AI and Ethics
https://doi.org/10.1007/s43681-021-00108-6

ORIGINAL RESEARCH

Bias, awareness, and ignorance in deep-learning-based face recognition

Samuel Wehrli[1] · **Corinna Hertweck**[2,6] · **Mohammadreza Amirian**[3,5] · **Stefan Glüge**[4] · **Thilo Stadelmann**[3,7]

Received: 30 July 2021 / Accepted: 4 October 2021

Abstract

Face Recognition (FR) is increasingly influencing our lives: we use it to unlock our phones; police uses it to identify suspects. Two main concerns are associated with this increase in facial recognition: (1) the fact that these systems are typically less accurate for marginalized groups, which can be described as "bias", and (2) the increased surveillance through these systems. Our paper is concerned with the first issue. Specifically, we explore an intuitive technique for reducing this bias, namely "blinding" models to sensitive features, such as gender or race, and show why this cannot be equated with reducing bias. Even when not designed for this task, facial recognition models can deduce sensitive features, such as gender or race, from pictures of faces—simply because they are trained to determine the "similarity" of pictures. This means that people with similar skin tones, similar hair length, etc. will be seen as similar by facial recognition models. When confronted with biased decision-making by humans, one approach taken in job application screening is to "blind" the human decision-makers to sensitive attributes such as gender and race by not showing pictures of the applicants. Based on a similar idea, one might think that if facial recognition models were less aware of these sensitive features, the difference in accuracy between groups would decrease. We evaluate this assumption—which has already penetrated into the scientific literature as a valid de-biasing method—by measuring how "aware" models are of sensitive features and correlating this with differences in accuracy. In particular, we blind pre-trained models to make them less aware of sensitive attributes. We find that awareness and accuracy do not positively correlate, i.e., that *bias* ≠ *awareness*. In fact, blinding barely affects accuracy in our experiments. The seemingly simple solution of decreasing bias in facial recognition rates by reducing awareness of sensitive features does thus not work in practice: trying to ignore sensitive attributes is *not* a viable concept for less biased FR.

Keywords Fairness · Convolutional neural networks · Discrimination · Ethnic bias · Gender bias

1 Introduction

FR has improved considerably and constantly over the last decade [1–4], giving rise to numerous applications ranging from services on mobile consumer devices to the use by law enforcement agencies [5–7]. The increased deployment has triggered an intense debate on the dangers of the pervasive use of biometrics [8–11] up to the point where regulation [12] and bans on the technology are discussed [13] and partially enforced [14, 15]. Several civil rights groups oppose facial recognition tools as they can easily be used for mass surveillance [13].

Besides these fears of surveillance, another critical issue is that facial recognition tools have been shown to perform at different levels of accuracy depending on which sociodemographic group a subject belongs to. In a seminal study of commercial face recognition software, Buolamwini and

✉ Samuel Wehrli
samuel.wehrli@zhaw.ch

1 School of Social Work, Zurich University of Applied Sciences, Pfingstweidstrasse 96, 8005 Zurich, Switzerland

2 Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland

3 Centre for Artificial Intelligence, Zurich University of Applied Sciences, Winterthur, Switzerland

4 Institute for Applied Simulation, Zurich University of Applied Sciences, Wädenswil, Switzerland

5 Institute of Neural Information Processing, Ulm University, Ulm, Germany

6 Department of Informatics, University of Zurich, Zurich, Switzerland

7 ECLT European Centre for Living Technology, Venice, Italy

Published online: 27 October 2021

Gebru [16] showed that these tools tend to misclassify darker-skinned women more often than lighter-skinned men. As face recognition is increasingly relied on to grant individuals access to services and locations, and to predict people's behavior, bias against certain socio-demographic groups easily results in these groups being more likely to be excluded from such services and locations. Bias becomes even more problematic when FR is used to identify suspects in a crime. When the FR algorithm misidentifies a person, this can have severe consequences: the misidentified person might unjustly be investigated or even charged with a crime they did not commit (as in the case of an American university student who was wrongly accused of terrorism by Sri Lankan police [17] or in the case of a black man who was wrongfully arrested in Michigan [18]).

Hence, the different levels of accuracy can be understood as an issue of bias: we expect FR to show approximately equal levels of accuracy for all socio-demographic groups and call it "biased" if it does not. One reason for the unequal levels of accuracy in FR is that the huge diversity in the appearance of human faces is not properly represent in the data used to train such models. Existing datasets tend to overrepresent lighter-skinned male faces, while other socio-demographic groups are underrepresented [16]. While bias also occurs when humans are the ones responsible for recognizing faces, the issue is more severe when conducted by machines as algorithmic decisions scale in speed, extent, and scope.

When one wants to avoid biased decisions made by humans, a standard approach is to make sensitive attributes unavailable. An every-day example are resumes: in the US, age and gender information as well as images are omitted in resumes. If the recruiters in charge are not aware of ethnicity, gender, and age—so the thinking goes—then decisions made by them cannot be biased by these sensitive features. Such biases could lead to strong candidates being wrongfully omitted. Ignorance of sensitive attributes is thus seen as a way of finding better candidates while mitigating discrimination. The methodology of being blind toward sensitive features is not new: As the "veil of ignorance," it is part of John Rawls's influential book "A Theory of Justice" (1971) [19] which deals with the political philosophy of just distribution and fairness. Behind this "veil of ignorance," people do not know their own identity and circumstances of life (gender, job, health, etc.). Rawls uses this concept as part of a thought experiment to find the principles based on which society and its institutions should be designed. Because people are biased by their situation in life (e.g., by knowing that they are born as a cisgendered white man), asking people for their ideas for such principles would most likely lead to biased principles. Therefore, Rawls asks people to imagine themselves behind this "veil of ignorance" in this newly constructed world. John Rawls demands ignorance of our own identity when imagining the "ideal" society and its institutions—with the goal of reducing bias and creating better results.

With both recruiting and the "veil of ignorance," the assumption is that humans' awareness of sensitive features leads them to make biased decisions, which harms marginalized groups. As Nyarko et al. [20] show, people are quick to apply this concept of removing sensitive attributes to machine learning models—despite the potentially harmful consequences for the disadvantaged group (see, e.g., [21–23]). The underlying assumption of those people is that removing sensitive attributes would reduce bias and thus help the disadvantaged group.

Considering the case of FR, this would mean that to reduce bias, "awareness" of sensitive features has to be removed. As stated above, we define bias as notably different level of accuracy between socio-demographic groups. We will refer to "awareness" as the extent to which a machine (e.g., a FR model) is aware of the presence of sensitive features (e.g., gender or ethnicity). In the literature, awareness is often equated with bias in FR: the idea is that removing awareness (i.e., decorrelating facial representations and sensitive attributes) simultaneously reduces bias—similar to how it is assumed that hiding sensitive features removes humans' tendency to make prejudiced decisions [24–28].

In this paper, we explore what this removal of awareness means on a technical level and demonstrate why it cannot be equated with reducing bias. We thus argue that $bias \neq awareness$ with the important consequence that dealing with awareness in FR models does not necessarily reduce bias in any desired way. The rest of the paper is organized as follows: Sect. 2 describes how this work relates to other studies in this field. Section 3 explains the methods and data as well as the existing face recognition models that we use to experimentally examine the relationship of awareness and bias in face recognition. We then present and discuss the results of these experiments in Sects. 4 and 5 and draw conclusions in Sect. 6.

