Analyzing Demographic Bias in Artificially Generated Facial Pictures

Joni Salminen
Qatar Computing Research Institute; and University of Turku
jsalminen@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute
bjansen@hbku.edu.qa

Soon-gyo Jung
Qatar Computing Research Institute
sjung@hbku.edu.qa

Shammur Chowdhury
Qatar Computing Research Institute
shchowdhury@hbku.edu.qa

Abstract
Artificial generation of facial images is increasingly popular, with machine learning achieving photo-realistic results. Yet, there is a concern that the generated images might not fairly represent all demographic groups. We use a state-of-the-art method to generate 10,000 facial images and find that the generated images are skewed towards young people, especially white women. We provide recommendations to reduce demographic bias in artificial image generation.

CHI 2020 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://10.1145/3334480.3382791

Author Keywords
Image generation; faces; demographics; bias

CSS Concepts
• Human-centered computing~Human computer interaction (HCI)

Introduction
Potential for artificial intelligence (AI)-generated images is tremendous. They could be implemented for various purposes where acquiring real photographs is not feasible due to cost, time, or scalability. An example is a system for data-driven persona creation that automatically generates personas, which are fictitious people representing behavioral user segments [3]. Additional use cases for artificial facial pictures include advertising [14], virtual avatars [1], fashion [16], and so on. Providing realistic faces can enhance the user experience and immersion, while mitigating issues of copyright, privacy, and financial cost.

Despite the potential of AI-generated images, there has been an increasing concern over the various ethical aspects relating to AI systems, such as transparency,

Figure 1: The problem: when not generating faces of minority groups, biased image generation diminishes individuals' abilities "to be seen and heard". Image credit: Patrick Doheny (https://www.flickr.com/photos/14132971@N05/2432071644)

fairness, and accountability [4, 6, 7, 18, 20]. One risk is that the outputs and decisions of the algorithms might work for majority views and groups, while excluding minority groups [23]. In the context of AI-generated images, these concerns relate to the demographic attributes of the output images. If an algorithm outputs, for example, only white males, that can be considered as an issue for diversity when applying the imagery to actual systems.

Artificial image creation can manifest forms of algorithmic bias, defined as systematic and unfair discrimination against certain groups in favor of others [8]. Essentially, the question is of demographic parity [10], i.e., the principle that all ages, genders, and races should have an equal opportunity of being "chosen" by the algorithm (Figure 1). The lack of representation can make minority groups "invisible" when using the artificially generated images in real applications. This can enhance stereotypes and biased decision making.

While scholars have awakened to the call of ethics in algorithmic systems [18], investigating demographic bias in AI-generated images has not been conducted. In this research, we analyze the demographic bias (specifically, age, gender, race) in AI-generated images. Our results indicate that there is a need for more diverse outputs in the AI-generated images, as in terms of age, young people are predominant, in terms of gender, young females are predominant, and in terms of race, white people are predominant.

Related Work

Previous research in artificial image generation has largely ignored the analysis of demographic bias in generated facial images. Rather, the research has focused on technical metrics [25], user perceptions, and authors' subjective evaluation.

Li et al. [16] evaluated makeup test images for quality, realism and makeup style similarity. They did not consider demographic bias. Lee et al. [15] asked users to rate realism of pictures from different generation methods. Again, they did not investigate demographic bias. Choi et al. [5] asked crowd workers to rank the generated images based on realism, quality of attribute transfer, and preservation of the person's original identity, without investigating demographic bias. Zhang et al. [26] recruited crowd raters to evaluate the similarity of generated and reference pictures, overlooking demographic variables. Izuka et al. [11] recruited 10 volunteers to evaluate the "naturalness" of the generated pictures; the volunteers were asked to guess if a picture was real or generated. Yin et al. [24] asked students to compare generated pictures with original pictures using saliency, quality, and identity.

These previous studies illustrate the general lack of analyzing the possible demographic bias in artificially generated images. The major exceptions are "Fairness GAN", a methodology for ensuring demographic parity in the generated images [21] and DebFace, a "de-biasing adversarial network that learns to extract disentangled feature representations for both unbiased face recognition and demographics estimation" (p. 1). However, neither of these studies focus on StyleGAN. Analyzing StyleGAN is important because it is made publicly available and it generates photo-realistic images. Moreover, the DebFace study is focused on image recognition, not generation [9].

