Gender Slopes: Counterfactual Fairness for Computer Vision Models by Attribute Manipulation

Jungseock Joo
UCLA

Kimmo Kärkkäinen
UCLA

ABSTRACT

Automated computer vision systems have been applied in many domains including security, law enforcement, and personal devices, but recent reports suggest that these systems may produce biased results, discriminating against people in certain demographic groups. Diagnosing and understanding the underlying true causes of model biases, however, are challenging tasks because modern computer vision systems rely on complex black-box models whose behaviors are hard to decode. We propose to use an encoder-decoder network developed for image attribute manipulation to synthesize facial images varying in the dimensions of gender and race while keeping other signals intact. We use these synthesized images to measure counterfactual fairness of commercial computer vision classifiers by examining the degree to which these classifiers are affected by gender and racial cues controlled in the images, e.g., feminine faces may elicit higher scores for the concept of nurse and lower scores for STEM-related concepts.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; *Neural networks*; • **Applied computing** → **Sociology**.

KEYWORDS

Model Bias, Generative Models, Facial Attribute Manipulation

ACM Reference Format:
Jungseock Joo and Kimmo Kärkkäinen. 2020. Gender Slopes: Counterfactual Fairness for Computer Vision Models by Attribute Manipulation. In *2nd International Workshop on Fairness: Affiliation; Accountability: Affiliation; Transparency and Ethics in MultiMedia (FATE/MM'20), October 12, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3422841.3423533

1 INTRODUCTION

Artificial Intelligence has made remarkable progress in the past decade. Numerous AI-based products have already become prevalent in the market, ranging from robotic surgical assistants to self-driving vehicles. The accuracy of AI systems has surpassed human capability in challenging tasks, such as face recognition [21], lung cancer screening [2] and pigmented skin lesion diagnosis [22]. These practical applications of AI systems have prompted attention and support from industry, academia, and government.

FATE/MM'20, October 12, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8148-2/20/10.
https://doi.org/10.1145/3422841.3423533

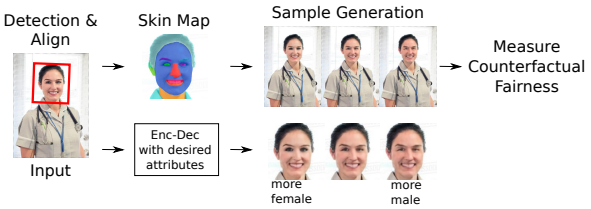

Figure 1: Overview of our method for counterfactual image synthesis.

While AI technologies have contributed to increased work productivity and efficiency, a number of reports have also been made on the algorithmic biases and discrimination caused by data-driven decision making in AI systems. For example, COMPAS, an automated risk assessment tool used in criminal justice [4], was reported to contain bias against Black defendants by assigning higher risk scores to Black defendants than White defendants [1]. Another recent study also reports the racial and gender bias in computer vision APIs for facial image analysis, which were shown less accurate on certain race or gender groups [5].

While previous reports have shown that popular computer vision and machine learning models contain biases and exhibit disparate accuracies on different subpopulations, it is still difficult to identify true causes of these biases. This is because one cannot know to which variable or factor the model responds. If we wish to verify if a model indeed discriminates against a sensitive variable, e.g., gender, we need to isolate the factor of gender and intervene its value for **counterfactual** analysis [8].

The objective of our paper is to adopt an encoder-decoder architecture for facial attribute manipulation [15] and generate counterfactual images which vary along the dimensions of sensitive attributes: gender and race. These synthesized examples are then used to measure counterfactual fairness of black-box image classifiers offered by commercial providers. Figure 1 shows the overall process of our approach. Given an input image, we detect a face and generate a series of novel images by manipulating the target sensitive attributes while maintaining other attributes. We summarize our main contributions as follows.

(1) We propose to use an encoder-decoder network [15] to modify cues in face images, which allows counterfactual interventions. Unlike previous methods [6], our method explicitly isolates the factors for sensitive attributes, which is critical in identifying true causes to model biases.
(2) We construct a novel image dataset which consists of 64,500 original images collected from web search and more than 300,000 synthesized images manipulated from the original images. These images describe people in diverse occupations and can be used for studies on bias measurement or mitigation. Both the code and data will be made publicly available.

(3) Using new methods and data, we measure counterfactual fairness of commercial computer vision classifiers and report whether and how sensitive these classifiers are affected along with attributes being manipulated by our model.