2 Related work

Racial biases are an issue across different sub-domains of computer vision: besides the field of FR, image classification models have, for example, been criticized for mis-labeling black men as "primates" [29]. Through the work of Joy Buolamwini and Timnit Gebru [16], FR's bias

problem became a topic of public debate [30]. A follow-up study of their work revealed that Microsoft, IBM, and Face++ released new versions of their API that improved the audited metrics [31]. IBM also removed the facial detection from its API in September 2019.[1] In the public sector, San Francisco was the first US city to ban the usage of FR technology in 2019 [32] as a consequence of discovering biases in FR [33]. Several other cities followed. Since then, a tremendous amount of research has *measured* [25] and *reduced* [34] biases in FR technologies. The remainder of this section presents recent works on measuring racial biases and methods for removing biases. To reduce the racial bias, researchers followed these main directions: 1) balancing datasets, 2) model selection and loss design, and 3) removing the sensitive feature in the representations of faces used for identification. This paper relates to the third category and investigates the effect of a blinding method to remove sensitive features from face matching techniques. Systematic reviews of the recent attempts to tackle biases in machine learning and FR are represented in [35, 36].

Identifying and quantifying the amount of bias in FR technology are the initial step toward less-biased FR. Garcia et al. showed that face matching confidence of FR models correlates with gender and ethnicity, thus revealing demographic bias [37]. Cavazos et al. demonstrated that different thresholds are needed to equalize false accept rates (FARs) and the recognition accuracy [38]. Serna et al. quantified FR bias using normalized overall activation of the models for various races [39].

The first direction to overcome racial bias in FR is addressing the bias in the data: e.g., measurement error (systematic errors in the measurements of variables for specific groups) or representation bias (not everyone has the same probability of being in the dataset, meaning that the training data do not represent the real world's diversity) [35]. To suppress the effects of imbalance in datasets, Kortylewski et al. proposed using synthetic data [40]. Robinson et al. introduce a racially balanced dataset [34]. With that, they were able to show how the performance gaps in FR for various races decrease when adapting the decision thresholds for each race.

Modifying the model choices, e.g., the training process and the target, is the second venue researchers explored to remove racial bias in FR [41]. Yu et al. adapted the selection of face samples for training based on the data distribution and model bias [42], while Wang et al. attempted to transfer knowledge from the source domain (Caucasian) to target domains (other races) through learning facial features with adequate generalizing across different races [43].

The last approach is removing sensitive information related to races [24]. Generative Adversarial Networks (GANs) inspired several researchers to blind models and/or reduce the correlation between sensitive features and facial attributes for face recognition [26, 27, 44, 45]. Adeli et al. proposed an adversarial loss to minimize the correlation between model representations and sensitive information (races) and statistical dependency of the learned features and source of bias (racial group) [46]. In this paper, we use a blinding technique that applies to every trained model to remove sensitive information from model representations and demonstrate that removing awareness does not necessarily remove the racial bias in several FR technologies.

3 Materials and methods

We start this section with a short recapitulation of the fundamentals of deep FR models. Afterwards we introduce the FR models and the evaluation dataset we chose for the present work. Finally, we detail the methods we used to quantify/measure the awareness of deep FR models regarding specific socio-demographic groups and to remove information in the models' embeddings with respect to these groups.

3.1 Basics of deep face recognition models

Wang et al. [47] provide a comprehensive survey on deep FR methods, including algorithms, databases, training protocols, and applications. In this section, we limit ourselves to a short review of the basic algorithm that yields a compact representation of a face in a feature space, a.k.a. embedding.

In image processing, deep learning methods, such as Convolutional Neural Network (CNN), use a cascade of multiple layers of processing units for feature extraction and transformation. It was shown that each layer learns multiple levels of representations which correspond to different levels of abstraction depending on the task at hand. Regarding face recognition, a major advantage of this hierarchy of concepts is a strong invariance to changes in face pose, lighting, and expression changes. Figure 1 shows the general structure of a CNN on the task of face identification, together with the features learned on different levels of the network hierarchy. Contrastive and triplet loss functions are use to train the CNNs of deep FR models [2]. They optimize the CNN, such that embeddings (i.e., features) of positive image pairs (same identity) are close to each other, whereas embeddings of negative pairs (different identity) are pushed apart.

To summarize, every face image I (160×160 pixel corresponding to 25,600 features) is processed in a deep CNN that

[1] https://cloud.ibm.com/docs/visual-recognition?topic=visual-recognition-release-notes.

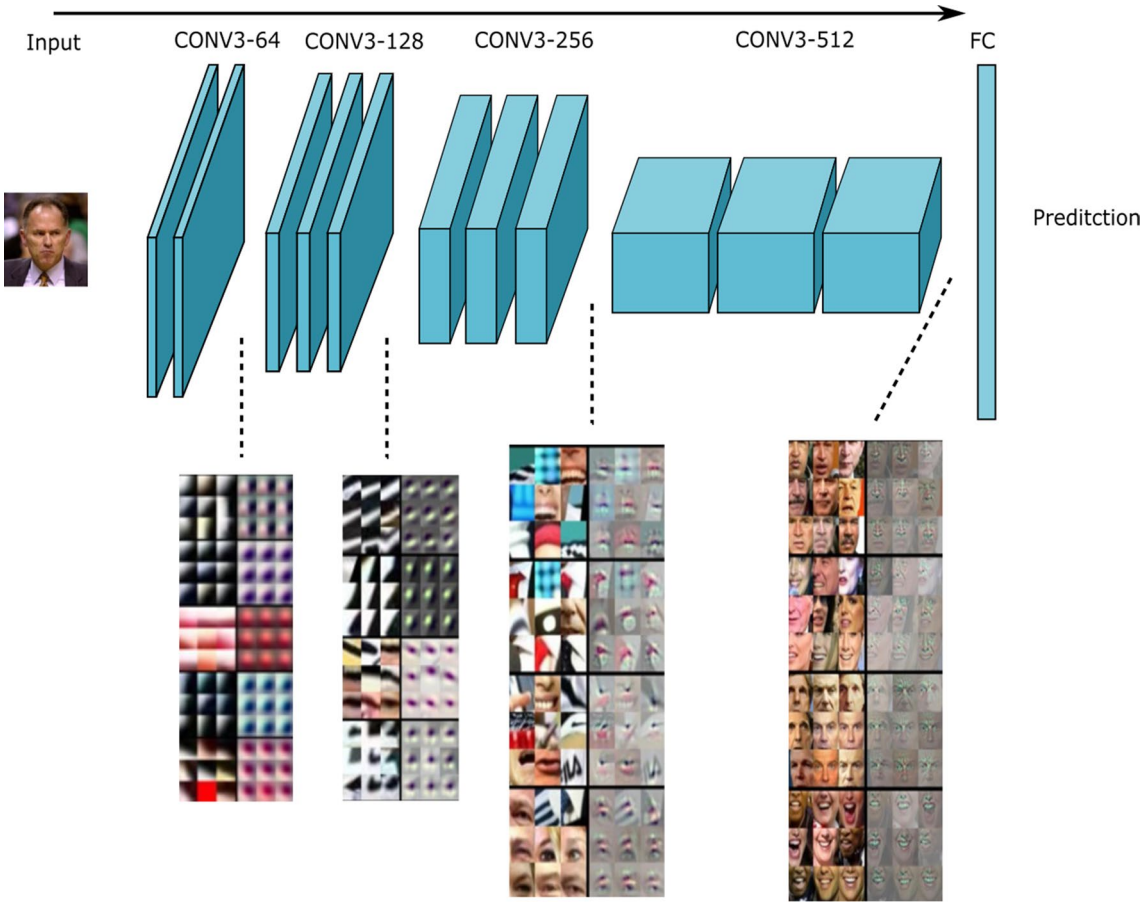

Fig. 1 Simplified example illustrating a hierarchical CNN architecture trained to convert pixels of the input faces into compact face representations at the last fully connected (FC) layer. The model consists of multiple layers that convolute and pool the input (CONV3 layers). Each block of convolutions works on a differently sized part of the input image, a.k.a. receptive field. To promote the learning of basic features at the bottom layers and more complex features (eyes, mouth etc.) at the top, the receptive field is enlarged every block further up the hierarchy. The output is a compressed representation of the face which can directly be used to make a prediction about the identity of the person