Instructions to Crowd Raters
You are shown a facial picture of a person. Look at the picture and choose how well it represents a real person. The options:
• **5: Perfect** - the picture is indistinguishable from a real person. • **4: High quality** - the picture has minor defects, but overall it's pretty close to a real person. • **3: Medium quality** - the picture has some flaws that suggest it's not a real person. • **2: Low quality** - the picture has severe malformations or defects that instantly show it's a fake picture. • **1: Unusable** - the picture does not represent a person at all.

Sidebar 1: Rating instructions.

Methodology

Image Generation

In this research, we use a pretrained model from the creators of StyleGAN, a state-of-the-art image generator [13]. The model we use was trained on CelebA-HQ and FFHQ datasets using eight Tesla V100 GPUs. It is implemented in TensorFlow, a machine learning (ML) library and is open sourced in a GitHub repository[1] that contains the model and the required source code to run it. We use this pre-trained model to generate 10,000 facial pictures, a number we deem sufficient to adequately evaluate the demographic properties in the output images. Technical details of how StyleGAN works can be found in Karras et al. [13]. Figure 2 shows examples of the output generated for this study, with the full set available upon request.

Figure 2: Examples of the AI-generated images of this study.

Demographic Tagging

To analyze the generated images, we tagged them for demographic attributes using Face++[2]. Even though there are multiple services for automatic tagging of face images (see [12]), Face++ is currently the only publicly available service that outputs race in addition to age group and gender. Previous research also has shown that the Face++ provides a satisfactory detection of a person's age and gender from facial pictures [12]. For these reasons, we use Face++ for the demographic tagging. We sent the 10K images to Face++ via the API and received an output of each picture's age group, gender, and race. One of the images did not provide a result for any of the attributes, and 381 images did not return an age or gender, most likely because of the picture representing a younger than 13-year-old person. Given this, we use 96.2% (N=9,618) of the images for the analysis.

Quality Rating

We randomly selected a sample of 1,000 (10% of the total) images for quality evaluation using crowdsourced ratings. Crowd raters were instructed (see Sidebar 1). We used the Figure Eight platform to collect the ratings, as this platform has been widely used for gathering annotations [2] [19] in various subdomains of computer science. The pictures were shown in the full 1024x1024 pixel format to provide the crowd raters enough detail for a valid evaluation, as well as examples from each quality category. For each image, we obtained ratings from three independent crowd raters. Results of the demographic tagging and quality ratings are shown in the following section.

[1] https://github.com/NVlabs/stylegan

[2] https://www.faceplusplus.com/

Age group	MtF
18-24	**0.485**
25-34	**0.505**
35-44	**0.772**
55-64	1.047
45-54	1.101
65-	1.137
13-17	1.149

Table 1: Male-to-Female (MtF) rations for age groups. Age groups higher female prevalence are bolded.

	Count	Share
White	**7,263**	**72.6%**
Asian	1,382	13.8%
Black	1,010	10.1%
Indian	344	3.4%

Table 2: Generated pictures by Race. The largest group bolded.

Results

Age and Gender

Gender of the generated images is roughly balanced in the dataset, with females accounting for 56.9% (N=5,690) and males 43.1% (N=4,309) of the pictures. In terms of Age, there is a bias towards young people, as the total share of young people (between 18-34) is 42.1% (N=4,054) of the pictures, accounting for almost half of the pictures (see Figure 3).

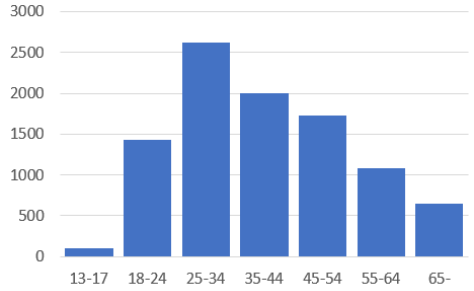

Figure 3: Number of generated images per age group.

