# Data as a mirror of society: Lessons from the emerging science of fairness in machine learning

**Arvind Narayanan**

Machine learning is a set of techniques for discovering patterns in existing data and reproducing them when classifying new inputs. Training data usually reflects human society, including stereotypes, biases and historical prejudices. As a result, demographic disparities in machine learning systems are the rule and not the exception. Algorithmic systems for word analogy generation[1], so-called toxicity detection[2], image search[3], image classification[4], and many, many other tasks have been found to reflect cultural stereotypes. As algorithmic decision making finds their way into criminal justice, hiring, and other consequential settings, it is crucial to understand and mitigate potentially discriminatory outcomes.

These observations give rise to two broad research agendas for social scientists, political scientists, and psychologists. First, an understanding of human culture helps identify possible harmful stereotypes that might be reflected in statistical models and algorithmic systems. For example, a recent paper that I coauthored on stereotypes in word embeddings[5] was made possible by my coauthor's fluency with the Implicit Association Test as a measure of implicit attitudes in people. We developed, essentially, a version of the IAT for machines.

The second research agenda is the converse of the first: if data is a mirror of society, machine learning can be used as a magnifying lens into human culture. A neat example, again using word embeddings, is the finding that they capture 100 years of gender and ethnic stereotypes.[6] In the visual domain, using machine learning on Google street view images can predict the demographic makeup and voting patterns of U.S. neighborhoods.[7] Compared to more traditional statistical methods used in the text-as-data community, the use of some of these cutting edge but less understood methods brings new possibilities but also new pitfalls.

---

[1] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357) 2016.
[2] West, J. "I tested 14 sentences for "perceived toxicity" using Perspectives." Tweet, 2017.
[3] Kay, M., Matuszek, C., & Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3819-3828). ACM, 2015.
[4] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. EMNLP, 2017.
[5] Caliskan, A., Bryson, J. J., & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. Science, 2017.
[6] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), 2018.
[7] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences, 2017.