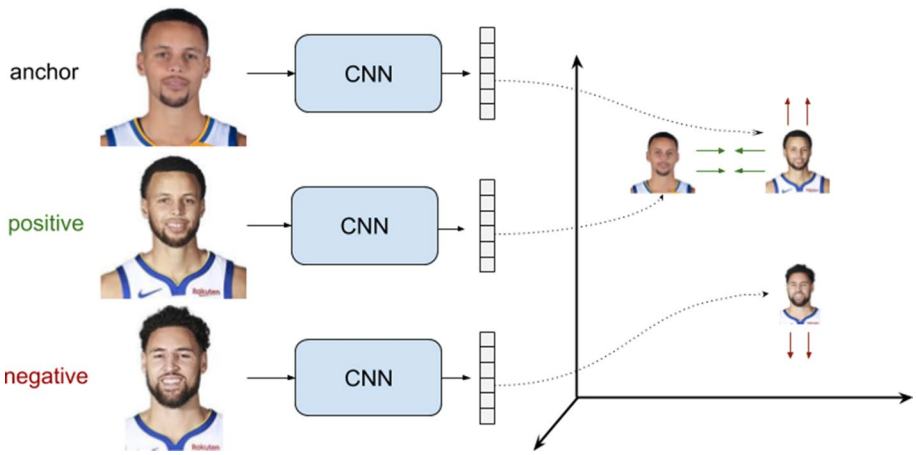

Fig. 2 General sketch of a deep FR model with triplet loss[2]. Each image is processed in the CNN model and further mapped into the embedding space, such that similar faces are close together

learned a hierarchy of features ranging from basic concepts like edges up to complex concepts like eyes. On top of these features, a compact vector representation $\boldsymbol{x}$ (typical length of 128) is learned. In general, we refer to these compact representation as embedding. Furthermore, the model is trained to generate embeddings, such that $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are close together

if the input images I_i, I_j belong to the same person. Figure 2 depicts this concept.[2]

3.2 Face recognition models

As described above, we are operating on the embedding space of FR models. This allows us to easily compare different readily-trained models. For our comparison, we chose two different openly available FR models/architectures: first, a model from the popular Visual Geometry Group (VGG) [48] family and second, FaceNet [2].

Together with the VGGFace2 dataset [48], the authors provide trained models[3] that we applied in our study. The models were pre-trained on the MS-Celeb-1M [49] dataset and then fine-tuned on VGGFace2. These SE-ResNet-50 models follow the architectural configuration in [50]. Besides the architecture, the models differ in the lower dimensional embedding layer (128D/256D) which is stacked on top of the original final feature layer (2048D) adjacent to the classifier. All models were trained with standard softmax loss. In our experiments, we did not see any significant difference regarding different sizes of the embedding layer. Hence, we show the results for the VGG128 with an embedding layer of size 128.

FaceNet is a face recognition system that was described by Florian Schroff, et al. at Google [2]. In our experiments, we use a model that is based on the GoogLeNet style Inception models [51] with approximately 23M trainable parameters. The trained model is freely available[4] and was pre-trained on the MS-Celeb-1M [49] dataset. The embedding layer is also of size 128.

3.3 Evaluation dataset to study racial bias

In the following, we look at the sensitive attributes ethnicity and gender. The Racial Faces in-the-Wild (RFW) dataset was designed to study racial bias in FR systems [43]. Images are annotated with one of four labels, namely *Caucasian*, *Asian*, *Indian*, and *African*, which form four ethnic clusters. Each subset contains about 10k images of 3k individuals for face verification. According to [43], the labels for Caucasian and African ethnicity were assigned by the Face++ API [52] and those for Asians and Indians from the nationality attribute in FreeBase celebrities [53]. To avoid the negative effects caused by the biased Face++ tool, the authors manually checked some images with low confidence scores from Face++.

Table 1 Number of samples per cluster regarding different facial characteristics in the RFW dataset that are associated with bias

	Caucasian	Indian	Asian	African	Total
Male	6921	7419	5647	10,053	30,040
Female	3178	2802	3955	344	10,279
Total	10,099	10,221	9,602	10,397	40,319

The dataset is balanced with respect to ethnicity but skewed with respect to gender (75% male, 25% female). African women are strongly underrepresented and constitute less than 1% of all samples

In addition to ethnicity, we added a gender label to each image using a Wide Residual Network trained on the UTKFace [54] and IMDB-WIKI [55] datasets.[5] The model's performance is reported to be around 88% accuracy [56, 57]. The gender prediction is accessed in terms of a continuous score s_{gender} between 0 and 1, where lower values indicate male and higher values indicate female. To create fixed clusters of samples, the gender score was split at $s_{\text{gender}} < 0.5$ for male and $s_{\text{gender}} > 0.5$ for female. As we have multiple face images for each person in the dataset, each person was labeled to be male/female based on their mean score. Table 1 gives an overview of the resulting number of samples per cluster with respect to ethnicity and gender.

3.4 Awareness of sensitive features

We investigate how well machine learning models can predict the sensitive features, such as ethnicity and gender, based on the face embedding. The intuition is that an FR model is "aware" of a sensitive feature if it can be predicted from the embedding vectors produced by the FR model. This inference is a classification task and the performance depends on the classification model at hand. If simple models, more precisely models with a low number of parameters, can properly infer the sensitive features, awareness is high. If models with a high number of parameters are required, then awareness is lower. There is no awareness if the features cannot be better predicted than random guessing.

3.5 Blinding

As stated above, deep FR models map images of faces into an embedding vectors $\boldsymbol{x}_i$. Given a data set with labels for sensitive attributes such as ethnicity and gender, the images can be grouped into clusters based on these labels. To investigate the influence of this clustering on the face recognition rates, we propose a blinding procedure to remove the information related to the separation of these clusters in the

[2] Attribution: https://www.pinterest.ch/pin/663999538792168278/, CC BY-SA 4.0, via Wikimedia Commons.

[3] https://github.com/ox-vgg/vgg_face2.

[4] https://github.com/nyoki-mtl/keras-facenet.

[5] https://github.com/yu4u/age-gender-estimation.

embedding space. The procedure is a linear operation and uses the following steps:

1. Compute the centroids of the clusters defined by the sensitive attributes in the embedding space.
2. Use a one-vs-rest (OvR) approach to calculate the "directions of discrimination" given by the centroids of each cluster relative to the other centroids.
3. Apply singular value decomposition (SVD) on the "directions of discrimination" to find an orthonormal basis spanning the "discriminatory subspace."
4. Remove projections onto the "discriminatory subspace" from the embedding vectors. This results in new embedding vectors whose centroids fall on top of each other, i.e., the separation due to sensitive attributes is removed—hence the term "blinding".