2 RELATED WORK

Fairness in computer vision is becoming more critical as many systems are being adapted in real world applications. For example, face recognition systems such as Amazon's Rekognition are being used by law enforcement to identify criminal suspects [9]. If the system produces biased results (e.g., higher false alarm on Black suspects), then it may lead to a disproportionate arrest rate on certain demographic groups. In order to address this issue, scholars have attempted to identify biased representations of gender and race in public image dataset and computer vision models [11, 12, 17, 18]. Buolamwini and Gebru [5] have shown that commercial computer vision gender classification APIs are biased and thus perform least accurately on dark-skinned female photographs. [14] has also reported that image classification APIs may produce different results on faces in different gender and race. These studies, however, used the existing images without interventions, and thus it is difficult to identify whether the classifiers responded to the sensitive attributes or to the other visual cues. [14] used the headshots of people with clean white background, but this hinders the classifiers from producing many comparable tags.

Our paper is most closely related to Denton et al. [6], who use a generative adversarial network (GAN) [7] to generate face images to measure counterfactual fairness. Their framework incorporates a GAN trained from a face image dataset called CelebA [16], and generates a series of synthesized samples by modifying the latent code in the embedding space to the direction that would increase the strength in a given attribute (e.g., smile). Our paper differs from this work for the following reasons. First, we use a different method to examine the essential concept of counterfactual fairness by generating samples that separate the signals of the sensitive attributes out from the rest of the images. Second, our research incorporates the generated data to measure the bias of black-box image classification APIs whereas [6] measures the bias of a dataset open to public [16]. Using our distinct method and data, we aim to identify the internal biases of models trained from unknown data.

3 COUNTERFACTUAL DATA SYNTHESIS

3.1 Problem Formulation

The objective of our paper is to measure counterfactual fairness of a predictor Y, a function of an image x. This predictor is an image classifier that automatically labels the content of input images. Without the loss of generality, we consider a binary classifier, $Y(x) = \{True, False\}$. This function classifies, for example, whether the image displays a doctor or not. We also define a sensitive attribute, A, gender and race. Typically, A is a binary variable in the training data, but it can take a continuous value in our experiment since we can manipulate the value without restriction. Following [8], this predictor satisfies counterfactual fairness if $P(Y_{A\leftarrow a}(x) = y|x) = P(Y_{A\leftarrow a'}(x) = y|x)$ for all y and any a and a', where $A \leftarrow a$ indicates an intervention on the sensitive attribute, A. We now explain how this is achieved by an encoder-detector network.

The goal of this intervention is to manipulate an input image such that it changes the cue related to the sensitive attribute while retaining all the other signals. We consider two sensitive attributes: gender and race. We manipulate facial appearance because face is the strongest cue for gender and race identification [19].

3.2 Counterfactual Data Synthesis

Before we elaborate our proposed method for manipulating sensitive attributes, we briefly explain why such a method is necessary to show if a model achieves counterfactual fairness. For an in-depth introduction to the framework of counterfactual fairness, we refer the reader to Kusner et al. [13].

Many studies have reported skewed classification accuracy of existing computer vision models and APIs between gender and racial groups [5, 12, 14, 26]. However, these findings are based on a comparative analysis, which directly compares the classifier outputs between male and female images (or White and non-White) in a given dataset. The limitation of the method is that it is difficult to identify true sources of biased model outputs due to hidden confounding factors. Even though one can empirically show differences between gender groups, such differences may have been caused by non-gender cues such as hair style or image backgrounds (see [20], for example). Since there exists an infinite number of possible confounding factors, it will be very difficult to control for all of them.

Consequently, recent works in bias measurement or mitigation have adopted generative models which can synthesize or manipulate text or image data [6, 27]. These methods generate hypothetical data in which only sensitive attributes are switched. These data can be used to measure counterfactual fairness but also augment samples in existing biased datasets.

3.3 Face Attribute Synthesis

From the existing methods available for face attribute manipulation [3, 10, 24], we chose FaderNetwork [15] as our base model. FaderNetwork is a computationally efficient model that produces plausible results, but we made a few changes to make it more suitable for our study.

Specifically, FaderNetwork is based on an encoder-decoder network with two special properties. First, it separates the sensitive attribute, a, from its encoder output, $E(x)$, and both are fed into the decoder, such that it can reconstruct the original image, i.e., $D(E(x), a) \approx x$. Second, it makes $E(x)$ invariant to a by using adversarial training such that the discriminator cannot predict the correct value for a given $E(x)$. At test time, an arbitrary value for a can be given to obtain an image with a modified attribute value.

