Mitigating Gender Bias in Face Recognition Using the von Mises-Fisher Mixture Model

Jean-Rémy Conti [*1 2] **Nathan Noiry** [*1] **Vincent Despiegel** [2] **Stéphane Gentric** [2] **Stéphan Clémençon** [1]

Abstract

In spite of the high performance and reliability of deep learning algorithms in a wide range of everyday applications, many investigations tend to show that a lot of models exhibit biases, discriminating against specific subgroups of the population (*e.g.* gender, ethnicity). This urges the practitioner to develop fair systems with a uniform/comparable performance across sensitive groups. In this work, we investigate the gender bias of deep Face Recognition networks. In order to measure this bias, we introduce two new metrics, BFAR and BFRR, that better reflect the inherent deployment needs of Face Recognition systems. Motivated by geometric considerations, we mitigate gender bias through a new post-processing methodology which transforms the deep embeddings of a pre-trained model to give more representation power to discriminated subgroups. It consists in training a shallow neural network by minimizing a *Fair von Mises-Fisher loss* whose hyperparameters account for the intra-class variance of each gender. Interestingly, we empirically observe that these hyperparameters are correlated with our fairness metrics. In fact, extensive numerical experiments on a variety of datasets show that a careful selection significantly reduces gender bias.

1. Introduction

In the past few years, Face Recognition (FR) systems have reached extremely high levels of performance, paving the way to a broader range of applications, where the reliability levels were previously prohibitive to consider automation. This is mainly due to the adoption of deep learning techniques in computer vision since the famous breakthrough of (Krizhevsky et al., 2012). The increasing use of deep FR systems has however raised concerns as any technological flaw could have a strong societal impact. Besides recent punctual events[1] that received significant media coverage, the academic community has studied the bias of FR systems for many years (dating back at least to (Phillips et al., 2003) who investigated the racial bias of non-deep FR algorithms). In (Abdurrahim et al., 2018) three sources of biases are identified: race (understood as biological attributes such as skin color), age and gender (available gender labels from FR datasets are males and females). The National Institute of Standards and Technology (Grother et al., 2019) conducted a thorough analysis of the performance of several FR algorithms depending on these attributes and revealed high disparities. For instance, some of the top state-of-the-art algorithms in absolute performance have more than seven times more false acceptances for females than for males. In this paper, we introduce a novel methodology to mitigate gender bias for FR. Though focusing on a single source of bias has obvious limitations regarding intersectional effects (Buolamwini & Gebru, 2018), it is a first step to gain insights into the mechanisms at work, before turning to more complex situations. Actually, the method promoted in this paper, much more general than the application considered here, could possibly alleviate many other types of bias. This will be the subject of a future work.

The topic corresponding to the study of different types of bias and to the elaboration of methods to alleviate them is referred to as *fairness* in machine learning, which has received increasing attention in recent years, see *e.g.* (Mehrabi et al., 2019), (Caton & Haas, 2020), (Du et al., 2020). Roughly speaking, achieving fairness means learning a decision rule that does not mistreat some predefined subgroups, while still exhibiting a good predictive performance on the overall population: in general, a trade-off has to be found between fair treatment and pure accuracy[2]. In this regard, one needs

[*]Equal contribution [1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]Idemia. Correspondence to: Jean-Rémy Conti <jean-remy.conti@telecom-paris.fr>, Nathan Noiry <nathan.noiry@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]See for instance the study conducted by the American Civil Liberties Union.

[2]This dichotomy somewhat simplifies the problem since an increase in accuracy could also lead to a better treatment of each subgroup of the population.

to carefully define what will be the relevant *fairness metric*. From a theoretical viewpoint, several ones have been introduced, see *e.g.* (Garg et al., 2020) or (Castelnovo et al., 2021) among others, depending on how the concept of equity of treatment is understood. In practice, these very refined notions can be inadequate, as they ignore specific use case issues, and one thus needs to adapt them carefully. This is particularly the case in FR, where high security standards cannot be negotiated. The goal of this article is twofold: novel fairness metrics, relevant in FR applications in particular, are introduced at length and empirically shown to have room for improvement by means of appropriate/flexible representation models.

Contribution 1. We propose two new metrics, BFAR and BFRR, that incorporate the needs for both security and fairness (see section 2.2). More precisely, the BFAR (resp. BFRR) metric accounts for the disparity between false acceptance (resp. rejection) rates between subgroups of interest, computed at an operating point such that each subgroup has a false acceptance rate lower than a false acceptance level of reference.

