found in real applications of machine learning. We used the largest of the four corpora provided—the "Common Crawl" corpus obtained from a large-scale crawl of the Internet, containing 840 billion tokens (roughly, words). Tokens in this corpus are case sensitive, resulting in 2.2 million different ones. Each word corresponds to a 300-dimensional vector derived from counts of other words that co-occur with it in a 10-word window.

In the supplementary materials, we also present substantially similar results using an alternative corpus and word embedding.

The details of the WEAT are as follows. Borrowing terminology from the IAT literature, consider two sets of target words (e.g., programmer, engineer, scientist; and nurse, teacher, librarian) and two sets of attribute words (e.g., man, male; and woman, female). The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words. The permutation test measures the (un)likelihood of the null hypothesis by computing the probability that a random permutation of the attribute words would produce the observed (or greater) difference in sample means.

In formal terms, let X and Y be two sets of target words of equal size, and A,B the two sets of attribute words. Let $\cos(\vec{a}, \vec{b})$ denote the cosine of the angle between vectors $\vec{a}$ and $\vec{b}$. The test statistic is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

In other words, $s(w,\text{A,B})$ measures the association of w with the attribute, and $s(X,Y,A,B)$ measures the differential association of the two sets of target words with the attribute.

Let $\{(X_i, Y_i)\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. The one-sided P value of the permutation test is

$$\Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

The effect size is

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)}$$

This is a normalized measure of how separated the two distributions (of associations between the target and attribute) are. We reiterate that these P values and effect sizes do not have the same interpretation as the IAT because the "subjects" in our experiments are words, not people.

The WEFAT allows us to further examine how word embeddings capture empirical information about the world embedded in text corpora. Consider a set of target concepts, such as occupations, and a real-valued, factual property of the world associated with each concept, such as the percentage of workers in the occupation who are women. We would like to investigate whether the vectors corresponding to the concepts embed knowledge of the property—that is, whether there is an algorithm that can extract or predict the property, given the vector. In principle, we could use any algorithm, but in this work we tested the association of the target concept with some set of attribute words, analogous to the WEAT.

Formally, consider a single set of target words W and two sets of attribute words A, B. There is a property p_w associated with each word $w \in W$.

Table 1. Summary of Word-Embedding Association Tests. We replicated eight well-known IAT findings using word embeddings (rows 1 to 3 and 6 to 10); we also help explain prejudiced human behavior concerning hiring in the same way (rows 4 and 5). Each result compares two sets of words from target concepts about which we are attempting to learn with two sets of attribute words. In each case, the first target is found compatible with the first attribute, and the second target with the second attribute. Throughout, we use word lists from the studies we seek to replicate. *N*, number of subjects; N_T, number of target words; N_A, number of attribute words. We report the effect sizes (*d*) and *P* values (*P*, rounded up) to emphasize that the statistical and substantive significance of both sets of results is uniformly high; we do not imply that our numbers are directly comparable with those of human studies. For the online IATs (rows 6, 7, and 10), *P* values were not reported but are known to be below the significance threshold of 10^{-2}. Rows 1 to 8 are discussed in the text; for completeness, this table also includes the two other IATs for which we were able to find suitable word lists (rows 9 and 10). We found similar results with word2vec, another algorithm for creating word embeddings, trained on a different corpus, Google News (see the supplementary materials).

Target words	Attribute words	Original finding				Our finding			
		Ref.	*N*	*d*	*P*	N_T	N_A	*d*	*P*
Flowers vs. insects	Pleasant vs. unpleasant	(5)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(5)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(5)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (5)	(7)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (9)	(7)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(9)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(9)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(9)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

The statistic associated with each word vector is a normalized association score of the word with the attribute

$$s(w,A,B) = \frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std\_dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})}$$

The null hypothesis is that there is no association between $s(w,A,B)$ and p_w. We tested the null hypothesis using a linear regression analysis to predict the latter from the former.

We elaborate on further implications of our results. In psychology, our results add to the credence of the IAT by replicating its results in such a different setting. Further, our methods may yield an efficient way to explore previously unknown implicit associations. Researchers who conjecture implicit associations might first test them using the WEAT on a suitable corpus before testing human subjects. Similarly, our methods could be used to quickly find differences in bias between demographic groups, given large corpora authored by members of the respective groups. If substantiated through testing and replication, the WEAT may also give us access to implicit associations of groups not available for testing, such as historic populations.

We have demonstrated that word embeddings encode not only stereotyped biases but also other knowledge, such as the visceral pleasantness of flowers or the gender distribution of occupations. These results lend support to the distributional hypothesis in linguistics, namely that the statistical contexts of words capture much of what we mean by meaning (*16*). Our findings are also sure to contribute to the debate concerning the Sapir-Whorf hypothesis (*17*), because our work suggests that behavior can be driven by cultural history embedded in a term's historic use. Such histories can evidently vary between languages.

