IBM Response to "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification"

January 2018

The paper "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" by Joy Buolamwini and Timnit Gebru, that will be presented at the Conference on Fairness, Accountability, and Transparency (FAT*) in February 2018, evaluates three commercial API-based classifiers of gender from facial images, including IBM Watson Visual Recognition. The study finds these services to have recognition capabilities that are not balanced over genders and skin tones [1]. In particular, the authors show that the highest error involves images of dark-skinned women, while the most accurate result is for light-skinned men.

To conduct their study, the authors constructed a new facial image dataset, called Pilot Parliaments Benchmark, which is highly balanced across skin phenotype and gender. Besides the use of this dataset for this study, the dataset provides the computer vision research and development community an excellent resource for investigating disparities across multiple dimensions utilizing the Fitzpatrick scale, a numerical classification schema for human skin color.

For the past nine months, IBM has been working towards substantially increasing the accuracy of its new Watson Visual recognition for facial analysis, which now uses different training data and different recognition capabilities than the service evaluated in this study conducted in April 2017. IBM will be bringing this service out in production over next several weeks. As with all of IBM's publicly available software services, we are constantly evaluating and updating them with new features and capabilities that continue to significantly enhance their overall performance.

In the attempt to evaluate our new service in a way that aligns with the one presented in the above paper, IBM Research crawled the face images of parliamentarians from Finland, Iceland, Rwanda, Senegal, South Africa and Sweden to evaluate our new system over this dataset, which is very similar to the Pilot Parliaments Benchmark. However, the experiment conducted by IBM Research is slightly different that the one in [1] in two aspects: 1) the dataset is slightly different as a new election in Senegal has changed the member photos, and 2) Rwanda has a smaller number of photos. Thus, IBM Research has 1,217 faces versus 1,270 reported in the paper. Secondly, we labeled the "lighter" and "darker" classes for the faces manually without using the Fitzpatrick score. We believe that these two minute differences do not impact the conclusions of the experiment. The table below illustrates our results.

As it can be seen, the error rates of IBM's upcoming visual recognition service are significantly lower than those of the three systems presented in the paper. While it is still true that the "darker" category has higher error rates than the lighter one, the highest error rate (which is still for darker skin women) is now much smaller (3.46%). This reflects a nearly ten-fold increase in accuracy. There was no new training or fine-tuning done based on the dataset.

More precisely, the average error over the full dataset is 1.23% while the average error for male subjects in 1.005% vs 1.535% for the female subjects. When we breakdown the errors in male subjects, the "lighter male" class had 0.253% and the "darker male" class had 1.99%. For female subjects, the "lighter female" class had no errors while the "darker female" class had an error rate of 3.46%. It must be noted that these errors are on a small dataset equivalent to the PPB dataset with a total size of 1,217 faces, and may show different error rates when the sample size is larger. The "darker female" class has the lowest number of subjects of all the four classes in this image collection.

Country	Total	Male	Female	Lighter Male	Darker Male	Lighter Female	Darker Female
Finland	194	113	81	113	0	81	0
Iceland	63	39	24	39	0	24	0
Rwanda	26	16	10	0	16	0	10
Senegal	161	95	66	0	95	0	66
South Africa	424	246	178	63	183	27	151
Sweden	349	187	162	180	7	158	4
All	1217	696	521	395	301	290	231
Errors @ score threshold = 0.99	15	7	8	1	6	0	8
Error as %	1.23%	1.005%	1.535%	0.253%	1.99%	0	3.46%

IBM is deeply committed to delivering services that are unbiased, explainable, value aligned, and transparent. To deal with possible sources of bias, we have several ongoing projects to address dataset bias in facial analysis – including not only gender and skin type, but also bias related to age groups, different ethnicities, and factors such as pose, illumination, resolution, expression, and decoration. We are currently in the process of creating a million-scale dataset of face images annotated with attributes and identity, leveraging geo-tags from Flickr images to balance data from multiple countries, and active learning tools to reduce sample selection bias. We intend to make this data publicly available, and propose a challenge at ECCV 2018 to encourage the research community to improve their algorithms with respect to bias in facial analysis. In addition, as a longer-term project, we are planning to conduct research on cycle-consistent adversarial networks to synthetically generate new training samples with specific attributes to reduce dataset bias across race, gender, and age.