Since we want to minimize the change by the model to dimensions other than the sensitive attributes, we added two additional steps as follows. First, we segment the facial skin region from an input face by [25] and only retain changes within the region. This prevents the model from affecting background or hair regions. Second, we control for the effects of other attributes (e.g., smiling or young) which may be correlated with the main sensitive attribute, such that their values remain intact while being manipulated. This

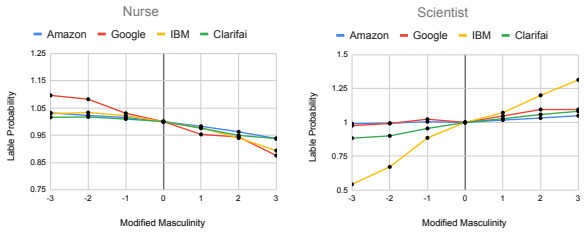

Figure 2: The sensitivity of image classification APIs for Nurse and Scientist to the modified facial gender cues.

was achieved by first modeling these attributes as the main sensitive attributes along with y in training and fixing their values at testing time.

4 EXPERIMENTS

4.1 Computer Vision APIs

We measured counterfactual fairness of commercial computer vision APIs which provide label classification for a large number of visual concepts, including Google Vision API, Amazon Rekognition, IBM Watson Visual Recognition, and Clarifai. These APIs are widely used in commercial products as well as academic research [23]. While undoubtedly useful, these APIs have not been fully verified for their fairness. They may be more likely to generate more "positive" labels for people in certain demographic groups. These labels may include highly-paid and competitive occupations such as "doctor" or "engineer" or personal traits such as "leadership" or "attractive". We measure the sensitivity of these APIs using counterfactual samples generated by our models.

4.2 Occupational Images

We constructed the baseline data that can be used to synthesize samples. We are especially interested in the effects of gender and race changes on the profession related labels provided by the APIs, and thus collected a new dataset of images related to various professions. We first obtained a list of 129 job titles from the Bureau of Labor Statistics (BLS) website and used Google Image search to download images. To obtain more diverse images, we additionally combined six different keywords (male, female, African American, Asian, Caucasian, and Hispanic). This results in around 250 images per keyword.

We also used datasets for training our model. For the gender manipulation model, we used CelebA [16], a popular face attribute dataset with 40 labels. This dataset mostly contains the faces of White people, and thus is not suitable for the race manipulation model. There is no publicly available dataset with a sufficiently large number of African Americans. Instead, we obtained the list of the names of celebrities for each gender and each ethnicity from an online website, FamousFix (roughly 5,000 people for each group). Then we used Google Image search to download up to 30 images for each celebrity. We estimated the true gender and race of each face by a model trained from a public dataset [12] and manually verified examples with lower confidences. Finally, this dataset was combined with CelebA to train the race manipulation model.

After training, two models (gender and race) were applied to the profession dataset to generate a series of manipulated images for each input image. Since many images contain multiple faces, we only manipulated the face closest to the center of it for these images. These faces are pasted into the original image, only on the facial skin region, and passed to each of the 4 APIs we tested. All the APIs provide both the presence of each label (binary) and the continuous classification confidence if the concept is present in the image.

4.3 Results

The sensitivity of a classifier with respect to the changes in gender or race cues of images is measured as a slope estimated from the assigned attribute value, a, and the model output, $Y(x(a))$, where $x(a)$ is a synthesized image with its attribute manipulated to the value a. The range of a was set to $(-2, 2)$. The center, i.e., gender-neutral face, is 0. $(-1, 1)$ is the range observed in training, and $(-2, 2)$ will extrapolate images beyond the training set. In practice, this still results in natural and plausible samples. From this range, we sampled 7 evenly spaced images for gender manipulation and 5 images for race manipulation.Let us denote x^i, the i-th input image, and $\{x_1^i, .., x_K^i\}$, the set of K synthesized images ($K = 7$). For each label in Y, we obtain 7 scores. From the entire image set $\{x^i\}$, we obtain a normalized classifier output vector:

$$y_k = \frac{1}{n} \sum_i \mathbb{1}\{Y(x_k^i) = True\}, k \in \{1, .., K\},$$
$$z_k = y_k / y_c, c = (K + 1)/2.$$

That is, we normalize the vector such that z_c is always 1 to allow comparisons across concepts. The slope b is obtained by linear regression with ordinary least squares. The magnitude of b determines the sensitivity of the classifier against a, and its sign indicates the direction.

