IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 2, 2021, accepted December 16, 2021, date of publication December 20, 2021,
date of current version December 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3136898

Implicit Stereotypes in Pre-Trained Classifiers

NASSIM DEHOUCHE
Mahidol University International College, Salaya 73170, Thailand
e-mail: nassim.deh@mahidol.edu

ABSTRACT Pre-trained deep learning models underpin many public-facing applications, and their propensity to reproduce implicit racial and gender stereotypes is an increasing source of concern. The risk of large-scale, unfair outcomes resulting from their use thus raises the need for technical tools to test and audit these systems. In this work, a dataset of 10, 000 portrait photographs was generated and classified, using CLIP (Contrastive Language–Image Pretraining), according to six pairs of opposing labels describing a subject's gender, ethnicity, attractiveness, friendliness, wealth, and intelligence. Label correlation was analyzed and significant associations, corresponding to common implicit stereotypes in culture and society, were found at the 99% significance level. A strong positive correlation was notably found between labels *Female* and *Attractive*, *Male* and *Rich*, as well as *White Person* and *Attractive*. These results are used to highlight the risk of more innocuous labels being used as partial euphemisms for protected attributes. Moreover, some limitations of common definitions of algorithmic fairness as they apply to general-purpose, pre-trained systems are analyzed, and the idea of controlling for bias at the point of deployment of these systems rather than during data collection and training is put forward as a possible circumvention.

INDEX TERMS Algorithmic fairness, facial recognition, deep learning, zero-shot classification, CLIP.

DATA AND CODE AVAILABILITY STATEMENT

The dataset used in this work was made publicly available via the DOI of reference [1]. The code is detailed in the Appendices of this paper.

I. INTRODUCTION

Numerous artificial intelligence applications in business and government [2] rely on deep learning. This algorithmic approach has rapidly set the state-of-the-art in challenging problems such as computer vision [3] and natural language processing [4]. However, deep learning typically requires large, manually annotated datasets that are highly cost and labor-intensive to produce. Moreover, applications are limited to the narrow, specific tasks defined by this labelling. For instance, *ImageNet*, one of the largest existing benchmark datasets for computer vision [5], required more than 25,000 workers [6] to label 14,197,122 images. These practical difficulties explain the popularity of pre-trained models. These are deep learning models that were trained on large benchmark datasets [7] and are commonly reused for similar problems. Moreover, advances in the last couple of years have made the manual labelling of data, particularly image-text pairing, plainly unnecessary. Indeed, January 2021 has seen the introduction of CLIP (Contrastive Language–Image Pretraining), a general-purpose image-text model which was able to learn from 400 million of text–image pairs collected from the internet, and achieved state-of-the-art performance on benchmark datasets it was not explicitly trained for [8]. This pre-trained model is thus capable of performing image classification out-of-the-box, on virtually any user-provided labels.

Though no commercial application of CLIP has been announced yet, OpenAI, the private research firm responsible for its development, reports that GPT-3 [9], a natural language processing model with comparable zero-shot performance it released in May 2020, is now used in more than 300 large-scale commercial applications. The authors of GPT-3 foresee several potentially harmful uses of their pre-trained language system (misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting) and state that the ability of their software represents a ''concerning milestone''. The authors of CLIP also note that its vast range of capabilities, many of which are made clear only after testing for them, raise ethical challenges similar to those found in GPT-3 [8], [9].

This paper argues that the performance, generality, opacity, and potential scale of distribution of CLIP additionally create a risk of reproducing and amplifying implicit gender and

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

racial stereotypes present in society and culture. We show that this pre-trained system, trained on the whole English corpus, not only has learned implicit stereotypes pertaining to race and gender, but renders many common definitions of fairness particularly challenging to apply.

Following this introductory Section, Section II reviews notable findings and concepts from the extent literature that are of relevance to this work; Section III describes the proposed dataset and the model used for its classification; Section IV presents the results of a correlation analysis of this classification and discusses its implications. Finally, Section V concludes this paper with perspectives for further research on algorithmic fairness in pre-trained systems such as CLIP.

II. RELATED WORK

Alongside the development of machine learning models, there has been a growing body of research on their ethical implications, particularly in relation to fairness, recent reviews of which can be found in [10] and [11].

For instance, based on an evaluation of 219 natural language processing systems, on a corpus of 8, 640 sentences specifically chosen to tease out gender and race biases in natural language processing systems, [12] found that more than 75% of the systems tend to mark sentences involving one gender/race with higher intensity scores than the sentences involving the other gender/race. Bolukbasi *et al.* [13] demonstrated that the vector representations of words in these model contain biases in their geometry that reflect gender stereotypes present in broader society. Moreover, due to their wide-spread usage as basic features, language models not only reflect such stereotypes but can also amplify them. Debiasing algorithms [13], i.e the identification of pairs/sets of words over which to neutralize or attenuate bias is a common strategy to enforce fairness. This approach can, for instance, force gender-neutral words such as ''babysit'' to be equally distant from gendered words such ''grandmother'' and grandfather''. However, large, pre-trained systems such as CLIP pose unprecedented challenges in this regard due to the very large volume of words and associations it has learned.

Moreover, [14] argue that a certain focus on dataset bias in critical investigations of machine vision paints an incomplete picture. Indeed, as noted by Hinton [15], implicit stereotypes learned by an individual do not reflect cognitive biases but associations prevalent within their culture. The author recommends examining these association, as they are communicated within culture, rather than considering them as cognitive defects. The same reasoning can be extended to deep learning systems. Although a relatively nascent approach, Tsamados *et al.* [11] see the use of technical/statistical tools to test and audit algorithmic systems and decision-making as a potentially promising way to enforce fairness. In this line of thinking, three statistical definitions of fairness for machine-learning-based decision-making, (*anti-classification*, *statistical parity*, and *calibration*), have been described and analyzed by [16] and [17]. If *calibration* (i.e. the proportion of individuals re-offending remaining uniform across protected groups) is mainly geared towards judiciary applications and is outside the more general scope of this work. we consider the first two of these definitions.

Let $x_i \in \mathbb{R}^p$ be a vector of features which can be partitioned into protected (e.g. ethnicity, gender) and unprotected attributes, i.e. $x_i = (x_i^p, x_i^u)$, describing an individual i, and $d : \mathbb{R}^p \rightarrow \{0, 1\}$, a decision rule, with 0 and 1 respectively representing a negative and positive outcome.

1) *Anticlassification*, also known as *unawareness*, requires that decisions be independent of protected attributes. Formally, a decision rule satisfies anticlassification, if and only if it verifies:

$$d(x_i) = d(x_j), \quad \forall i, j, \quad \text{such that } x_i^u = x_j^u.$$

A commonly noted limitation of this concept of fairness is that correlated unprotected attributes (education level, salary, life-expectancy, etc.) might act as proxies for protected attributes, effectively resulting in indirect discrimination [18]. Ethnicity may, for instance, be correlated with education levels [17] and ZIP codes because of the demographic makeup of residential areas [10].

2) *Demographic parity*, a form of *classification parity*, requires that positive decisions be equally likely for all demographic groups defined by protected attributes. Formally, a decision rule satisfies demographic parity, if and only if it verifies:

$$P(d(x_i) = 1|x_i^p) = P(d(x_i) = 1), \quad \forall i$$

As noted in [16], demographic parity is closely related to anti-classification, and the same reservations pertaining to the use of proxies for protected attributes can be extended to it.

Common precautions concerning the use of non-protected attributes as proxies concern attributes that effectively show such a correlation in real distributions. We show that CLIP has inherited implicit stereotypes present in culture, resulting in significant associations between protected attributes and superficial attributes in generated photographs (e.g. attractiveness), as well as plainly inexistent attributes (e.g. the wealth of the persons represented). Therefore, seemingly unrelated attributes such as attractiveness can effectively serve as proxies for attributes of gender and ethnicity.