The procedure outlined above operates on the embedding vectors $\boldsymbol{x}_i$ where i denotes the sample. Associated with each sample is a cluster label $k \in \{1, \ldots, K\}$. As a first step, we define the centroids of each cluster by the average

$$\bar{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_{i \in C_k} \boldsymbol{x}_i, \tag{1}$$

where C_k is the set of embedding vectors associated with cluster k and n_k is the corresponding size. Following an OvR approach, the normalized direction of discrimination of each cluster k to the other clusters is given by the vectors

$$\boldsymbol{u}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|} \quad \text{with} \quad \boldsymbol{v}_k = \bar{\boldsymbol{x}}_k - \frac{1}{K-1} \sum_{k' \neq k} \bar{\boldsymbol{x}}_{k'}, \tag{2}$$

where K is the number of clusters. By construction, the vectors $\boldsymbol{u}_k$ are not linearly independent, but span a subspace of rank $K-1$. This can by verified by applying a SVD on the matrix $U = [\boldsymbol{u}_1 \ldots \boldsymbol{u}_K]$. SVD also provides an orthonormal basis $B = [\boldsymbol{e}_1 \ldots \boldsymbol{e}_{K-1}]$ of the corresponding subspace. The final step is to remove the projections onto this subspace by

$$\boldsymbol{x}_i^b = \boldsymbol{x}_i - \sum_{j=1}^{K-1} (\boldsymbol{x}_i \cdot \boldsymbol{e}_j)\, \boldsymbol{e}_j, \tag{3}$$

where $(\boldsymbol{x}_i \cdot \boldsymbol{e}_j)$ is the dot (or scalar) product. Equation (3) yields new embedding vectors $\boldsymbol{x}_i^b$ with the same shape as the original ones. The upper index b stands for "blinded" inspired by the fact that information with regard to the discriminatory dimension has been removed.

3.6 Experimental setup

We extract the embeddings of the approximately 40k face images from the RFW testset for the VGG and FaceNet models. Face detection and alignment is done using the MTCNN approach[6] proposed by Zhang et al. [58]. Based on the embeddings, we analyze the models' awareness regarding sensitive attributes ethnicity and gender using the classifier performance within the embedding space (cf. Sect. 3.4). Furthermore, we report the bias of the models based on the actual FR rates. In a second step, we apply the proposed blinding procedure (cf. Sect. 3.5) to the embedding spaces and again report the models' awareness and bias.

4 Results

In this section, we present the results of our experiments. Specifically, we compare the model's awareness of sensitive features as well as bias with respect to ethnicity and gender before and after the blinding procedures.

4.1 Awareness of sensitive attributes

To visualize the structure of the embedding space (128 dimensions), we use t-distributed Stochastic Neighbor Embedding (t-SNE) [59] as an unsupervised way to reduce the dimensionality to a 2D representation. t-SNE maps points from a high-dimensional space to a lower dimensional space, preserving local distances. Points which are close to each other in the high-dimensional space remain close to each other in the low-dimensional space. This property is particularly suitable for the FR task where images are mapped to the embedding space and decisions are based on distances in this space. Figure 3 shows the 2D visualization of the embeddings of the RFW test set. As one can see, the dimensionality reduction reveals well-separated clusters defined by ethnicity and gender. The fact that the data are skewed with respect to gender is reflected in the figures, in agreement with the sample counts shown in Table 1. The t-SNE visualizations for the VGG128 and FaceNet models are surprisingly similar: Caucasian men lie in the center, surrounded by the remaining groups in similar positions. As stated, t-SNE is an unsupervised clustering method and groups embedding vectors solely based on their respective distances without any direct information about the sensitive features. Nonetheless, the clusters appear very nicely separated, suggesting that the embedding space is separated into different sectors corresponding to different ethnicities and genders and that the models under considerations are therefore highly "aware" of the sensitive features.

To associate this vague intuition of "awareness" with a more formal approach, we investigate how well different classifiers predict the sensitive features based on the face embeddings. The intuition is that "awareness" toward

[6] https://github.com/YYuanAnyVision/mxnet_mtcnn_face_detection.

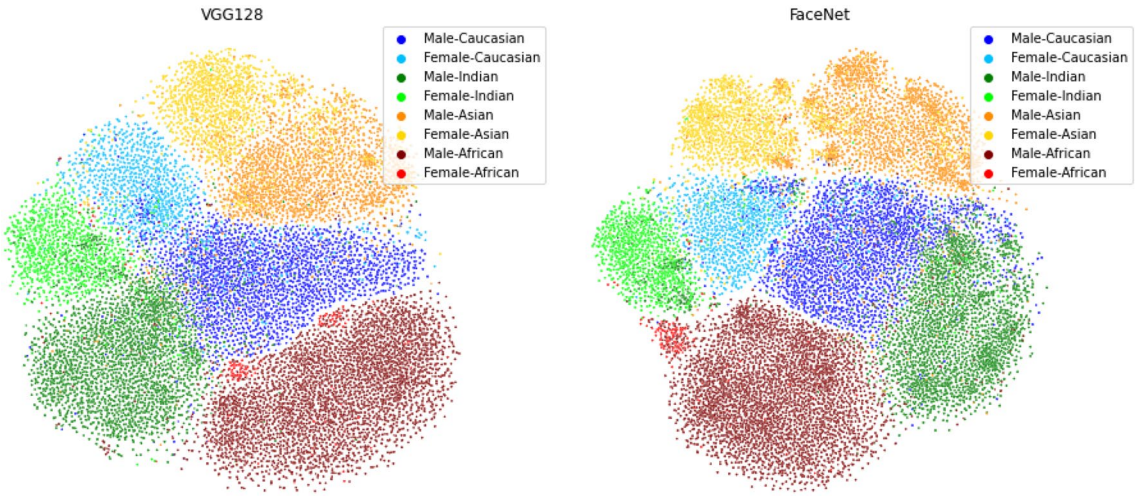

Fig. 3 t-SNE visualization (2D) of the embedding space of the RFW test set samples. The coloring is based on different labels corresponding to the sensitive features ethnicity (color) and gender (light versus dark). Left: VGG128 model. Right: FaceNet model (color figure online)

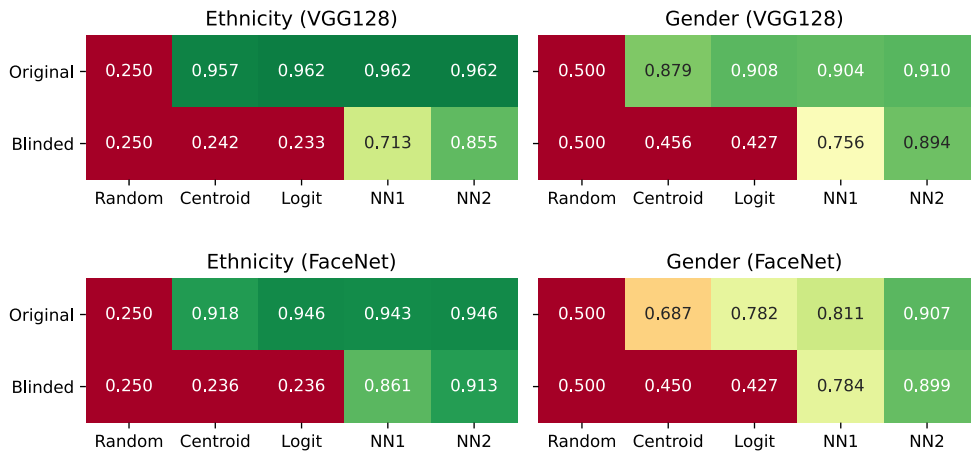

Fig. 4 "Awareness" as represented by the macro-averaged F1-scores for various classifiers, predicting sensitive features based on the face embeddings for the VGG128 model (upper panel) and the FaceNet model (lower panel). The following classifiers from scikit-learn were used: (Random) theoretical value of random guessing; (Centroid) nearest centroid classifier; (Logit) logistic regression; (NN-1,2) neural network with one or two hidden layers (100 nodes) and relu activation. A train/test split of two-thirds/one-third was used. The upper rows represent the scores using the original embeddings. The lower rows show the scores for the blinded embeddings. The scores are colored, ranging from red for random guessing to green for a perfect score of 1 (color figure online)