Even though the aggregated results show that gender is balanced, in terms of Age and Gender, we observe that the model is biased toward generating Young Females. This can be seen in Figure 4 that shows that in the Age groups between 18-44 years, the Gender ratio is skewed towards females. After that, the Gender ratio becomes more balanced, with males being slightly more predominant in the 45+ age groups.

To quantify the bias, we compute the Male-to-Female Ratios (MtF = count of males / count of females) for different groups. These ratios (see Table 1) corroborate the finding that women are more prevalent in the younger age groups and less prevalent in the older age groups (as well as in 13-17 age group). Women were also more common in the output for all Races (see also Figure 5), apart from Indian, for which males were much more common (MtF = 2.02). For White, the ratio was the most balanced, i.e., closest to one (MtF = 0.83), while for Asian (MtF = 0.46) and Black (MtF = 0.54) females were much more common.

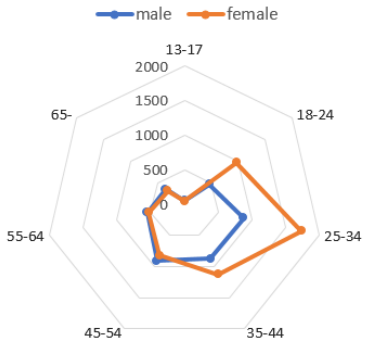

Figure 4: Gender counts of the generated pictures by age.

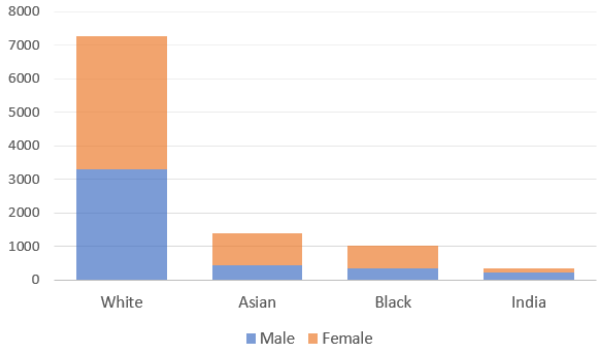

Figure 5: Counts of generated 10K images by Race.

Race

Results (see Table 2; Figure 5) indicate a racial bias among the generated pictures, with close to three-thirds (72.6%) of the pictures representing White people. Asian (13.8%) and Black (10.1%) are considerably less frequent, while Indians represent only a minor fraction of the pictures (3.4%). Note that, technically, Indians could be considered as Asian (as India is part of the Asian continent), but since Face++ separates Indians from other Asians, we present this category separately.

If we adopt the concept of demographic parity, we would expect that an unbiased algorithm would produce pictures with the same probability for each race. This is clearly not the case for StyleGAN, as the network is much more likely to generate pictures of White people. The prevalence rate of White is P = (7263 / 9999) / (1/4) = 2.9 times the expected random probability. In turn, the prevalence rate for Asian is 0.55 times the expected rate, 0.40 for Black, and 0.14 for Indian.

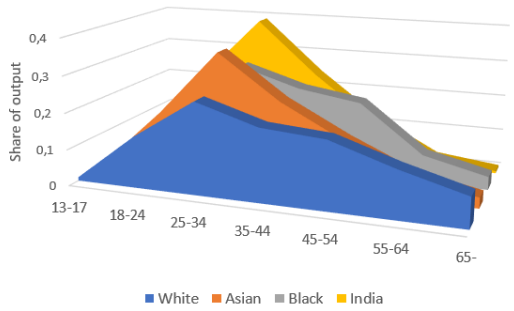

Figure 6: Frequency of images by Age and Race.

In terms of Race, White and Black are the most balanced concerning age distribution (less peaking), while Asian and Indian output pictures are more skewed toward younger faces (see Figure 6).