III. DATA AND METHODS

A. IMAGE GENERATION

Generative Adversarial Networks (GAN) currently constitute the state-of-the-art for image generation [19]. This class of methods relies on the opposition of two neural networks; a generative network generates candidate solutions by mapping learning data to a latent space of all possible outputs, while a discriminative network attempts to distinguish this output from the true data distribution. *thispersondoesnotexist.com* is a software application written by Phillip Wang, for the generation of fake portrait photographs [20]. *StyleGAN2*, the GAN underlying *thispersondoesnotexist.com* is based on the work of Karras *et al.* [21] at NVIDIA Research. The fake, yet convincing photographs generated by this GAN have raised academic concern pertaining to their use for political deception and identity scams [22], and have been used to gauge the accuracy of methods intending to distinguish them from photographs of real persons [23]. However, these images being mere vectors of coordinates, randomly sampled from the latent space of all possible representations of a person in *StyleGAN*2, they also offer an interesting avenue for the analysis of implicit stereotypes in classifiers. As opposed to photographs of real persons, one can be certain that no real distribution of attributes exists prior to their labeling. Therefore, any reference to, say, the inexistent socioeconomic status or intelligence of the persons they represent, would solely result from the classifier rather than possibly meaningful hidden information in the data. Moreover, fake photographs allow for fruitful illustrations without revealing any personally identifiable information of real persons.

Thus, in this work, we have generated a dataset of 10, 000 portrait photographs through random API calls (Cf. Appendix A) to *thispersondoesnotexist.com*, in order to study implicit stereotypes in the form of associations between the labels ascribed by probabilistic classifiers to the persons it represents. The 10, 000 photographs used in this work were made publicly available [1].

B. ATTRIBUTE DEFINITION

We are interested in studying the association of two protected attributes, gender and ethnicity, with conventionally desirable attributes that are attractiveness, friendliness, wealth and intelligence. We define each attribute by a pair of opposing labels, representing a spectrum of belonging probabilities for each image. The pairs of labels considered are the following:

- (*Male*, *Female*),
- (*White Person*, *Person of Color*),
- (*Attractive*, *Unattractive*),
- (*Friendly*, *Unfriendly*),
- (*Rich*, *Poor*),
- (*Intelligent*, *Unintelligent*).

Variations of these attributes could, for instance, be part of an automated screening system for job or credit applicants. The script presented in Appendix B describes this model in the Python language. In addition to the two protected attributes of gender and ethnicity, attractiveness and friendliness can be decently inferred from photographs of even fake persons to verify the quality of the classification, while wealth and intelligence have been precisely chosen because they do not have any real basis and would thus constitute an interesting way to explore implicit stereotypes in classifiers through their association with "visible" attributes, notably protected ones.

C. IMAGE CLASSIFICATION

Given an image as input, probabilistic classifiers [28] compute a probability distribution over a set of classes. These probabilities can be understood as the likelihood of the image belonging to each class. A discrimination threshold of 50% is commonly used [29] to infer a crisp classification from these probabilities.

A successful approach for this type of classification consists in learning a shared embedding space for images and texts, known as a visual-semantic embedding space [30], in which images and their corresponding textual descriptions would be adjacent. Belonging probabilities can then be calculated based on the proximity of images and texts. A key feature of CLIP [8] is that it learned this joint, visual-semantic embedding space from unprecedentedly large datasets of internet data that did not require manual labelling. Given a new image-text pair to classify, the image and text are first encoded separately into two vectors of different dimensions. The projection layer of CLIP then projects these vectors on the aforementioned joint embedding space, resulting in text and image embeddings of a similar dimension. It is therefore possible to measure their similarity, as we would do for two vectors of text, using cosine similarity. This methodology is graphically outlined in Figure 1, with a two-dimensional representation of the visual-semantic embedding space for the sake of simplicity.

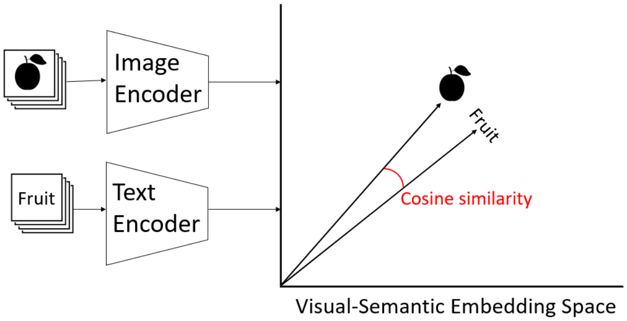


FIGURE 1. **Schematic view of the zero-shot classification methodology of CLIP.**

Lastly, the *softmax* function, which is commonly used as the last activation function of neural networks [31], is used to normalize this output to a probability distribution over predicted output classes. We have used the *Transformers* implementation of CLIP [32]. In this implementation, the default dimensions for both image and text embeddings consisting in vectors of 512 elements, as detailed in Appendices B and C.

For the set of attributes considered in this work, the output for each pair of labels, e.g. *Friendly*/*Unfriendly*, is of the form $p/1-p, p \in [0, 1]$. Each label can therefore be considered as a random variable encoding all information about the corresponding attribute. The belonging probability of an image to class *Unfriendly* is the complement of the belonging probability to *Friendly*, and can be deduced from it. Moreover, as a consequence of the bilinearity of covariance, the correlation of, say *Unfriendly* and *Unattractive* would be equal to that of *Friendly* and *Attractive* and equal in opposite sign to the correlation of *Friendly* and *Unattractive*. Therefore, labels of one polarity can be omitted, in each pair, without loss of generality.

We have used CLIP to classify the 10, 000 images in the generated dataset according to the six pairs of labels previously defined, used literally and without additional prompts, as detailed in Appendix B. The results of this probabilistic classification are provided in a comma-separated values (CSV) file [1]. For each pair of labels, CLIP determined complementary belonging probabilities of individual images to both labels in the pair.

A random sample of 1000 images was manually verified in terms of its crisp classification on protected attributes (gender and ethnicity), with a discrimination threshold of 50%. The classification produced by CLIP was found to perfectly correspond to manual classification.

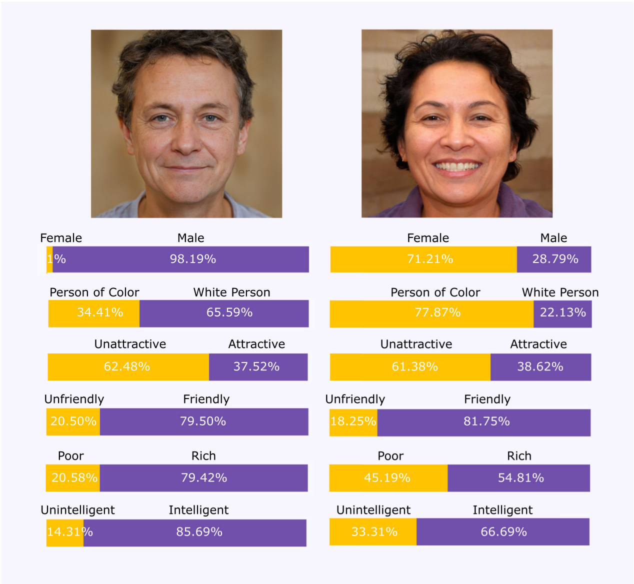


FIGURE 2. Classification of two images.

Figure 2 illustrates this classification with two images from the dataset (files 1054.jpg and 1055.jpg) and exemplifies some of the stereotypes analyzed in this paper. These two images are adjacent in the latent space and show two persons with relatively similar facial structures against a similarly neutral background. CLIP shows a reasonably accurate classification on the two protected attributes, as well as the two visible attributes, though the person on the right is seen as slightly more attractive, a point which will be explored more formally for the whole dataset in Section 4, and slightly friendlier, as a likely consequence of the more pronounced smile exhibited in the image. However, the person on the left of Figure 2, classified as Male and White, with the common discrimination threshold of 50%, is deemed 24.61% more likely to be rich and 19.00% more likely to be intelligent, without there being any features in the photographs (e.g. expensive accessories or clothing items) that could suggest such a discrepancy of socio-economic status and intelligence. The goal of the present analysis would be to formally investigate these associations and the proportion of the variance of visible and hidden attributes (wealth and intelligence) that can be explained by protected attributes, based on implicit stereotypes.