sensitive features is high if simple models (low number of parameters) can accurately predict these features. The ability to predict sensitive features based on the embedding vector is evaluated for different classifiers. In the case of ethnicity, it is a multi-class classification task (4 classes). Therefore, we use macro-averaged F1-scores as a performance measure, which is the unweighted mean of the F1-scores of each label, as shown in Fig. 4. The upper bound of this score is 1. A lower baseline is given by random guessing and is the inverse of the number of clusters for a given sensitive feature, i.e., 0.25 for ethnicity and 0.5 for gender. Figure 4 shows that the classification scores on the original embeddings are close to 1. Moreover, simple linear classifiers such as nearest centroid classification and logistic regression (both with a low number of parameters) show the same performance as the more advanced neural network classifiers. This confirms the hypothesis made above that the embedding space is structured into sectors given by the sensitive features. The fact that centroid classifiers work well means that these sectors are linearly separable and well-represented by their centroids. Therefore, we conclude that the models are indeed "highly aware" of ethnicity and gender.

Previous work [25] investigated this structure by means of various clustering scores such as the Silhoutte coefficient.

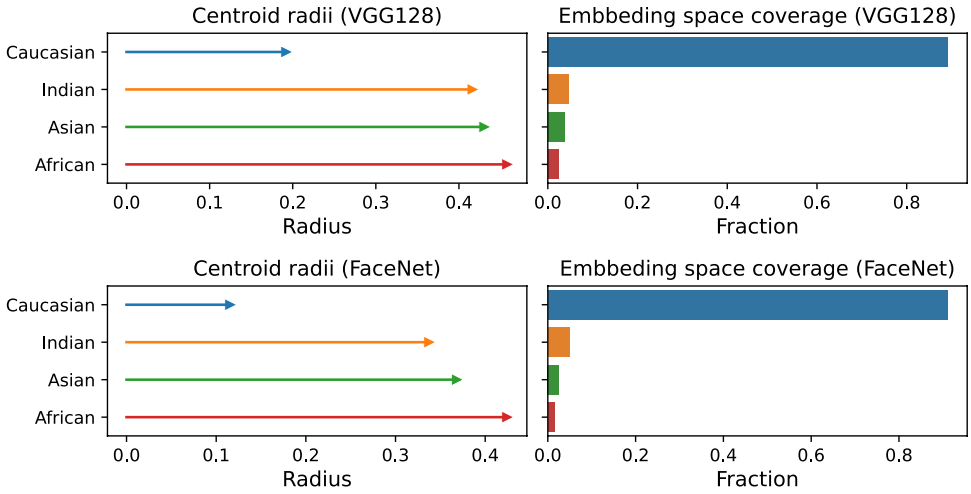

Fig. 5 Left side: radii of the origin of the embedding space to centroids of the clusters related to ethnicity. Right side: coverage of the embedding space associated with ethnicity. This is calculated by classifying (centroid classifier) randomly generated embedding vectors. The fractions add up to 1

These scores suggested negligible structuring in contrast to present findings. The discrepancy is due to the high dimensionality of the embedding space: In order for the linear classifiers to work, it is enough if the sectors related to ethnicity and gender are separated by a few dimensions out of many (128 in the models under considerations). In contrast, clustering scores give an average clustering for all dimensions and yield low scores if only a few dimensions contributed to the separation, which explains the discrepancy.

The fact that the cluster centroids are representative for the whole cluster lends itself for further analysis. The positions of the centroids in the high-dimensional space can obviously not be visualized. It is however interesting to look at the radii (norm) from the origin of the embedding space to the centroids. This is depicted on the left side of Fig. 5 for the ethnic clusters. The embedding vectors for different faces all have a radius 1, because they are the output of a soft-max layer. The centroids, which are averages of embedding vectors, therefore have radii < 1 as can be seen in the figure. Interestingly, the radius of the Caucasian cluster is roughly half of the other radii, confirming that Caucasians are indeed closer to the origin which is inline with the t-SNE visualization shown in Fig. 3. The centroid classifier can also be used to generate a measure of the relative "size" of the different sectors by randomly generating embedding vectors (with radius 1) and classifying them with the centroid classifier. The result is shown on the right side of Fig. 5 for ethnicity. It turns out that the Caucasian sector covers roughly 90% of the embedding space in this metric. This is a clear and remarkable result which shows that different ethnic groups are treated differently in the models under consideration. To put it dramatically: In the models under consideration, 90% of the embedding space serves the purpose of Caucasian face recognition.

The differences in centroid radii and the embedding space coverage may be related to the bias in face recognition discussed below. The relation can either be causal (the difference in radii causes the bias) or rather indicative (the difference is due to the same root cause as the bias). To differentiate between the scenarios, we propose the blinding procedure described above which removes the dimensions separating the centroids (cf. Sect. 3.5). In the case of ethnicity, there are 4 centroids which span a 3D subspace (similar to the well-known fact that three points span a 2D plane). By blinding, i.e., projecting out this 3D subspace, the centroids fall on top of each other. As a consequence, the sectors corresponding to different ethnicities will be shifted together and the clusters are no longer represented by their centroids. We apply this blinding procedure to both ethnicity and gender. Figure 4 shows that the performance of the linear classifiers drops to random guessing as expected. Hence, the clusters are no longer linearly separable. The more complex neural network classifiers are still able to predict the sensitive features. Therefore, after blinding, more complex models with a higher number of parameters are needed to predict the sensitive features. This means, in our terminology, that the "blinded" embeddings are indeed less "aware" of the sensitive features.

4.2 Bias with respect to sensitive attributes

Face recognition rates and bias are evaluated with the RFW dataset. The RFW dataset provides image (i.e., embedding) pairs. These pairs can either be two pictures of the same person (positive pair) or of two different people (negative pair). The resulting task is a binary classification of the pairs into positive and negative pairs. Note that all pairs have the same ethnicity label (only people with the same ethnicity label are compared). The recognition rate is the accuracy of the corresponding classification. Here, we investigate the error rate

$$\begin{aligned}\text{Error rate} &= 1 - \text{recognition rate} \\ &= \text{share of false positives} + \text{share of false negatives},\end{aligned} \tag{4}$$