Image Quality

To analyze the quality, we compare the quality rating distributions between the demographic categories. Results show no major differences in image quality by race (see Figure 7). While the raw numbers indicate Asian pictures are more often situated in the lower quality spectrum than other races, as medium and low-quality pictures comprise close to half (49%) of the generated Asian pictures, the chi-square result testing the association between Race and Quality is not significant (X^2 = 9.389; p = .402), indicating that quality of the pictures does not significantly differ by race. Quality does not significantly vary by gender either, with the two genders achieving very similar mean quality ratings (M_{MALE} = 3.72, M_{FEMALE} = 3.71). The same applies for age, with the quality not varying significantly by age group. Thus, we observe no significant demographic bias in terms of quality.

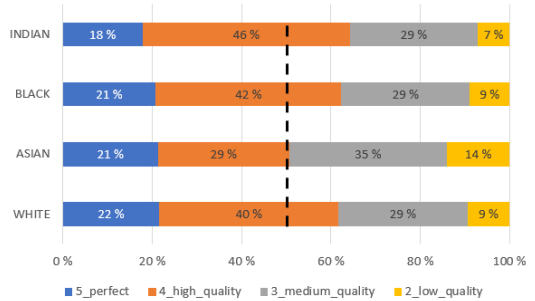

Figure 7: Quality ratings by Race. Line in the middle denotes center point (50%).

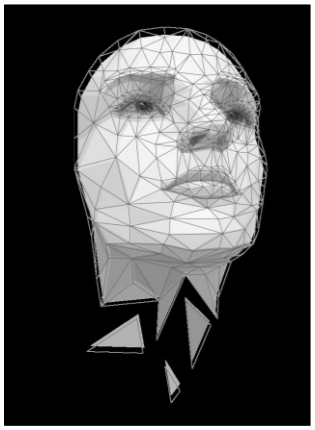

Figure 8: IBM's 'Diversity in Faces' dataset [16] is a good example of diversity-adjusted training data for artificial image generation. Image credit: IBM (https://www.research.ibm.com/artificial-intelligence/trusted-ai/diversity-in-faces/images/face2.png).

Discussion

Our results indicate room for improvement in the diversity of AI-generated images. At least for StyleGAN, there is a clear bias towards generating images containing young white adults (especially women) rather than elderly people and other races. In terms of gender, we find that although females and males are generated in similar proportion, in the younger age groups, females are more prevalent than males. Interestingly, the bias seems to be limited to the demographic attributes in the output, while not affecting the quality of the images. In other words, even though the model generates disproportionally more whites and young people, the quality of the images representing minority groups (other races and the elderly) is at par with the majority groups.

The bias in the generated images can most likely be explained by the properties of the training data [17] – that is, the training data is likely to contain more samples of young white women than other demographic groups. Although we did not specifically investigate the training set distributions (as the demographic labels are not available), given the general nature of machine learning and GANs in particular, this inference seems sensical, as the output distribution should follow the training distribution.

Interestingly, Karras et al. [13] specifically mention the risk of bias on the GitHub repository ("The images were crawled from Flickr, thus inheriting all the biases of that website"); yet, they evidently did not consider this as a major risk for the application of their model. This exhibits, according to our understanding of the ML field, the tendency of the authors to be primarily interested in showcasing the technical performance of their models, while overlooking the repercussions of actual systems implementing their models (see exception in Figure 8). Even though retraining StyleGAN from scratch can be done, developers may opt for using the pretrained model, because acquiring more training data is costly, and retraining large models can take weeks even with proper hardware. Therefore, researchers that make public their retrained models should consider investigating biases in the model outputs and openly state that any systems or applications implementing the pretrained models risk inheriting these biases in their output. As our analysis shows, the risk of demographic biases in AI-generated images is real.

In general, demographic bias in AI-generated images could be addressed in two ways: (a) *extending the training data to include demographically balanced representation of populations* (i.e., including an even number of samples from young and mature, as well as different races and age groups), and/or (b) *providing parameters for conditional generation of the images*, i.e., adjusting purposefully the demographic attributes of the generated images. Even in the latter case, some retraining is likely required, because the network was not able to produce any pictures from some demographic groups (i.e., Indian 13-17).