IV. RESULTS

A. DESCRIPTIVE STATISTICS

Table 1 presents the proportion of images assigned to each class, with a discrimination threshold of 50%, while Table 2 presents descriptive statistics for the distributions of belonging probabilities to labels *White Person*, *Attractive*, *Rich*, *Unintelligent*, *Male*, *Friendly* (the opposite labels are omitted without loss of generality, as explained in Section III). Table 2 additionally includes the individual images that achieve minima and maxima on each label, as well as their reference in the dataset. The proportions of images in each class are determined using a discrimination threshold of 50%. The shapes of the distributions of belonging probabilities are illustrated by histograms (a) to (f) in Figure 3. Overall, the two protected attributes show more clearly bimodal distributions, while the classes *Rich* and *Attractive* are bell-shaped and classes *Intelligent* and *Friendly* are right-skewed, a possible consequence of the fact that the learning data are mainly comprised of smiling portrait photographs.

Standard deviation inversely measures the concentration of values around their mean. In the context of binary probabilistic classification, the standard deviation of class belonging probabilities reflects the certainty of the classifier in assigning labels. Classification certainty would be minimum if all objects have a 50% belonging probability (i.e. no object can be assigned to either class), resulting in a Dirac distribution of belonging probabilities with a standard deviation of zero. Certainty in classification would be maximum if objects are assigned a belonging probability of either 100% or 0% (the latter meaning a belonging probability of 100% to the opposite class). In this case, standard deviation would be at its highest with 50%, for perfectly balanced classes. In between these two extremes, a Gaussian distribution of belonging probabilities centered in 50% would exhibit a standard deviation of 12.5%. Because of their visible and fairly objective nature, the two protected attributes, unsurprisingly, show the highest standard deviations and thus certainty in classification. Gender (standard deviation of 35.09%) is determined with more certainty than ethnicity (24.31%). However, although the proportion of images classified in the two genders considered in this study are roughly equal (50.74% male and 49.26% female), it can be seen in Figure 3a that assignment to the class *Male* is performed with more certainty than to the class *Female*. In other words, a female subject is more likely to be misclassified as male, than a male subject to be misclassified as female. This asymmetry has been previously observed in the literature. Classifiers tend to gender white men correctly, while they are most likely to misgender women [33], particularly darker skinned ones [34]. We thus find this asymmetry reflected in the histogram of gender belonging probabilities in Figure 3a, and observe a mixture of distributions, with a bell-shaped distribution of belonging probabilities in the class *Female* and a skewed distribution in the class *Male*. Moreover, there is a 63.8% proportion of white persons in the generated images. This proportion is not representative of the global population and seems rather roughly representative of the population of the USA, with a percentage of "Non-Hispanic Whites" of 60.7% [24]. It justifies not further breaking-down the *Person of Color* class, in order to maintain relative balance between the two classes. Though there exist datasets of more representative real images, such as *FairFace* [25], a dataset of Flickr images covering seven ethnicities, they would reproduce wealth and other bias that are present in society. As of the time of this writing, generating faces that are ethnically representative is still an open problem, with promising recent (May 2021) contributions [26]. The fact that these six random variables represent normalized scores and their thin-tailed distributions justify the use of Pearson's correlation coefficient as a metric for the assessment of their associations.

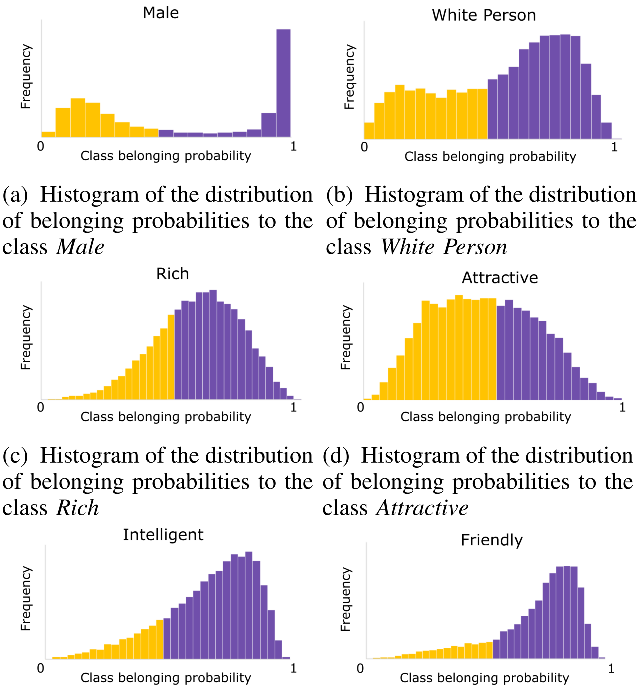
B. LABEL CORRELATION

We have computed pairwise correlation coefficient for the six labels considered. Results are presented in Figure 4, while Table 3 and Table 4 respectively present the corresponding t-statistics and p-values [27], with hypothesis H_0: "The correlation is not significant". For a significance threshold of 0.05, H_0 cannot be rejected for the association between gender and ethnicity, i.e. between labels *Male* and *White Person* (as well as the three other combinations of *Male*/*Female* and *White Person*/*Person of Color* as explained in Section III-C). For a significance threshold of 0.01, the association between ethnicity and intelligence does not appear significant, a pleasantly surprising result, given the long history of scientific racism [39] that CLIP could have learned from. Though, it is possible that this result is an artefact of the binary definition of ethnicity used in this study, and more detailed ethnicity labels might have uncovered other associations between intelligence and ethnicity.

All remaining pairs of attribute show significant correlations for this significance threshold. A high negative correlation of -0.52 has been found between labels *Male* and *Attractive* (and therefore a correlation of $+0.52$ between *Female* and *Attractive*). Another strong correlation is found between labels *Male* and *Rich* at $+0.4$, meaning that 20% of the variance of the latter, "inexistent" attribute is explained by gender alone. Lastly, a correlation of $+0.33$ was found between labels *White* and *Attractive*, which means that about 10% of the variance in attractiveness is explained by whiteness alone. In addition to the 25% of variance explained by the uncorrelated dimension of gender, these results suggest that about 35% of attractiveness, according to CLIP, can be explained by the extent to which a person is white and female.

TABLE 1. Class proportions for the 10, 000 images in the dataset, for a discrimination threshold of 50%.

Labels	Proportions
Male - Female	50.74% - 49.26%
White Person - Person of Color	63.80% - 36.20%
Rich - Poor	76.46% - 23.54%
Attractive - Unattractive	42.48% - 57.52%
Intelligent - Unintelligent	84.77% - 15.23%
Friendly - Unfriendly	86.68% - 13.32%



(a) Histogram of the distribution of belonging probabilities to the class *Male*

(b) Histogram of the distribution of belonging probabilities to the class *White Person*

(c) Histogram of the distribution of belonging probabilities to the class *Rich*

(d) Histogram of the distribution of belonging probabilities to the class *Attractive*

(e) Histogram of the distribution of belonging probabilities to the class *Unintelligent*

(f) Histogram of the distribution of belonging probabilities to the class *Friendly*

FIGURE 3. Histograms of the distributions of belonging probabilities.

C. FINE-GRAINED ANALYSIS

To further elucidate these three remarkable pairwise associations and rule out Simpson's paradox, data were pooled according to the crisp classification of protected attributes (with a discrimination threshold of 50%). No reversal in the sign of the corresponding coefficients of correlation when measured for each pool was found. However, some significant discrepancies in their absolute values have been found, as detailed in this section.

