Input to the Office of the UN High Commissioner for Human Rights

The potential impact of emerging technologies on gender equality

Amit Datta, Anupam Datta, and Michael Carl Tschantz

Increasingly, automated systems make critical decisions about our lives. Often these systems utilize machine learning and other statistical methods over big data sets of personal information. The widespread use of these systems in a broad swath of sectors — credit, education, hiring, online search and advertising, criminal justice, and many others — has fueled concerns over their effects on societal values.

Indeed, recent studies indicate that existing gender inequalities could be amplified by these systems [4, 2, 6]. Society must urgently address this problem. Solutions will have to combine ongoing technical advances with robust public policy that encourages data processors to adopt such solutions. We argue that making *automated decision systems accountable* is a critical requirement to effectively address this problem. We use the term "accountable" to refer to computational mechanisms that can be used to "account for" behaviors of systems to support detection of fairness violations (including gender-based discrimination or bias), as well as explain how they came about. This understanding is then leveraged to repair systems to avoid future violations.

Gender bias in automated systems. We briefly describe three studies that demonstrate how automated systems that employ machine learning and other statistical methods can amplify existing gender inequalities.

A study from Carnegie Mellon University and the International Computer Science Institute found evidence of gender-based discrimination in the targeting of job-related ads [4]. Simulated male and female users with identical job-seeking browsing behavior received significantly different job-related ads. In particular, Google showed the simulated males ads from a career coaching agency that promised large salaries more frequently than the simulated females, a finding suggestive of discrimination. This finding is concerning from a societal standpoint as these ads seem to encourage only men to seek high-paying jobs, which may sustain the existing gender pay gap.

Studies have shown that the use of certain words in job advertisements lead to more or less appeal to certain genders, which can sustain gender inequality in jobs [5]. Researchers at Princeton University uncovered biases in existing text corpora [2]. They found that male attributes were more associated with science, math, and career words, whereas female attributes were associated with arts and family words. Such corpora are often used to train algorithms that automatically produce text and articles, which allows these biases to be incorporated into the models these algorithms learn. When these models are used to write job advertisements, the trend of gender disparity in jobs will continue.

Another study found that Google search images for certain professions were skewed for gender [6]. They uncovered stereotype exaggeration (e.g., only 11 percent of image search results for "CEO" showed women, compared to the 27 percent of U.S. CEOs who are women) as well as slight underrepresentation of women in search results. They also find that skewing the gender representation in image search results can skew people's perceptions about real-world distributions.

The case for accountability. Data subjects who are subjected to decisions by automated systems expect protection from the kinds of discrimination harms described above. These expectations are in tension with the utility goals of data processors (e.g., revenue maximization in online advertising). Accountability provides a means to resolve this tension. By making systems answerable for behavior that is indicative of threats to fairness, it ensures that the interests of data subjects are protected. Further, by detecting and explaining violations, it provides a path toward repairing systems to avoid future violations while minimizing the impact on utility goals.

Enabling accountability often requires a higher level of access to systems than currently available. The studies of the Google ad ecosystem described above were performed with black-box access to the system.

They could detect violations but were unable to explain why they came about. For example, were the gender disparity results of Datta et al. [4] caused by advertisers bidding preferences, the machine learning component picking up on the signal that more men than women were clicking on the high paying job-related ads, or some other factor?

Answering these questions requires new tools for explaining decisions of data-driven systems and correcting their biases that technologists are actively working on [3, 1]. These tools require greater access to the internal components and data used by the systems. External auditors can use these tools to make systems accountable. But for that to happen public policy and regulatory efforts should be directed toward ensuring greater access to decision systems. This development will significantly help combat gender disparity introduced by technology.

Background of authors. We are a group of researchers at Carnegie Mellon University and the International Computer Science Institute. Amit Datta is a doctoral student at CMU, Anupam Datta is an associate professor at CMU, and Michael Carl Tschantz is a research scientist at ICSI. We have been approached by the Office of the United Nations High Commissioner for Human Rights to provide input on the potential impact of big data and emerging technologies such as AI and machine learning on gender equality. This report is in response to that request.

References

[1] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (2016), pp. 4349–4357.

[2] Caliskan-Islam, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora necessarily contain human biases. *ArXiv preprint arXiv:1608.07187* (2016).

[3] Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on* (2016), IEEE, pp. 598–617.

[4] Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies 2015*, 1 (2015), 92–112.

[5] Gaucher, D., Friesen, J., and Kay, A. C. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology 101*, 1 (2011), 109.

[6] Kay, M., Matuszek, C., and 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* (2015), CHI '15.