More broadly, we are also developing algorithms for detecting, rating, and correcting bias and discrimination across modalities, both for data and for models. For example,

- At FAT* 2018, we present a new text, image and video dataset we have constructed from Bollywood films and analyze it for gender bias in various ways [2].
- In another paper at AIES 2018, we present a composable bias and fairness ratings system and architecture for API-based AI services (including all of the commercial classifiers studied by Buolamwini and Gebru) and demonstrate its applicability in the domain of language translation [3].

- In a paper at NIPS 2017, we presented a flexible optimization approach for transforming a training dataset into one that is more fair according to given protected attributes and can be used by any downstream AI system [4].

We are actively working on transferring these and other research contributions into IBM's core AI offerings. In doing so, we are taking a holistic view in which fairness detections and corrections occur throughout an overall data pipeline in an auditable and transparent manner; this perspective is summarized in a paper presented at the 2017 Data for Good Exchange conference [5].

Moreover, we do not view AI ethics simply from the perspective of easily quantifiable distributive fairness results such as accuracy disparities across gender and skin tone. We are actively pursuing a research agenda that includes explainability, computational morality, value alignment, and other topics that will also be translated into product and service offerings.

In addition, as a founding member of the Partnership on AI (PAI) to Benefit People and Society, IBM is happy to see that researchers such as Buolamwini and Gebru, as well as many others of the global community, are assuming the mantle to audit algorithms and illuminate disparities in artificial intelligence technologies. At IBM, we see these inquiries and the discussions they prompt as essential to promoting the responsible advancement of AI. Further, the PAI has just started a working group on Fair, Transparent, and Accountable AI. Since the MIT Media Lab is a partner of the PAI, we hope that the authors will actively join this working group and work with the other partners to define the research agenda on these topics.

We have also participated in several workshops devoted to bias specifically and AI ethics in general, such as the Beneficial AI conference (Asilomar, Jan.2017), the NYU Algorithms and Explanation conference (New York, April 2017), the AI for Good Global Summit (Geneva, June 2017), the G7 roundtable Towards a Beneficial AI in a Digital Society (Turin, Italy, Sept. 2017), the NSF-sponsored Auditing Algorithms Workshop (Ann Arbor, MI, Sept. 2017), the AI & Society conference (Tokyo, Oct. 2017), the Future of Life Institute workshop on The Ethics of Value Alignment (Laguna Beach, CA, Dec. 2017). We also regularly organize scholarly venues for the discussion and dissemination of such work, including the 2016 ICML Workshop on #Data4Good, the 2018 AAAI Spring Symposium on AI and Society, and the inaugural 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES) taking place next week.

Even before the deployment of AI, IBM believes that organizations that collect, store, manage and process data have an obligation to handle it responsibly. That belief – embodied in our century-long commitment to trust and responsibility in all relationships – is why the world's largest enterprises trust IBM as a steward of their most valuable data. We take that trust seriously and earn it every day by following beliefs and practices outlined in our description of Data Responsibility@IBM:
https://www.ibm.com/blogs/policy/dataresponsibility-at-ibm/
and by working to continually enhance and improve the technologies we bring to the world.

In conclusion, we want to sincerely thank the authors of this paper for their dedication to ensuring a more diverse and inclusive world. As the company that hired our first women and black employees in 1899; our first disabled employee in 1914; our first female VP, Ruth Leach in 1943; wrote our first equal opportunity policy in 1953; stated publicly our non-discrimination on basis of sexual orientation in 1984; and introduced domestic partner benefits in 1996, diversity and inclusion is in our DNA.

Data ethics and AI has to be a conversation and commitment that transcends any one company and we're grateful for your important contribution. Thank you again.

References:

[1]J. Buolamwini and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Conference on Fairness, Accountability, and Transparency*, New York, NY, February 2018.

[2]N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena. "Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies," *Conference on Fairness, Accountability, and Transparency*, New York, NY, February 2018.

[3]B. Srivastava and F. Rossi. "Towards Composable Bias Rating of AI Services," *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, New Orleans, LA, February 2018.

[4]F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurty, and K. R. Varshney. "Optimized Pre-Processing for Discrimination Prevention," *Advances in Neural Information Processing Systems*, Long Beach, CA, December 2017.

[5]S. Shaikh, H. Vishwakarma, S. Mehta, K. R. Varshney, K. N. Ramamurthy, and D. Wei. "An End-To-End Machine Learning Pipeline That Ensures Fairness Policies," *Data for Good Exchange Conference*, New York, NY, September 2017.