Thus, for the labels *White Person* and *Attractive*, data were pooled according to the former label, using the same classification threshold of 50%, as represented in Figure 5. The correlation between ethnicity and attractiveness was separately measured for individuals classified as *White Person* (in purple) and *Persons of Color* (in yellow). The least square trend lines for each pool have been included to

TABLE 2. Descriptive statistics for the probabilistic classifications of 10, 000 images with CLIP.

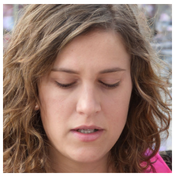
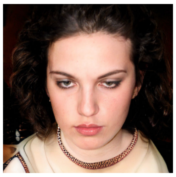
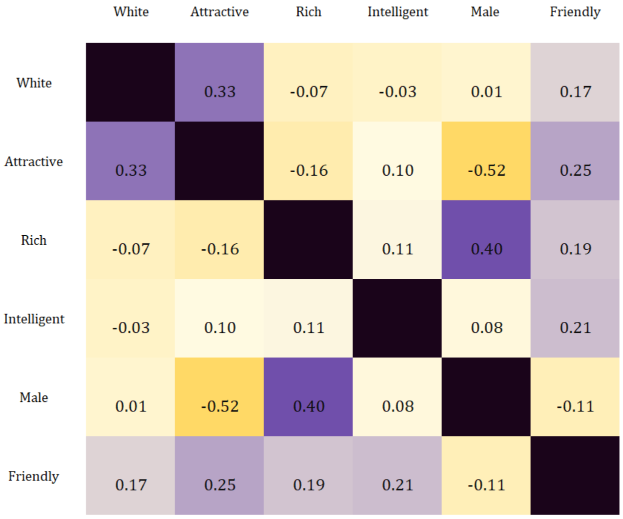
	White Person	Attractive	Rich	Intelligent	Male	Friendly
Mean	56.49%	46.60%	60.71%	68.81%	58.26%	70.68%
Standard Deviation	24.31%	17.37%	15.10%	17.24%	35.09%	17.23%
Skewness	-0.4457	0.1208	-0.3575	-0.8240	0.0053	-1.2398
Minimum	3.47% 1758.jpg	5.92% 5860.jpg	4.93%  2216.jpg	6.85% 724.jpg	4.42% 5954.jpg	2.88%  5009.jpg
Maximum	98.32% 4687.jpg	92.66% 3425.jpg	95.85% 1840.jpg	97.65% 348.jpg	99.40% 7613.jpg	97.34% 1756.jpg
Class proportion	63.80%	42.48%	76.46%	84.77%	50.74%	86.68%

TABLE 3. t-statistics for pairwise correlation coefficients.

	White	Attractive	Rich	Intelligent	Male	Friendly
White	-	35.53	-6.87	-2.88	0.94	16.95
Attractive	35.53	-	-16.02	9.76	-61.14	25.72
Rich	-6.87	-16.02	-	10.60	43.27	19.62
Intelligent	-2.88	9.76	10.60	-	7.54	20.98
Male	0.94	-61.14	43.27	7.54	-	-10.96
Friendly	16.95	25.72	19.62	20.98	-10.96	-



	White	Attractive	Rich	Intelligent	Male	Friendly
White		0.33	-0.07	-0.03	0.01	0.17
Attractive	0.33		-0.16	0.10	-0.52	0.25
Rich	-0.07	-0.16		0.11	0.40	0.19
Intelligent	-0.03	0.10	0.11		0.08	0.21
Male	0.01	-0.52	0.40	0.08		-0.11
Friendly	0.17	0.25	0.19	0.21	-0.11	

FIGURE 4. Label correlation table.

aid visualization. Correlation between ethnicity and attractiveness was found to be positive and approximately equal for individuals classified into *White Person* and *Person of Color*.

In Figure 6, The correlation between gender and attractiveness was separately measured for individuals classified as *Male* (in purple) and *Female* (in yellow). Correlations for each of the two gender classes are positive, though there exists a higher correlation between gender belonging probabilities and attractiveness for individuals classified as *Female* (+0.3003 vs +0.1204 for those classified as *Male*), indicating that the variations of the belonging probabilities of these individuals explain a higher proportion of the variance in the belonging probabilities to the label *Attractive*. The same analysis was performed for gender and the label *Rich*, in Figure 7, using the same color coding for the crisp classification *Male* and *Female*, though the axis was reversed to focus on positive correlations. Here, there is an even greater gap in correlation for each cohort. Indeed, the correlation between gender and wealth is non-significant for individuals classified as *Female*, indicating that the degree of femaleness does not explain any of the variance in belonging probabilities to the class *Rich*.

TABLE 4. p-values for pairwise correlation coefficients.

	White	Attractive	Rich	Intelligent	Male	Friendly
White	-	$1.41 \cdot 10^{-260}$	$6.76 \cdot 10^{-12}$	0.003873	0.345209	$1.31 \cdot 10^{-63}$
Attractive	$1.41 \cdot 10^{-260}$	-	$4.32 \cdot 10^{-57}$	$2.07 \cdot 10^{-22}$	0	$2.47 \cdot 10^{-14}$
Rich	$6.76 \cdot 10^{-12}$	$4.32 \cdot 10^{-57}$	-	$4.09 \cdot 10^{-26}$	0	$3.65 \cdot 10^{-84}$
Intelligent	0.003873	$2.07 \cdot 10^{-22}$	$4.09 \cdot 10^{-26}$	-	$4.99 \cdot 10^{-14}$	$1.12 \cdot 10^{-95}$
Male	0.345209	0	0	$4.99 \cdot 10^{-14}$	-	$7.69 \cdot 10^{-28}$
Friendly	$1.31 \cdot 10^{-63}$	$2.47 \cdot 10^{-141}$	$3.65 \cdot 10^{-84}$	$1.12 \cdot 10^{-95}$	$7.69 \cdot 10^{-28}$	-

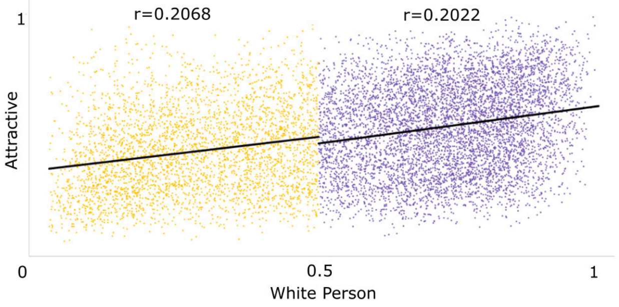


FIGURE 5. Scatter diagram of belonging probabilities for labels *White Person* and *Attractive*.

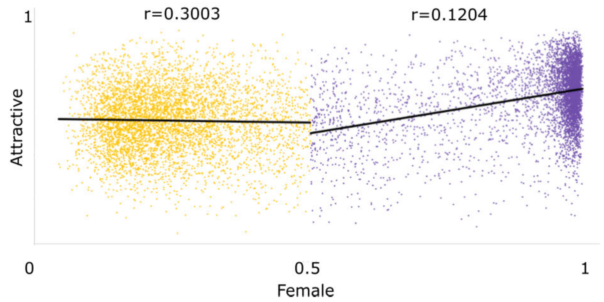


FIGURE 6. Scatter diagram of belonging probabilities for labels *Female* and *Attractive*.

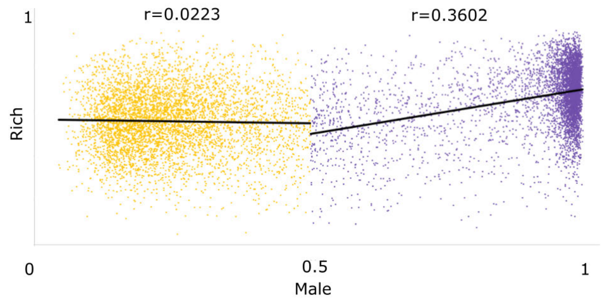


FIGURE 7. Scatter diagram of belonging probabilities for labels *Male* and *Rich*.