Table 1 and 2 show the list of labels returned by each API, more frequently activated with images manipulated to be closer to women and to men, respectively. Not surprisingly, we found the models behave in a closely related way to the actual gender gap in many occupations such as nurses or scientists (see Figure 2, too). One can imagine this bias was induced at least in part due to the bias in the online media and web, from which the commercial models have been trained. Table 3 shows skewed gender and race representations in our main dataset of peoples' occupations. Indeed, many occupations such as nurse or engineer exhibit very sharp gender contrast, and this may explain the behaviors of the image classifiers. Figure 3 shows example images and their label prediction scores.

Similarly, Table 4 shows the labels which are most sensitive to the race manipulation. The tables show all the dimensions which are significantly correlated with the model output ($p < 0.001$), except plain concepts such as "Face" or "Red color".

5 CONCLUSION

AI fairness is an increasingly important criterion to evaluate models and systems. In real world applications, especially for private models whose training processes or data are unknown, it is difficult to identify their biased behaviors or to understand the underlying causes. We introduced a novel method based on facial attribute manipulation by an encoder-decoder network to synthesize counterfactual samples, which can help isolate the effects of the main

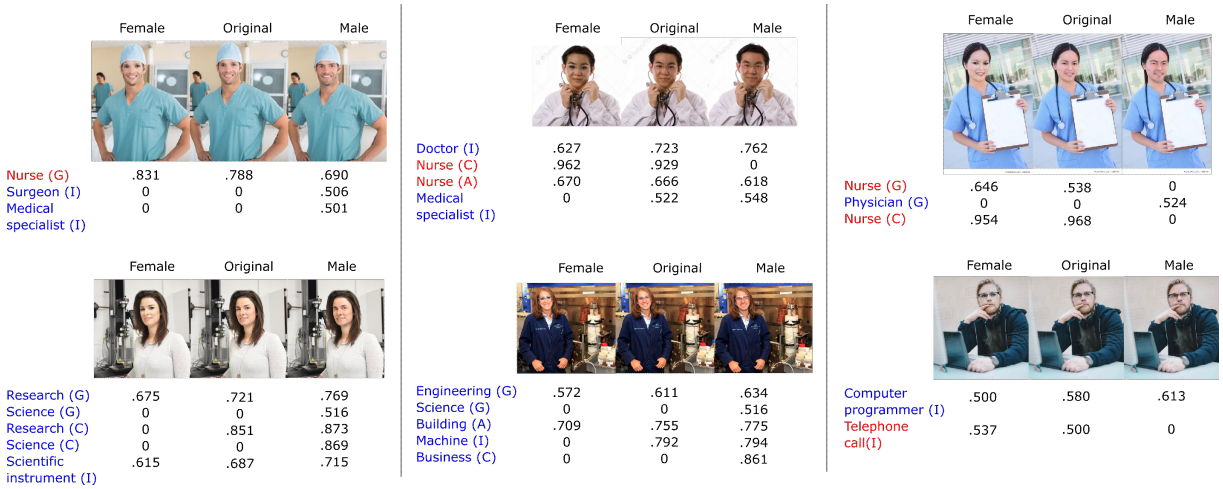

Figure 3: Example images and label prediction scores from APIs (G:Google, A:Amazon, I:IBM, C:Clarifai). "0" means the label was not detected. Blue labels indicate an increasing score with increasing masculinity (red for femininity). Some images were clipped to fit the space. Zoom in to see the details.

Table 1: The Sensitivity of Label Classification APIs against Gender Manipulation (Female). (All tables only show labels with p-value < 0.001 and | slope | > 0.03.)

API	Label	Slope	API	Label	Slope
Amazon	Nurse	-.031	IBM	Secretary of State	-.107
Google	Fashion model	-.262	IBM	gynecologist	-.099
Google	Model	-.261	IBM	celebrity	-.097
Google	Secretary	-.140	IBM	newsreader	-.090
Google	Nurse	-.073	IBM	cleaning person	-.081
IBM	anchorperson	-.213	IBM	nurse	-.046
IBM	television reporter	-.155	IBM	laborer	-.044
IBM	college student	-.151	IBM	workman	-.041
IBM	legal representative	-.147	IBM	entertainer	-.040
IBM	careerist	-.128	Clarifai	secretary	-.273
IBM	host	-.125	Clarifai	receptionist	-.268
IBM	steward	-.110	Clarifai	model	-.211

Table 2: The Sensitivity of Label Classification APIs against Gender Manipulation (Male).