As training algorithms is at least partly dependent on datasets in the wild, it remains an open question how to encourage participation by underprivileged groups in areas touched by the "digital divide" [22]. The consequence of white females posting more online photos also increases the changes of algorithms learning that these groups are more "important". Thus, the issue is linked to the larger context of online justice and equal participation.

References

[1] Ablanedo, J. et al. 2018. Is This Person Real? Avatar Stylization and Its Influence on Human Perception in a Counseling Training Environment. *International Conference on Virtual, Augmented and Mixed Reality* (2018), 279–289.

[2] Alam, F. et al. 2018. Processing social media images by combining human and machine computing during crises. *International Journal of Human–Computer Interaction*. 34, 4 (2018), 311–327.

[3] An, J. et al. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB)*. 12, 3 (2018).

[4] Chander, A. 2016. The racist algorithm. *Mich. L. Rev.* 115, (2016), 1023.

[5] Choi, Y. et al. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 8789–8797.

[6] Diakopoulos, N. and Koliska, M. 2017. Algorithmic transparency in the news media. *Digital Journalism*. 5, 7 (2017), 809–828.

[7] Eslami, M. et al. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), 432.

[8] Friedman, B. and Nissenbaum, H. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*. 14, 3 (1996), 330–347.

[9] Gong, S. et al. 2019. DebFace: De-biasing Face Recognition. *ArXiv*. abs/1911.08080, (2019).

[10] Hardt, M. et al. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* (2016), 3315–3323.

[11] Iizuka, S. et al. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* 36, 4 (Jul. 2017), 107:1–107:14. DOI:https://doi.org/10.1145/3072959.3073659.

[12] Jung, S.-G. et al. 2017. Inferring social media users' demographics from profile pictures: A Face++ analysis on Twitter users. *Proceedings of 17th International Conference on Electronic Business* (Dubai, Dec. 2017).

[13] Karras, T. et al. 2018. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*. (2018).

[14] Kietzmann, J. et al. 2018. Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey. *Journal of Advertising Research*. 58, 3 (2018), 263–267.

[15] Lee, H.-Y. et al. 2018. Diverse Image-to-Image Translation via Disentangled Representations. *arXiv:1808.00948 [cs]*. (Aug. 2018).

[16] Li, T. et al. 2018. BeautyGAN: Instance-level Facial Makeup Transfer with Deep Generative Adversarial Network. *Proceedings of the 26th ACM International Conference on Multimedia* (New York, NY, USA, 2018), 645–653.

[17] Merler, M. et al. 2019. Diversity in Faces. *arXiv:1901.10436 [cs]*. (Apr. 2019).

[18] Mittelstadt, B.D. et al. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*. 3, 2 (Nov. 2016), 2053951716679679. DOI:https://doi.org/10.1177/2053951716679679.

[19] Salminen, J. et al. 2018. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)* (Valencia, Spain, Oct. 2018).

[20] Salminen, J. et al. 2019. Persona Transparency: Analyzing the Impact of Explanations on Perceptions of Data-Driven Personas. *International Journal of Human–Computer Interaction*. 0, 0 (Nov. 2019), 1–13. DOI:https://doi.org/10.1080/10447318.2019.1688946.

[21] Sattigeri, P. et al. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*. 63, 4/5 (Jul. 2019), 3:1-3:9. DOI:https://doi.org/10.1147/JRD.2019.2945519.

[22] Van Dijk, J. and Hacker, K. 2003. The digital divide as a complex and dynamic phenomenon. *The information society*. 19, 4 (2003), 315–326.

[23] Williams, B.A. et al. 2018. How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*. 8, (2018), 78–115. DOI:https://doi.org/10.5325/jinfopoli.8.2018.0078.

[24] Yin, W. et al. 2017. Semi-Latent GAN: Learning to generate and modify facial images from attributes. *arXiv:1704.02166 [cs]*. (Apr. 2017).

[25] Yuan, Y. et al. 2019. Locally and multiply distorted image quality assessment via multi-stage CNNs. *Information Processing & Management*. (Nov. 2019), 102175. DOI:https://doi.org/10.1016/j.ipm.2019.102175.

[26] Zhang, R. et al. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv:1801.03924 [cs]*. (Jan. 2018).