However, past the 50% threshold, i.e. for individuals classified as *Male*, a significant positive correlation of +0.3602 is found. Past, the 50% threshold, the more likely one is to be *Male*, the more likely they would be to be classified as *Rich*. It should be noted that the previous associations with gender exist within each gender group (i.e. images classified as *Male* and *Female*). They are, therefore, unaffected by the mixed nature of the distribution of gender belonging probabilities highlighted in the histogram in Figure 3a.

More so than other facial attributes (e.g. ethnicity), the latent space of gender attributes in GANs is well-studied [38]. These previous associations with gender can be illustrated, more significantly, using so-called *gender swap* procedures.

We have used a commercial implementation of the approach by Viazovetskyi *et al.* [40] with default parameters to swap the gender of the images and tested the effect of the gender swap on the CLIP classification according to the six pairs of labels considered. Though there is no univocal way to swap gender, and these methods are known to result in impurities, the result of the gender swap was manually inspected for 120 randomly-selected pairs and found to be accurate. The set of pairs of gender-swapped images as well as the classification results can be verified in [1] in the archive ‘‘Gender Swap.zip’’. Figure 8 shows the classification of an example of such a pair images (files 47.jpg and 47M.jpg).

We are interested in the average belonging probability to each class label, for the female and male version of each image and also in measuring bias in the distribution of the differences in the belonging probability to each label.

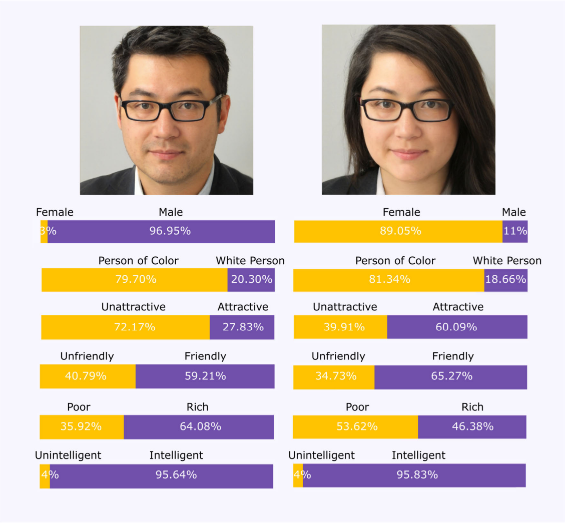


FIGURE 8. Classification of a pair of gender-swapped images.

TABLE 5. Mean belonging probabilities and Mean difference in the classification of gender-swapped pairs of images.

	White Person	Attractive	Friendly	Rich	Intelligent
Female versions	55.54%	54.64%	71.49%	47.93%	72.77%
Male versions	53.96%	42.62%	66.38%	67.70%	75.37%
Mean Difference	1.58%	12.02%	5.10%	−19.76%	−2.59%

Given a label and a pair of gender-swapped images, the mean difference (MD) or difference of means was used to estimate the average difference between the belonging probability of the female and the male version of an image. Formally, let $p_i^f(l)$ and $p_i^m(l)$ respectively be the belonging probability of the female and male version of the $i-th$ pair of images, to a class label l. The mean difference is given by $\mathrm{MD} = \sum_{i=1}^{n} \frac{p_i^f(l)}{n} - \sum_{i=1}^{n} \frac{p_i^m(l)}{n} = \sum_{i=1}^{n} \frac{(p_i^f(l) - p_i^m(l))}{n}$. Because actual rather than absolute values of differences are used, positive and negative differences can offset each other. Consequently, the MD can be used as a measure of bias. The results in Table 5 correspond to the previously identified associations. Notably, the female version of an image is, on average, deemed 19.76% less likely to be rich and 12.02% more likely to be attractive than its male counterpart.

D. IMPLICIT STEREOTYPES IN LANGUAGE AND CULTURE

Though theoretically anomalous, between labels that could reasonably be assumed to be independent (e.g. males and females, as a whole, being equally attractive to each other), the previous correlations are nonetheless representative of the society and culture that CLIP learned from. As observed by Hinton [15], implicit stereotype are not much of cognitive bias than associations prevalent within culture. Indeed, European standards of beauty are notoriously preeminent, including among people of non-European ancestry [41], [42], and according to Silvestrini [43], this tendency is still perpetuated by media. Moreover, there exists a widely-studied cultural propensity to romanticize women as objects of heterosexual affection (including self-objectification) and they tend to be judged on their appearance more than men [35]–[37]. As for gender economic inequality, it is both an implicit cultural stereotype and a global economic reality [44]–[46].

A prosaic yet meaningful illustration of these gender stereotypes in culture, as they relate to natural language processing, can be seen in the prevalence of corresponding label associations in the English language. The frequencies of usage of eight combinations of labels pertaining to gender with attractiveness and wealth in the 2019 English corpus language of the Google Books Ngram Viewer, a search engine that measures the annual frequencies of any set of strings in publications printed between 1500 and 2019, were compared. The prevalence of the terms ''Rich Man'', ''Rich Woman'', ''Poor Man'', and ''Poor Woman'', as well as ''Attractive Man'', ''Attractive Woman'', ''Unattractive Man'', and ''Unattractive Woman'' were compared. Results are presented in Figure 9 concerning attractiveness, and in Figure 10 for wealth. Throughout history, each of the terms *Rich* and *Poor* have been used significantly more to qualify men than women, whereas each of *Attractive* and *Unattractive* have been used more to qualify women than men, in a form of benevolent sexism [47].

E. DEBIASING CLIP

This section applies an influential debiasing algorithm introduced by Bolukbasi *et al.* [13] to CLIP. For the sake of simplicity of the presentation, we focus on the previously identified gender stereotypes, namely the association of gender with wealth and attractiveness. However, this approach could be similarly extended to the stereotypes associated with ethnicity. As explained in Section III-C, a feature of CLIP and similar models is the computation of a joint embedding space for images and text (512-elements vector embeddings in this application). This allows us to naturally extend the aforementioned algorithm, devised for gender bias in text alone, to CLIP. Debiasing first requires geometrically identifying a gender subspace in text and image embeddings and quantifying its contribution to similarity scores. This gender subspace will later serve as a basis to the neutralization of bias. Let $g \in \mathbb{R}^{512}$ be this subspace. Similarly to [13], we define g by aggregating the differences in the embeddings of a set of ten pairs of words denoting gender. These are:

- (*Male*, *Female*),
- (*Man*, *Woman*),
- (*Boy*, *Girl*),
- (*He*, *She*),
- (*His*, *Her*),
- (*Himself*, *Herself*),
- (*Father*, *Mother*),
- (*Son*, *Daughter*),
- (*Guy*, *Gal*),
- (*John*, *Mary*).

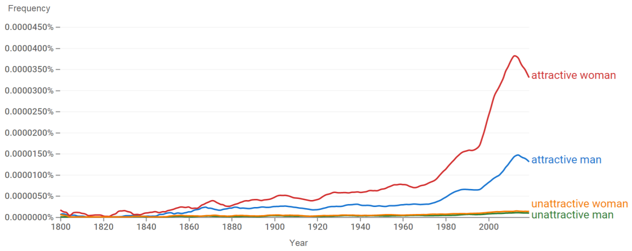


FIGURE 9. Case-insensitive incidence of the combinations of terms attractive/unattractive man/woman in the English (2019) corpus of Google Books Ngram Viewer.

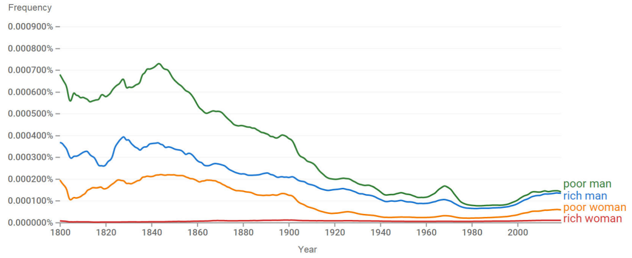


FIGURE 10. Case-insensitive incidence of the combinations of the terms rich/poor man/woman in the English (2019) corpus of Google Books Ngram Viewer.