API	Label	Slope	API	Label	Slope
Amazon	Executive	.055	IBM	repairer	.151
Amazon	Attorney	.113	IBM	resident commissioner	.159
Google	Engineer	.038	IBM	sports announcer	.164
Google	Spokesperson	.040	IBM	investigator	.174
Google	Blue-collar worker	.056	IBM	sociologist	.213
IBM	subcontractor	.043	IBM	scientist	.254
IBM	official	.049	Clarifai	politician	.037
IBM	medical specialist	.050	Clarifai	athlete	.048
IBM	contractor	.061	Clarifai	construction worker	.053
IBM	player	.061	Clarifai	police	.054
IBM	diplomat	.063	Clarifai	singer	.056
IBM	detective	.081	Clarifai	musician	.056
IBM	radiologist	.082	Clarifai	scientist	.070
IBM	military officer	.107	Clarifai	worker	.078
IBM	biographer	.109	Clarifai	waiter	.082
IBM	Secretary of the Int.	.114	Clarifai	inspector	.085
IBM	internist	.119	Clarifai	surgeon	.087
IBM	speaker	.122	Clarifai	repairman	.125
IBM	security consultant	.131	Clarifai	writer	.153
IBM	high commissioner	.134	Clarifai	machinist	.192
IBM	cardiologist	.140	Clarifai	film director	.342
IBM	Representative	.142			

Table 3: Skewed gender representations in Google Image search result

Occupation	Female %	Occupation	Male %
nutritionist	.921	pest control worker	.971
flight attendant	.891	handyman	.964
hair stylist	.884	logging worker	.950
nurse	.860	basketball player	.925
medical assistant	.847	businessperson	.920
dental assistant	.835	chief executive officer	.917
merchandise displayer	.821	lawn service worker	.909
nursing assistant	.821	electrician	.901
dental hygienist	.815	barber	.901
veterinarian	.784	repair worker	.900

Table 4: The Sensitivity of Label Classification APIs against Race Manipulation.

API	Label (Black)	Slope	API	Label (White)	Slope
IBM	woman orator	-.690	IBM	careerist	.179
IBM	President of the U.S.	-.367	IBM	dermatologist (doctor)	.127
IBM	first lady	-.323	IBM	legal representative	.111
IBM	high commissioner	-.284	IBM	business man	.093
IBM	Representative	-.225	Clarifai	repair	.074
IBM	scientist	-.183	Clarifai	beautiful	.074
IBM	worker	-.131	Clarifai	repairman	.073
IBM	resident commissioner	-.116	Google	Beauty	.060
IBM	sociologist	-.099	Clarifai	writer	.054
IBM	analyst	-.090	Clarifai	physician	.053
Clarifai	democracy	-.090	Clarifai	work	.051
IBM	call center	-.085	Clarifai	professional person	.050
IBM	diplomat	-.085	Clarifai	contractor	.050
Clarifai	musician	-.063	Clarifai	fine-looking	.048
Clarifai	singer	-.046	Clarifai	skillful	.044
Clarifai	cheerful	-.044	Clarifai	pretty	.039
Clarifai	happiness	-.034	IBM	entertainer	.034
Clarifai	music	-.033			
Clarifai	confidence	-.032			

sensitive variables on the model outcomes. Using this methodology, we were able to identify hidden biases of commercial computer vision APIs on gender and race. These biases, likely caused by the skewed representation in online media, should be adequately addressed in order to make these services more reliable and trustworthy.

REFERENCES

[1] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2019. Machine Bias. *ProPublica* (Mar 2019).

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* 25, 6 (2019), 954.

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*. 2745–2754.

[4] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36, 1 (2009), 21–40.

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[6] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting Bias with Generative Counterfactual Face Attribute Augmentation. *arXiv preprint arXiv:1906.06439* (2019).

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[8] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[9] Drew Harwell. 2019. Oregon Became A Testing Ground For Amazon's Facial-Recognition Policing. But What If Rekognition Gets It Wrong. *Washington Post* (2019).

[10] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.

[11] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.

[12] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913* (2019).

[13] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[14] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 313–322.

[15] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*. 5967–5976.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[17] Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit Bias Discovery in Visual Question Answering Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9562–9571.

[18] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. 2019. Characterizing Bias in Classifiers using Generative Models. *arXiv preprint arXiv:1906.11891* (2019).

[19] Baback Moghaddam and Ming-Hsuan Yang. 2002. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (2002), 707–711.

[20] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. 2018. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099* (2018).

[21] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[22] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology* (2019).

[23] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2019. Understanding the Political Ideology of Legislators from Social Media Images. *arXiv preprint arXiv:1907.09594* (2019).

[24] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*. Springer, 776–791.

[25] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 325–341.

[26] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989.

[27] Ran Zmigrod, Sebastian J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *arXiv preprint arXiv:1906.04571* (2019).