We stress that we replicated every association documented via the IAT that we tested. The number, variety, and substantive importance of our results raise the possibility that all implicit human biases are reflected in the statistical properties of language. Further research is needed to test this hypothesis and to compare language with other modalities, especially the visual, to see if they have similarly strong explanatory power.

Our results also suggest a null hypothesis for explaining origins of prejudicial behavior in humans, namely, the implicit transmission of ingroup/outgroup identity information through language. That is, before providing an explicit or institutional explanation for why individuals make prejudiced decisions, one must show that it was not a simple outcome of unthinking reproduction of statistical regularities absorbed with language. Similarly, before positing complex models for how stereotyped attitudes perpetuate from one generation to the next or from one group to another, we must check whether simply learning language is sufficient to explain (some of) the observed transmission of prejudice.

Our work has implications for AI and machine learning because of the concern that these technologies may perpetuate cultural stereotypes (*18*). Our findings suggest that if we build an intelligent system that learns enough about the properties of language to be able to understand and produce it, in the process it will also acquire historical cultural associations, some of which can be objectionable. Already, popular online translation systems incorporate some of the biases we study (see the supplementary materials). Further concerns may arise as AI is given agency in our society. If machine-learning technologies used for, say, résumé screening were to imbibe cultural stereotypes, it may result in prejudiced outcomes. We recommend addressing this through the explicit characterization of acceptable behavior. One such approach is seen in the nascent field of fairness in machine learning, which specifies and enforces mathematical formulations of nondiscrimination in decision-making (*19*, *20*). Another approach can be found in modular AI architectures, such as cognitive systems, in which implicit learning of statistical regularities can be compartmentalized and augmented with explicit instruction of rules of appropriate conduct (*21*, *22*). Certainly, caution must be used in incorporating modules constructed via unsupervised machine learning into decision-making systems.

REFERENCES AND NOTES

1. M. Stubbs, *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture* (Blackwell, Oxford, 1996).
2. J. A. Bullinaria, J. P. Levy, *Behav. Res. Methods* **39**, 510–526 (2007).
3. T. Mikolov, J. Dean, *Adv. Neural Inf. Process. Syst.* **2013**, 3111–3119 (2013).
4. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, London, 2006).
5. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
6. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, *Adv. Neural Inf. Process. Syst.* **2016**, 4349–4357 (2016).
7. M. Bertrand, S. Mullainathan, *Am. Econ. Rev.* **94**, 991–1013 (2004).
8. M. Bertrand, D. Chugh, S. Mullainathan, *Am. Econ. Rev.* **95**, 94–98 (2005).
9. B. A. Nosek, M. Banaji, A. G. Greenwald, *Group Dyn.* **6**, 101–115 (2002).
10. B. A. Nosek, M. R. Banaji, A. G. Greenwald, *J. Pers. Soc. Psychol.* **83**, 44–59 (2002).

11. B. A. Nosek *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10593–10597 (2009).
12. P. D. Turney, P. Pantel, *J. Artif. Intell. Res.* **37**, 141 (2010).
13. J. Pennington, R. Socher, C. D. Manning, *EMNLP* **14**, 1532–1543 (2014).
14. T. MacFarlane, Extracting semantics from the Enron corpus, University of Bath, Department of Computer Science Technical Report Series; CSBU-2013-08; http://opus.bath.ac.uk/37916/ (2013).
15. W. Lowe, S. McDonald, The direct route: Mediated priming in semantic space, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (LEA, 2000), pp. 806–811.
16. M. Sahlgren, *Ital. J. Linguist.* **20**, 33 (2008).
17. G. Lupyan, *Lang. Learn.* **66**, 516–553 (2016).
18. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 2477899 (2014).
19. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (ACM, 2012), pp. 214–226.
20. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
21. K. R. Thórisson, *Minds Mach.* **17**, 11–25 (2007).
22. M. Hanheide *et al*., *Artif. Intell.* **2015**, j.artint.2015.08.008 (2015).
23. L. L. Monteith, J. W. Pettit, *J. Soc. Clin. Psychol.* **30**, 484–505 (2011).

ACKNOWLEDGMENTS

We are grateful to W. Lowe for substantial assistance in the design of our significance tests; T. Macfarlane for pilot research as a part of his undergraduate dissertation; and S. Barocas, M. Brundage, K. Crawford, C. Lai, and M. Salganik for extremely useful comments on a draft of this paper. We have archived the code and data on Harvard Dataverse (doi: 10.7910/DVN/DX4VWP).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/356/6334/183/suppl/DC1
Materials and Methods
Supplementary Text
Table S1
References

17 November 2016; accepted 9 March 2017
10.1126/science.aal4230