The gender subspace g is defined by a linear aggregation of the differences in the embeddings of words denoting two polarities of gender. Formally, for a pair of words, such as *She*, let $\overrightarrow{She}$ denote its embedding. The gender subspace is defined by the directions explaining most of the variance in the difference vectors, such as $\overrightarrow{He} - \overrightarrow{She}$. To this end, the principal components (PC) of the ten difference vectors were computed. Figure 11 shows the proportion of variance in the normalized vector differences that is explained by the first ten principal components (PC1 to PC10). The code used for this computation can be found in reference *#Identifying the gender subspace* of Appendix C.

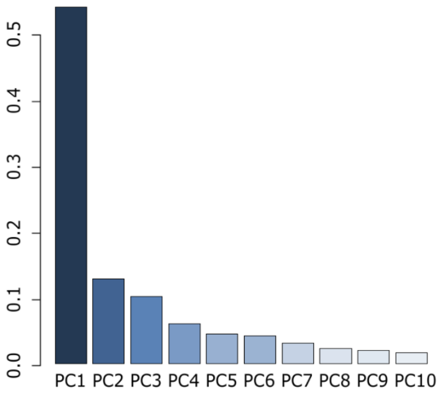


FIGURE 11. Percentage of variance explained by the first ten principal components of the normalized vector differences.

As previously observed by Bolukbasi *et al.*, the first principal component explains a disproportionate percentage of variance (54.57%) and can be considered to largely capture the geometry of gender. Therefore, we consider the linear combination of the ten difference vector defining the first principal component to be the gender direction g. This gender direction is used to quantify direct bias (i.e. the contribution of gender to the embedding of a supposedly gender-neutral class label) and indirect bias (i.e. the contribution of gender to the cosine similarity of images and labels).

Considering that the class labels $N = \{Rich, Poor, Attractive, Unattractive\}$ are supposed to be neutral with regard to gender, we can measure direct bias in their embeddings and indirect bias in their association with other words and images.

Given a neutral word $w \in N$, the direct gender bias in its embedding $\overrightarrow{w}$ is defined by [13] as:

$$Direct\ Bias_c = \frac{1}{|N|} \cdot \sum_{w \in N} |cos(\overrightarrow{w}, g)|^c$$

where c is a parameter that determines the strictness of the measurement of bias. Like, Bolukbasi et al, we set $c = 1$, for the sake of simplicity. The direct bias in the embeddings of the four class labels N was found to be $DirectBias_1 = 0.2719$, which confirms the very significant contribution of the gender direction to the embeddings of these supposedly gender-neutral words. For comparison, the set of 327 occupations considered in [13] were found to have $DirectBias_1 = 0.08$. The code used for this computation can be found in reference *#Direct bias in word embeddings* of Appendix C.

Thus, the gender subspace g also allows us to quantify the contribution of gender to the similarities between pairs of words and words and images. Formally, a vector embedding $\overrightarrow{e} \in \mathbb{R}^{512}$ (be it that of a word or an image in the CLIP model), can be decomposed into $\overrightarrow{e} = \overrightarrow{e_g} + \overrightarrow{e_\perp}$, where $\overrightarrow{e_g} = (\overrightarrow{e} \cdot g)g$ is the contribution from gender, and $\overrightarrow{e_\perp} = \overrightarrow{e} - \overrightarrow{e_g}$ is the part of the embedding that is independent from gender. Therefore, indirect bias, i.e. the contribution of gender to the similarity between two embeddings $\overrightarrow{e}$ and $\overrightarrow{f}$, denoted β, can be computed as follows[1]:

$$\beta(\overrightarrow{e}, \overrightarrow{f}) = \left(\overrightarrow{e} \cdot \overrightarrow{f} - \frac{\overrightarrow{e_\perp} \cdot \overrightarrow{f_\perp}}{||\overrightarrow{e}||_2 ||\overrightarrow{f}||_2}\right) \Big/ \overrightarrow{e_\perp} \cdot \overrightarrow{f_\perp}$$

For the 10, 000 images classified, we find an average β of 0.1168 for labels (*Attractive*, *Unattractive*) and an average β of 0.1303 for labels (*Rich*, *Poor*). Interestingly, indirect bias shows an almost zero standard deviation across images (0.0075 and 0.0051, for the two pairs of labels respectively). This shows not only the high contribution of gender in the computation of these similarities but also how systematic this bias is. The code used for this computation can be found in reference *#Indirect bias in image-text similarity* of Appendix C.

The last step of the debiasing algorithm ensures that gender neutral words are zero in the gender subspace. As previously defined, these words are $N = \{Rich, Poor, Attractive, Unattractive\}$. This procedure, known as *Neutralize and Equalize* in [13], corrects bias while ensuring that the desirable properties of the original embeddings are preserved.

Each word $w \in N$ is re-embedded to:

$$\overrightarrow{w} = \frac{\overrightarrow{w_\perp}}{||\overrightarrow{w_\perp}||} = \frac{\overrightarrow{w} - \overrightarrow{w_g}}{||\overrightarrow{w} - \overrightarrow{w_g}||}$$

where g is the previously identified gender subspace. This methodology is graphically outlined in Figure 12, with a two-dimensional representation of the visual-semantic embedding space for the sake of simplicity.

TABLE 6. Results of the debiasing algorithm.

	$DirectBias_1$	Average β	Abs. correlation with gender
Attractiveness	0.0034	0.0261	0.11
Wealth	0.0034	0.0407	0.07

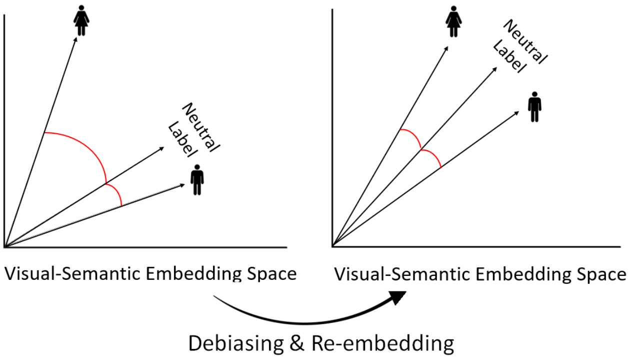

Neutral Label
Neutral Label
Visual-Semantic Embedding Space
Visual-Semantic Embedding Space
Debiasing & Re-embedding

FIGURE 12. Schematic view of the debiasing methodology.

The 10, 000 images were re-classified according to these new embeddings for the class labels in N. The effect of debiasing on direct and indirect bias, as well as the association of neutral labels with gender is presented in Table 6.

Direct bias and indirect bias are almost completely eliminated, and though not completely eliminated, the correlation of attractiveness and wealth with gender is reduced to a level close to that of intelligence or friendliness. An additional advantage of this procedure is that it does not affect the classification of images to labels outside N (e.g. *Male*, *Female*) and preserves the overall quality of the classification.

V. CONCLUSION AND PERSPECTIVES

Though it represents a remarkable advance in image classification, CLIP is prone to reproducing some implicit stereotypes present in the culture and society it learned from. This raises several issues in regard to conventional definitions of algorithmic fairness. CLIP trivially does not verify anticlassification and demographic parity, assuming that the conventionally desirable labels such as *Attractive* or *Rich* are positive outcomes. Moreover, the fact that CLIP and its

[1] These two embeddings can correspond to a pair of words such as *Male* and *Rich*, or an image and a word, such as when computing the belonging probability of an image to the class *Rich*.

underlying language model, are able to comprehend virtually any term of the English language, without additional training, makes the verification of the two concepts of fairness, for all possible terms directly pertaining to protected attributes for the astronomical number of possible labels and their associated stereotypes. This work focused on one set of labels (that could be used, for instance, out-of-the-box by a nightlife establishment for its door policy [48]) and found implicit stereotypes. Synonymous labels (e.g. *Affluent* instead of *Rich*, *Caucasian* for *White Person*) may have come with different stereotypes. As a perspective of future development, different synonyms of the labels considered in this work could be used to test their effect on the identified associations.