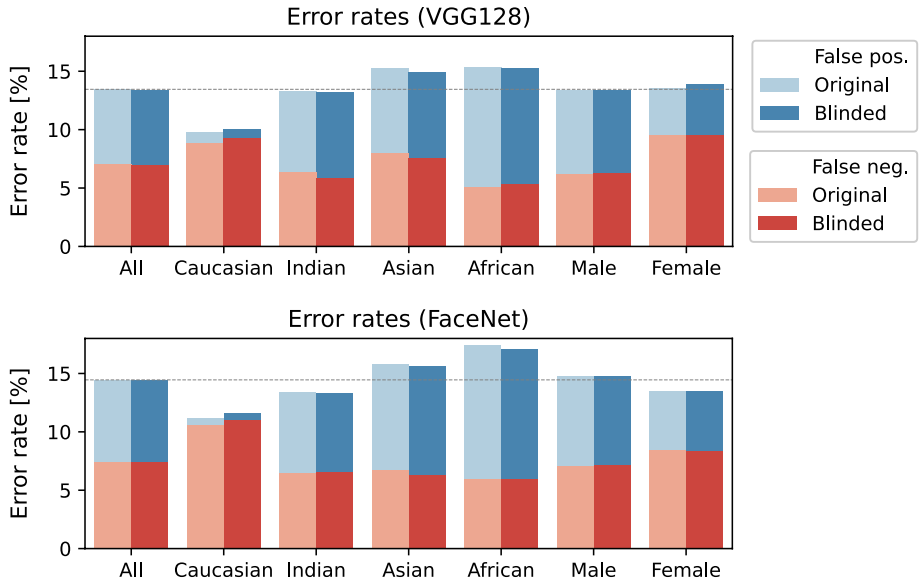

Fig. 6 Relative face recognition error rates for the VGG128 and FaceNet models. The error rates are given for *all* image pairs, for the different ethnic groups (*Caucasian, Indian, Asian*, and *African*) as well as for gender (*male, female*). The horizontal line helps to indicate whether a specific group performs better or worse than the overall average. The colors distinguish the two types of errors: False positives (blue) are pairs of different identities which are mistakenly predicted as identical, whereas false negatives (red) are identical faces mistakenly predicted as different. The brightness indicates the type of embedding. Light: original embeddings. Dark: blinded embedding (color figure online)

which has two contributions due to the two types of misclassification

- false negatives: same identity predicted as different ones (red in the figures);
- false positives: different identities predicted as the same one (blue in the figures).

The classification itself is done by calculating the (Euclidean) intra-pair distance d between the embedding vectors of the images of the pair and comparing it to a threshold d_c. A positive (same) pair is predicted for $d < d_c$. Otherwise, a negative (different) pair is predicted. The error rates of these classifications are shown in Fig. 6 for both the original and the blinded embeddings. A substantial difference is apparent between the different ethnic groups and to a marginal extent for the gender groups. The error rate is lowest for the label "Caucasian" ($\approx 10\%$), whereas the error rate is highest for the label "African" ($\approx 15\%$). Moreover, for people with the label "Caucasian", false negatives are the most common error type. For people with the label "African", however, false positives are the most common error type. Apparently, we observe two types of bias

1. difference in the total error (or recognition) rate;
2. difference in ratio between false-positive and false-negative error rates.

As can be seen, removal of the separation of the sectors associated with different sensitive features (i.e., blinding) affects the error rates and bias only marginally. We conclude that the concept of awareness introduced above is different from bias. In addition, the difference in the centroid radii mentioned above does not cause the bias, as removing it by blinding does not affect it. t-SNE allows us to visualize where the misclassifications are located in the embedding space. Figure 7 shows that the misclassifications are randomly scattered. It is therefore not surprising that moving the different clusters on top of each other by blinding has only marginal effects on the recognition rate and bias.

5 Discussion

The results show that putting FR behind a "veil of ignorance"—where predicting people's assigned gender and ethnicity label becomes harder—does not have any notable influence on the accuracy of the FR model. In this way, dealing with biases in machines is strikingly different from dealing with human biases in, e.g., job application screening where ignorance is often deemed to be a necessary condition for unbiased decision-making. We argue that this is because the task of screening job applicants is fundamentally different from the task of FR technology, which is to decide whether two pictures show the same person. If humans take biased decisions when looking at CVs (e.g., accepting more

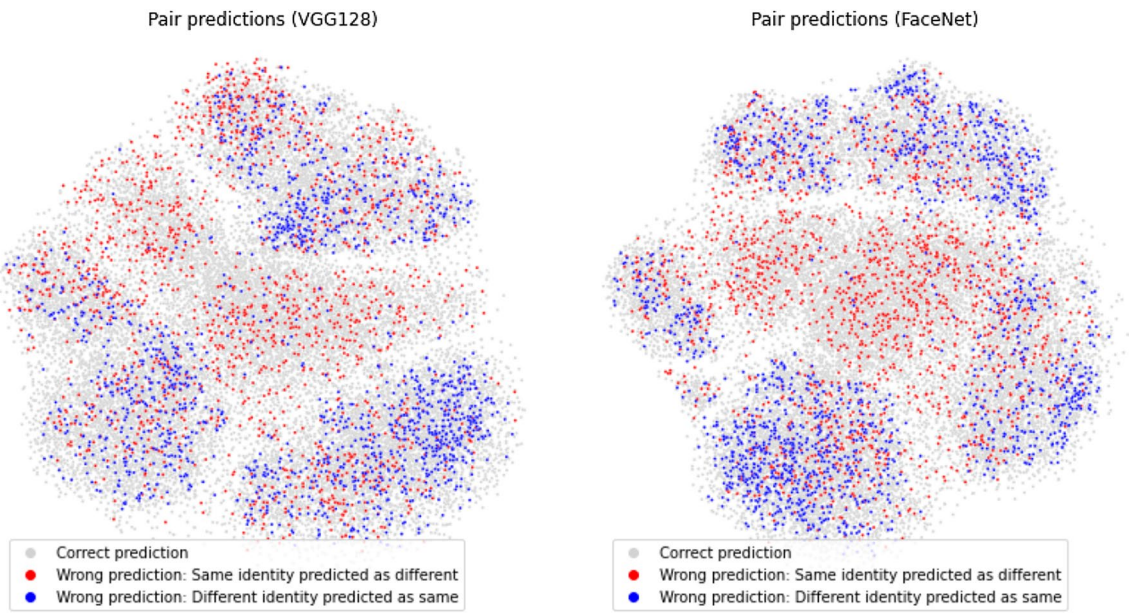

Fig. 7 This figure is derived from the t-SNE coordinates shown in Fig. 3. Each point in the plot represents the average coordinates of a pair, either positive with the same identity or negative with different identity. The grey points represent correct predictions of positive (same identity) or negative (different identity) pairs by the face recognition algorithm. Red points are the cases where positive pairs are mistakenly classified as negative pair. Blue points are the cases where negative pairs are mistakenly classified as positive pair. Left: VGG128 model. Right: FaceNet model (color figure online)

men than women without any reasonable justification), we assume that this is caused by implicit or explicit prejudices. When humans have to match faces, biases in the form of lower performance in recognizing members of certain groups might not be caused by prejudices, but by a lack of exposure to these faces—the so-called "cross-race effect" [60]. The same likely applies to the FR models in our experiments: the reason for their different performance levels is not awareness of groups' race or gender. Rather, it is because they have not "seen" as many people from that group before. We thus caution machine learning experts against trying to solve the issue of bias in FR by applying their intuition of fairness to FR. Simply ignoring sensitive attributes does not seem to be a proper mediator of unbiased FR. Instead, FR developers noticing biases in the performance rates of their models should avoid trying to "blind" their model, but should rather improve the training data of their model as suggested in previous work [16, 43].

However, the experiments also demonstrated how easy it is to predict the assigned gender and ethnicity label from the existing models, such as FaceNet. This is particularly worrisome considering recent reports of Russia deploying tools that detect people's ethnicity [61] and China using tools that detect Uighur faces [62]. One might therefore want to consider if such open-source models should be "blinded" before their release. While it does not improve the accuracy rates for any groups, it also does not harm the accuracy, but might potentially protect from such tools being used to assign ethnicity labels to people.