Further, associations with protected attributes, not only reproduce observed associations in real distributions (e.g. *Male* and *Rich*), but also reflect implicit stereotypes in culture and language (e.g. *Female* and *Attractive*, *White Person* and *Attractive*). Indeed, as this work has shown, such associations correspond neither to bias in data, nor algorithmic bias (i.e. bias that is not present in the input data and is added purely by the algorithm), as per the taxonomy of [10]. We see the distinction between these two types of correlations as important. If all implicit stereotypes are wrong, some of them are useful to model, so that they can precisely be corrected in real life. The existence of associations that reflect objective, unfair realities such as economic disparities between genders can be potentially valuable, depending on the intended usage of the system. For instance, you do want a classification system intended at identifying persons who are the most likely to benefit from financial aid to reproduce these associations. Other associations, such as *Female* and *Attractive* do not seem to have any basis, other than a certain benevolent sexism present in culture and language. Therefore, fairness is relative and depends on the intended application (i.e. what terms should have neutral associations with protected attributes). The responsibility of testing and ensuring fairness should thus be put at the point of deployment. For clearly defined neutral terms, this paper proposed an approach to quantify and attenuate bias reflecting implicit gender and racial stereotypes in CLIP. Fine-tuning with specifically-constructed additional training data (e.g. photographs of females labelled rich or males labelled attractive) risks reducing the accuracy of the classification system by over-fitting and reinforcing opposite stereotypes. The ability of CLIP's projection layer to abstractly represent images and text into a shared embedding space offers an elegant geometric solution, and permits the application of advances in debiasing made in the more mature field of Natural Language Processing. The debiasing approach used in this paper, adapted from [13] for bias in text models, proved promising for CLIP. It allowed for a geometric characterization of bias in the embedding space and its attenuation, while preserving the desirable properties of the classifier. This paper also highlighted the usefulness of GAN images, as opposed to real photographs, to audit classifiers and proposed a 10, 000 images dataset that can be used as a benchmark for such a task. However, this dataset, just like other existing GAN datasets suffers from a lack of representation for some groups (including the author's own North African ethnic group). This limited our study to binary labels for gender and ethnicity. Advances in Generative Adversarial Networks to improve their ability to generate more diverse ethnic and gender groups could help uncover other, more precise implicit racial and gender stereotypes in pre-trained classifiers.

APPENDIX A
PYTHON SCRIPT FOR IMAGE GENERATION

```
from PIL import Image
import tpdne
import base64
for k in range(10000):
 imagedata = base64.b64decode(tpdne.tpdne_base64())
 filename = str(k)+".jpg"
 with open(filename, ’wb’) as f:
 f.write(imagedata)
```

APPENDIX B
PYTHON SCRIPT FOR IMAGE CLASSIFICATION

```
import torch
from PIL import Image
import requests
import numpy
from transformers import CLIPProcessor, CLIPModel
import glob
import csv
model = CLIPModel.from_pretrained("openai/clip-vit-base-patch32")
processor = CLIPProcessor.from_pretrained("openai/clip-vit-base-patch32")
images=glob.glob("Images/*.jpg")
labels0 = [’white person’, ’person of color’]
labels1 = [’attractive’, ’unattractive’]
labels2 = [’rich’, ’poor’]
labels3 = [’friendly’, ’unfriendly’]
labels4 = [’intelligent’, ’unintelligent’]
labels5 = [’male’, ’female’]
for image in images:
 image=Image.open(image).resize((200, 200), Image.NEAREST)
 for i in range(6):
  globals()[’inputs
  return_tensors="pt", padding=True)
  globals()[’outputs
  #image-text similarity score
  globals()[’logits_per_image
  globals()[’probs
  globals()[’data
  with open("thisperson"+str(i)+".csv", "a") as fp:
   wr = csv.writer(fp, dialect=’excel’)
   wr.writerow(globals()[’data
```

APPENDIX C
PYTHON SCRIPT FOR DEBIASING

```
import torch
from PIL import Image
import requests
import numpy as np
import pandas as pd
import torch
from torch import nn
import torch.nn.functional as F
from transformers import CLIPProcessor, CLIPModel, CLIPTokenizer
import glob
import csv
images=glob.glob("Images/*.jpg")

#CLIP Tokenizer for embeddings
tokenizer = CLIPTokenizer.from_pretrained(’openai/clip-vit-base-patch32’)
model = CLIPModel.from_pretrained("openai/clip-vit-base-patch32")
processor = CLIPProcessor.from_pretrained("openai/clip-vit-base-patch32")
data = []

#Image embeddings, equal for all labels
ie = []
#Word embeddings, for neutral labels
we = []
for image in images:
 image=Image.open(image).resize((200, 200), Image.NEAREST)
 inputs0 = processor(text=["male", "female"], images=image, return_tensors="pt",
 padding=True)
 inputs1 = processor(text=["he", "she"], images=image, return_tensors="pt",
 padding=True)
 inputs2 = processor(text=["father", "mother"], images=image, return_tensors="pt",
 padding=True)
 inputs3 = processor(text=["son", "daughter"], images=image, return_tensors="pt",
 padding=True)
 inputs4 = processor(text=["John", "Mary"], images=image, return_tensors="pt",
 padding=True)
 inputs5 = processor(text=["man", "woman"], images=image, return_tensors="pt",
 padding=True)
 inputs6 = processor(text=["boy", "girl"], images=image, return_tensors="pt",
 padding=True)
 inputs7 = processor(text=["himself", "herself"], images=image, return_tensors="pt",
 padding=True)
 inputs8 = processor(text=["guy", "gal"], images=image, return_tensors="pt",
 padding=True)
 inputs9 = processor(text=["his", "her"], images=image, return_tensors="pt",
 padding=True)
 #Identifying the gender subspace
 for i in range(10):
  globals()[’outputs
  ie.append(outputs0.image_embeds.detach().numpy())
  #Text embeddings, for gendered labels
  globals()[’te
  #Differences in text embeddings of gendered labels
  globals()[’diff
  data.append(globals()[’diff
data = np.array(data)
ie = np.array(ie)
data=data / np.sqrt(np.sum(data**2))
df = pd.DataFrame(data=data)
df.to_csv(’out.csv’,index=False)
df = pd.DataFrame(data=data)
from sklearn.decomposition import PCA
pca=PCA(n_components=10)
pca.fit(df)
print(pca.explained_variance_ratio_)
g=pca.components_[0]
g=g / np.linalg.norm(g)

#Pairs of labels that are supposed to be gender-neutral for this application
pair0 = processor(text=["rich", "poor"], images=image, return_tensors="pt", padding=True)
pair1 = processor(text=["attractive", "unattractive"], images=image, return_tensors="pt",
padding=True)

#Direct bias in word embeddings
direct_bias=0
for i in range(2):
 for j in range(2):
  globals()[’out
  we=globals()[’out
  unit_vector = we / np.linalg.norm(we)
  direct_bias=direct_bias+abs(np.dot(g, unit_vector))
direct_bias=direct_bias/4
print(direct_bias)

beta = []
#Indirect bias in image-text similarity
for i in range(10000):
 for j in range(2):
  for k in range(2):
   globals()[’out
   te=globals()[’out
   te = te / np.linalg.norm(te)
   tg=np.dot(g, te)*g
   ie[i]=ie[i]/ np.linalg.norm(ie[i])
   ig=np.dot(g, ie[i][0])*g
   t_orthog=te-tg
   i_orthog=ie[i][0]-ig
   ratio=(np.dot(te,ie[i][0])-(np.dot(t_orthog,i_orthog)/(np.linalg.norm(t_orthog)
   *np.linalg.norm(i_orthog))))/np.dot(te,ie[i][0])
   beta.append(ratio)
beta = np.array(beta)
print(beta)
```

ACKNOWLEDGMENT

The author is grateful to Moein Shariatnia, for a helpful discussion on CLIP, as well as to two anonymous referees of IEEE Access, for their careful reading and constructive suggestions that improved the original draft of this article.