We also found that there is an imbalance in the type of errors that the models make. We assume that the aforementioned imbalance in the training data is what leads to this difference in the types of errors that the model makes for the different ethnic groups. Assuming that the training data consisted of a disproportionately high number of Caucasian faces, the trained model would essentially provide more "space" in the embedding space for Caucasian faces. As noted in Sect. 4.1, the Caucasian sector indeed covers a large part of the embedding space, namely about 90%. This means that two Caucasian faces will on average have a greater distance to each other than two pictures from another group. When we now use the same threshold to determine at which distance two pictures are classified as "identical" instead of "different", Caucasian faces—being generally more spread out—are more likely to be classified as "different". This explains the high false-negative rate for the Caucasian group. Pictures from other groups are generally closer to each other, which makes a classification of two pictures as showing the same person more likely and explains the comparatively higher false-positive rate. As suggested by Robinson et al. [34], a way to deal with these different levels of error types is to create ethnicity-specific thresholds for when two faces are classified as "identical". The distance up to which Caucasian faces are classified as "identical" would have to be higher than that of other groups. However, this requires a

classification of individuals into ethnic groups before checking potential matches, which is a at least questionable practice considering how the results of such an analysis could be misused. Therefore, while this solution makes sense on a technical level, we question if its benefits (equalization of error types across groups) outweigh the potentially harmful consequences (misuse of ethnicity classifications). Insofar, it once again appears to be the better option to address the root cause of the issue of the differing error rates: the training set.

The reason why the differences in error types can have dire consequences becomes evident when we think about a common application of facial recognition: policing. False negatives in the context of policing might mean not recognizing a suspect as the camera footage of them is not matched with their picture in the database. False positives are cases where camera footage of an innocent person is confused with pictures of the wanted person. As false negatives are far more common for Caucasian faces than false positives, the FR system errs on the side of not flagging Caucasian people as potential suspects. The situation is reversed for faces labeled as "African": here, the system is more likely to flag an innocent person as a suspect than it is to accidentally let a suspect pass by without being flagged. This is particularly problematic considering the discrimination that people of color face in many regions and, on the other hand, the privilege that white people have. Anna Lauren Hoffmann points out *"the different real-world consequences false results might have for different groups"* [63, p. 907] in the context of AI applications. *"A person of relative socioeconomic advantage is more likely to have the time or resources necessary to contest an unfair decision—an imbalance that persists regardless of the fact that differently-situated groups stood an equal chance of being falsely flagged within the system"* [63, p. 907]. When we are looking at error types for marginalized groups, we thus also have to consider how these groups deal with the different error types.

Finally, we want to emphasize that this paper only considered one of the two main issues of FR technology: its biases, i.e., different levels of accuracy and different error types. While we showed that a seemingly easy solution to this problem ("blinding" the model) does not work in practice, even a method that mitigates this issue would not address the second and arguably more pressing issue of FR: its potential usage for mass surveillance. Simply advocating for better performance rates might be dangerous as long as there are no policies that guide what FR can legally be used for. Moreover, we have to keep in mind that mass surveillance, just like other harmful technologies, disproportionately affects marginalized groups. Mohamed et al. [64], for example, describe several cases in which problematic AI applications are beta-tested in poorer communities as less resistance is expected due to their limited resources. An example of this is the case of a facial recognition systems that has been proposed to be used in a Brooklyn apartment complex whose tenants are mostly black. The tenants worry that the technology is not implemented for a safer environment for the tenants, but rather for their surveillance as such data could be abused [65]. Considering the potentially harmful consequences of FR technology as well as the systemic disadvantages marginalized communities face, simply advocating for equal accuracy rates is thus not enough. Instead, we need to think further about how such technologies are being used. This includes policy as the previously mentioned bans on facial recognition (see, e.g., [32]) and the EU's current attempt of regulating AI [66, 67].

6 Conclusion

In this paper, we operationalized "awareness" as a measure of how well different classifiers predict the sensitive features based on the face embeddings. The intuition is that "awareness" toward sensitive features is high if simple models (with a low number of parameters) are sufficient for accurately predicting the sensitive features based on the face representations learned by an FR neural network. For example, we would say that a model's "awareness" of gender is high if a simple linear model can accurately predict the gender of a test subject from its picture when trained on face embeddings and gender labels.

Inspired by the example of human job application screening and its early adoption in the FR literature, we introduced a *blinding* procedure to reduce awareness. We showed that this procedure allows us to reduce awareness in a controlled and selective way. Applying this procedure enabled us to answer the question of whether removing information about sensitive features helps to reduce bias as it is assumed to do in the human example. For this, we compared the models' awareness and bias before and after blinding. We came to the conclusion that indeed $bias \neq awareness$.

We further found that the models make different kinds of errors for different ethnic groups. For the Caucasian group, the models are more likely to identify two pictures of the same person as different people. The opposite is the case for faces labeled as "African." Again, blinding the models did little to change that. Instead, improving the training data seems to be the method that is more reliable when it comes to reducing bias in FR.

6.1 Limitations

Our study comes with certain limitations. The main limitation is the data used for our experiments. We tried to use a well-balanced dataset in terms of ethnic groups and gender groups. We picked the RFW dataset as it consists of approximately equally sized ethnic groups. However,

the dataset does not contain labels for other attributes such as "gender." We therefore had to annotate the dataset ourselves, for which we used a pre-trained classifier. We found that faces labeled as "female," and in particular such that are also labeled as "African," are underrepresented in the RFW dataset. It is thus unclear if our results are representative of these underrepresented groups (e.g., faces labeled as "female" and "African").

Regarding the calculated error rates, we only compared pairs of faces with equal labels, e.g., "female" to "female" and "Indian" to "Indian". We have not tested how likely the models are to make false-positive errors when given two faces with different labels. However, the false-positive rates of confusing faces from group A with faces from group B would be the same as the other way around. Therefore, such an analysis would be unlikely to give us new insights. The interesting analysis is thus mainly the in-group comparison that we focused on.

6.2 Future work

Importantly, considering the potentially harmful consequences of FR, future work should ask in what situations FR technology is inappropriate to use and the question of how misuse of these systems can be prevented, e.g., through policies. Only once these questions have been addressed can advocacy for less-biased FR be beneficial to marginalized communities.

As discussed, the easiest and least problematic way to improve error rates seems to be to improve the data that are used to train the FR models. This means ensuring that sensitive attributes and their intersections are more equally represented. Respective research would then need to confirm that this is actually sufficient. Existing work by, e.g., Buolamwini and Gebru [16], already created a dataset that is diverse in terms of skin color and gender. Future work could continue on this path and ensure diversity along other axes.

Funding Open access funding provided by ZHAW Zurich University of Applied Sciences. No funding was received from external resources besides academic institutions for conducting this study.

Availability of data and materials This research work reports results on a benchmark face recognition dataset that is available for academic research.

Code availability The source code is available and accessible via the following link: https://github.com/samuelwehrli/Face-Recognition-Bias.

Declarations

Conflict of interest The authors have no conflict of interest concerning this research.

References

1. Jain, A.K., Li, S.Z.: Handbook of face recognition, vol. 1. Springer (2011)
2. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. https://doi.org/10.1109/CVPR.2015.7298682 (2015)
3. Deng, J., et al.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
4. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Understand. **189**, 102805 (2019)
5. Stadelmann, T., et al.: Deep learning in the wild. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 17–38. Springer (2018)
6. Smith, D.F., Wiliem, A., Lovell, B.C.: Face recognition on consumer devices: Reflections on replay attacks. IEEE Trans. Inf. Forensics Secur. **10**(4), 736–745 (2015)
7. Robertson, D.J., et al.: Face recognition by metropolitan police superrecognisers. PLoS One **11**(2), e0150036 (2016)
8. Bernal, P.: Data gathering, surveillance and human rights: recasting the debate. J. Cyber Policy **1**(2), 243–264 (2016)
9. Norval, A., Prasopoulou, E.: Public faces? A critical exploration of the diffusion of face recognition technologies in online social networks. N. Media Soc. **19**(4), 637–654 (2017)
10. Mann, M., Smith, M.: Automated facial recognition technology: Recent developments and approaches to oversight. Univ. N. S. W. Law J. **40**(1), 121–145 (2017)
11. Royakkers, L., et al.: Societal and ethical issues of digitization. Ethics Inf. Technol. **20**(2), 127–142 (2018)
12. Learned-Miller, E., et al.: Facial Recognition Technologies in the Wild: A Call for a Federal Office. Tech. rep. Algorithmic Justice League (2020)
13. Harwell, D.: Civil rights groups ask Biden administration to oppose facial recognition. In: The Washington Post. https://www.washingtonpost.com/technology/2021/02/17/facial-recognition-biden/ (2021)
14. van Sant, S., Gonzales, R.: San Francisco approves ban on government's use of Facial recognition technology. In: NPR (2019). https://www.npr.org/2019/05/14/723193785/san-francisco-considers-ban-on-governments-use-of-facial-recognition-technology (2019)

15. Gershgorn, D.: Maine passes the strongest state facial recognition ban yet. In: The Verge. https://www.theverge.com/2021/6/30/22557516/maine-facial-recognition-ban-state-law (2021)
16. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. PMLR, pp. 77–91 (2018)
17. Farzan, A.N.: Sri Lankan police wrongly identify Brown University student as wanted suspect in terror attack. In: The Washington Post. https://www.washingtonpost.com/nation/2019/04/26/sri-lankan-police-wrongly-identify-brown-university-student-wanted-suspect-terror-attack/(2019)
18. Hill, K.: Wrongfully accused by an algorithm. In: The New York Times (2020)
19. John, R.A.: Theory of Justice, 1st edn., ISBN: 0-674-88014–5. Belknap Press of Harvard University Press, Cambridge, Massachussets (1971)
20. Nyarko, J., Goel, S., Sommers, R.: Breaking Taboos in Fair Machine Learning: An Experimental Study. Stanford University, Tech. rep (2020)
21. Žliobaitė, I., Custers, B.: Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artif. Intell. Law **24**(2), 183–201 (2016)
22. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023 (2018)
23. Kleinberg, J., et al.: Algorithmic fairness. In: Aea Papers and Proceedings, vol. 108, pp. 22–27 (2018)
24. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 556–572 (2018)
25. Glüge, S., et al.: How (not) to measure bias in face recognition networks. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 125–137. Springer (2020)
26. Wang, X., Huang, H.: Approaching machine learning fairness through adversarial network. arXiv:1909.03013 (2019)
27. Kim, B., et al.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9012–9020 (2019)
28. Gong, S., Liu, X., Jain, A.K.: Jointly de-biasing face recognition and demographic attribute estimation. In: European Conference on Computer Vision, pp. 330–347. Springer (2020)
29. Mac, R.: Facebook apologizes after A.I. Puts 'primates' label on video of black men. In: The New York Times (2021)
30. Lohr, S.: Facial recognition is accurate, if you're a white guy. In: The New York Times (2018)
31. Raji, I.D., Buolamwini, J.: Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435 (2019)
32. Conger, K., Fausset, R., Kovaleski, S.F.: San Francisco bans facial recognition technology. In: The New York Times, vol. 14 (2019)
33. Lunter, J.: Beating the bias in facial recognition technology. Biom. Technol. Today **2020**(9), 5–7 (2020)
34. Robinson, J.P., et al.: Face recognition: too bias, or not too bias? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–10 (2020)
35. Mehrabi, N., et al.: A survey on bias and fairness in machine learning. In: ACM Computing Surveys (CSUR), vol. 54(6), pp. 1–35 (2021)
36. Khalil, A., et al.: Investigating bias in facial analysis systems: A systematic review. IEEE Access **8**, 130751–130761 (2020)
37. Garcia, R.V., et al.: The harms of demographic bias in deep face recognition research. In: 2019 International Conference on Biometrics (ICB). IEEE, pp. 1–6 (2019)
38. Cavazos, J.G., et al.: Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Trans. Biom. Behav. Identity Sci. **3**(1), 101–111 (2020)
39. Serna, I., et al.: InsideBias: Measuring bias in deep networks and application to face gender biometrics. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 3720–3727 (2021)
40. Kortylewski, A., et al.: Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 2261–2268 (2019)
41. Hooker, S.: Moving beyond algorithmic bias is a data problem. Patterns **2**(4), 100241 (2021)
42. Baosheng, Y., et al.: Correcting the triplet selection bias for triplet loss. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 71–87 (2018)
43. Wang, M., et al.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 692–702 (2019)
44. Depeng, X., et al.: Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE, pp. 570–575 (2018)
45. Yucer, S., et al.: Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 18–19 (2020)
46. Adeli, E., et al.: Representation learning with statistical independence to mitigate bias. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2513–2523 (2021)
47. Wang, M., Deng, W.: Deep face recognition: A survey. In: Neurocomputing, vol. 429, pp. 215–244. ISSN: 0925-2312. https://doi.org/10.1016/j.neucom.2020.10.081. https://www.sciencedirect.com/science/article/pii/S0925231220316945 (2021)
48. Cao, Q., et al.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp. 67–74 (2018)
49. Guo, Y., et al.: Ms-celeb-1m: A dataset and benchmark for largescale face recognition. In: European Conference on Computer Vision, pp. 87–102. Springer (2016)
50. Jie, H., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
51. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
52. Face++ Cognitive Services. https://www.faceplusplus.com (2021)
53. Wikipedia contributors. Freebase (database) – Wikipedia, The Free Ency- clopedia. https://en.wikipedia.org/w/index.php?title=Freebase_(database)&oldid=1035240694 (2021). Accessed 17 Sept 2021
54. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR, pp. 4352–4360 (2017)
55. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. **126**(2–4), 144–157 (2018). https://doi.org/10.1007/s11263-016-0940-3
56. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. In: IEEE Transactions on Information Forensics and Security, vol. 9(12), pp. 2170–2179. ISSN: 1556-6013. https://doi.org/10.1109/TIFS.2014.2359646 (2014)

57. Lee, S.H., et al.: Age and gender estimation using deep residual learning network. In: 2018 International Workshop on Advanced Image Technology (IWAIT). IEEE, pp. 1–3 (2018)
58. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
59. van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)
60. Young, S.G., et al.: Perception and motivation in face recognition: A critical review of theories of the cross-race effect. Pers. Soc. Psychol. Rev. **4**, 116–142 (2021)
61. Maye, D.: Russian face rec suppliers offer ethnicity analytics, raising alarm. In: IPVM. https://ipvm.com/reports/russia-ethnicity-analytics (2021)
62. Harwell, D., Dou, E.: Huawei tested AI software that could recognize Uighur minorities and alert police, report says. In: The Washington Post. https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/ (2020)
63. Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf. Commun. Soc. **22**(7), 900–915 (2019)
64. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol. **33**(4), 659–684 (2020)
65. Durkin, E.: New York tenants fight as landlords embrace facial recognition cameras. In: The Guardian. https://www.theguardian.com/cities/2019/may/29/new-york-facial-recognition-cameras-apartment-complex (2019) Accessed 13 Jul 2021
66. European Commission. Proposal for a Regulation of the European Parliament and of the Council: Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (2021)
67. Veale, M., Borgesius, F.Z.: Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Comput. Law Rev. Int. **22**(4), 97–112 (2021)