REFERENCES

[1] N. Dehouche, "This person does not exist CLIP classification," Mendeley Data, V4, to be published, doi: 10.17632/92jk3fchm2.4.

[2] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm to machine learning," *Arch. Comput. Methods Eng.*, vol. 27, no. 4, pp. 1071–1092, Sep. 2019, doi: 10.1007/s11831-019-09344-w.

[3] X. Jiang, A. Hadid, Y. Pang, E. Granger, and X. Feng, Eds., *Deep Learning in Object Detection and Recognition*. Singapore: Springer, 2018, doi: 10.1007/978-981-10-5152-4.

[4] D. Li and L. Yang, Eds., *Deep Learning in Natural Language Processing*. Singapore: Springer, 2018, doi: 10.1007/978-981-10-5209-5.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[6] S. Ings, "Impaired visions," *New Scientist*, vol. 244, no. 3251, pp. 30–31, 2019, doi: 10.1016/S0262-4079(19)31917-7.

[7] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2017, *arXiv:1605.07678v4*.

[8] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[9] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019, *arXiv:1908.09635*.

[11] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi, "The ethics of algorithms: Key problems and solutions," *AI Soc.*, to be published, doi: 10.1007/s00146-021-01154-8.

[12] S. Kiritchenko and S. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," in *Proc. 7th Joint Conf. Lexical Comput. Semantics*, 2018, pp. 43–53.

[13] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4349–4357.

[14] F. Offert and P. Bell, "Perceptual bias and technical metapictures: Critical machine vision as a humanities challenge," *AI Soc.*, vol. 36, no. 4, pp. 1133–1144, Dec. 2021, doi: 10.1007/s00146-020-01058-z.

[15] P. Hinton, "Implicit stereotypes and the predictive brain: Cognition and culture in 'biased' person perception," *Palgrave Commun.*, vol. 3, no. 1, p. 17086, Oct. 2017, doi: 10.1057/palcomms.2017.86.

[16] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," 2018, *arXiv:1808.00023*.

[17] Y. R. Shrestha and Y. Yang, "Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems," *Algorithms*, vol. 12, no. 9, p. 199, Sep. 2019, doi: 10.3390/a12090199.

[18] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, "Exposing the probabilistic causal structure of discrimination," *Int. J. Data Sci. Anal.*, vol. 3, no. 1, pp. 1–21, Feb. 2017, doi: 10.1007/s41060-016-0040-z.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[20] BBC Technology. (2019). *AI Fake Face Website Launched*. Accessed: Apr. 19, 2021. [Online]. Available: https://www.bbc.com/news/technology-47296481

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," 2019, *arXiv:1912.04958*.

[22] K. de Vries, "You never fake alone. Creative AI in action," *Inf., Commun. Soc.*, vol. 23, no. 14, pp. 2110–2127, Dec. 2020, doi: 10.1080/1369118X.2020.1754877.

[23] H. Mansourifar and W. Shi, "One-shot GAN generated fake face detection," 2020, *arXiv:2003.12244*.

[24] US Census Bureau. (2017). *National Population Projections, Projections for the United States: 2017 to 2060*. Accessed: Apr. 16, 2021. [Online]. Available: https://www.census.gov/data.html

[25] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age," 2019, *arXiv:1908.04913*.

[26] M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic, "Mitigating demographic bias in facial datasets with style-based multi-attribute transfer," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2288–2307, Jul. 2021.

[27] D. Goldsman, *Hypothesis Testing*. Atlanta, Georgia: Georgia Institute of Technology Press, 2010.

[28] A. Garg and D. Roth, "Understanding probabilistic classifiers," in *Proc. Eur. Conf. Mach. Learn., (ECML)*, 2001, pp. 179–191.

[29] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Res.*, vol. 5, pp. 2–8, Sep. 2016, doi: 10.1016/j.bdr.2015.12.001.

[30] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2021–2030.

[31] I. B. Y. Goodfellow and A. Courville, "Deep learning," in *Deep Feedforward Networks*. Cambridge, MA, USA: MIT Press, 2016, pp. 180–184.

[32] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.

[33] M. K. Scheurman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. 144, pp. 1–33, 2019.

[34] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proc. Mach. Learn. Res.*, vol. 81, pp. 1–15, Jan. 2018.

[35] R. M. Calogero, S. Tantleff-Dunn, and J. K. Thompson, Eds., *Self-Objectification in Women: Causes, Consequences, and Counteractions*. Washington, DC, USA: American Psychological Association, 2011.

[36] L. D. Rhode, *The Beauty Bias*. New York, NY, USA: Oxford Univ. Press, 2010.

[37] D. Weichselbaumer, "Is it sex or personality? The impact of sex stereotypes on discrimination in applicant selection," *Eastern Econ. J.*, vol. 30, no. 2, pp. 159–186, 2014. [Online]. Available: https://www.jstor.org/stable/40326127.

[38] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," 2019, *arXiv:1912.01865*.

[39] A. J. Onwuegbuzie and C. E. Daley, "Racial differences in IQ revisited: A synthesis of nearly a century of research," *J. Black Psychol.*, vol. 27, no. 2, pp. 209–220, May 2001, doi: 10.1177/0095798401027002004.

[40] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "StyleGAN2 distillation for feed-forward image manipulation," 2020, *arXiv:2003.03581v2*.

[41] M. A. Poran, "Denying diversity: Perceptions of beauty and social comparison processes among Latina, black, and white women," *Sex Roles*, vol. 47, nos. 1–2, pp. 65–81, 2002.

[42] N. Akinro and L. Mbunyuza-Memani, "Black is not beautiful: Persistent messages and the globalization of 'whit' beauty in African women's magazines," *J. Int. Intercultural Commun.*, vol. 12, no. 4, pp. 308–324, Oct. 2019, doi: 10.1080/17513057.2019.1580380.

[43] M. Silvestrini, "'It's not something i can shak': The effect of racial stereotypes, beauty standards, and sexual racism on interracial attraction," *Sexuality Culture*, vol. 24, no. 1, pp. 305–325, Feb. 2020, doi: 10.1007/s12119-019-09644-0.

[44] E. Ortiz-Ospina and M. Roser. (2018). *Economic Inequality by Gender*. Accessed: Apr. 19, 2021. [Online]. Available: OurWorldInData.org

[45] J. Merikull, M. Kukk, and T. Rõõm, "What explains the gender gap in wealth? Evidence from administrative data," *Rev. Econ. Household*, vol. 19, no. 2, pp. 501–547, Jun. 2021, doi: 10.1007/s11150-020-09522-x.

[46] M. Denton and L. Boos, "The gender wealth gap: Structural and material constraints and implications for later life," *J. Women Aging*, vol. 19, nos. 3–4, pp. 105–120, Sep. 2007, doi: 10.1300/J074v19n03_08.

[47] G. B. Forbes, J. Jung, and K. B. Haas, "Benevolent sexism and cosmetic use: A replication with three college samples and one adult sample," *J. Social Psychol.*, vol. 146, no. 5, pp. 635–640, Oct. 2006, doi: 10.3200/SOCP.146.5.635-640.

[48] T. F. Søgaard, "Bouncers, policing and the (in)visibility of ethnicity in nightlife security governance," *Social Inclusion*, vol. 2, no. 3, pp. 40–51, Sep. 2014, doi: 10.17645/si.v2i3.34.