It turns out that state-of-the-art FR networks (*e.g.* ArcFace (Deng et al., 2019a)) exhibit poor fairness performance w.r.t. gender, both in terms of BFAR and BFRR. Different strategies could be considered to alleviate this gender bias: pre-, in- and post-processsing methods (Caton & Haas, 2020), depending on whether the practitioner "fairness" intervention occurs before, during or after the training phase. The first one, pre-processing, is not well suited for FR purposes as shown in (Albiero et al., 2020), while the second one, in-processing, has the major drawback to require a full retraining of a deep neural network. This encouraged us to design a post-processing method so as to mitigate gender bias of pre-trained FR models.

In order to improve BFAR and BFRR disparities, we crucially rely on the geometric structure of the last layer of state-of-the-art FR neural networks. The latter is a set of embeddings lying on a *hypersphere*. Those embeddings are obtained through two concurrent mechanisms at work during the learning process: *(i)* repel images of different identities and *(ii)* bring together images of a same identity.

Contribution 2. We set a von Mises-Fisher statistical mixture model on the last layer representation, which corresponds to a mixture of gaussian random variables conditioned to live on the hypersphere. Based on the maximum likelihood of this model, we introduce a new loss we call *Fair von Mises-Fisher*, that we use to supervise the training of a shallow neural network we call *Ethical Module*. Taking the variance parameters as hyperparameters that depend on the gender, this flexible model is able to capture the two previously mentioned mechanisms of repulsion / attraction, which we show are at the origin of the biases in FR. Indeed, our experiments remarkably exhibit a substantial correlation between these hyperparameters and our fairness metrics BFAR and BFRR, suggesting a hidden regularity captured by the model proposed. More precisely, we identify some regions of hyperparameters' values that *(i)* significantly improve BFAR while keeping a reasonable performance but degrading BFRR, *(ii)* significantly improve BFRR while keeping a reasonable performance but degrading BFAR and *(iii)* improve both BFAR and BFRR at the cost of little performance degradation. This third case actually achieves state-of-the-art results in terms of post-processing methods for gender bias mitigation in FR.

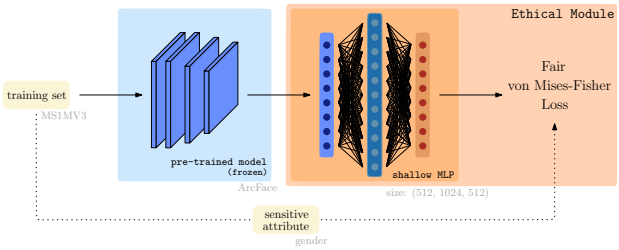


Figure 1: Illustration of the Ethical Module methodology. In gray: our experiment choices.

Besides a simple architecture and a fast training (few hours), the *Ethical Module* enjoys several benefits we would like to highlight.

Taking advantage of foundation models. In the recent survey (Bommasani et al., 2021), the authors judiciously point out a change of paradigm in deep learning: very efficient pre-trained models with billions of parameters they call *foundation models* are at our disposal such as BERT (Devlin et al., 2018) in NLP or ArcFace (Deng et al., 2019a) in FR. Many works rely on these powerful models and fine tune them, inheriting from both their strengths and weaknesses such as their biases. Hence the need to focus on methods to improve the fairness of foundation models: our method is in line with this approach.

No sensitive attribute used during deployment. Though the Ethical Module requires access to the sensitive label during its training phase, this label (*e.g.* gender) is not needed anymore, once the training is completed. This is compliant with the EU jurisdiction that forbids the use of protected attributes for prediction rules.

Organization of the paper. Section 2.1 presents the widely spread usage of FR and its main challenges. It is followed by section 2.2 where we discuss different fairness metrics that arise in FR and introduce two new ones we think are more relevant with regards to operational use cases. In section 3, we present the von Mises-Fisher loss that is used for

the training of the Ethical Module and discuss its benefits. Finally, in section 4, we present at length our numerical experiments, which partly consist in learning an Ethical Module on the ArcFace model, pre-trained on the MS1MV3 dataset (Deng et al., 2019b). Our results show that, remarkably, some specific choices of hyperparameters provide high performance and low fairness metrics both at the same time.

Related works. The correction of bias in FR has been the subject of several recent papers. (Liu et al., 2019) and (Wang & Deng, 2020) use reinforcement learning to learn fair decision rules but despite their mathematical relevance, such methods are computationally prohibitive. Another line of research followed by (Yin et al., 2019), (Wang et al., 2019a) and (Huang et al., 2019) assumes that bias comes from the unbalanced nature of FR datasets and builds on imbalanced and transfer learning methods. Unfortunately, these methods do dot completely remove bias and it has been recently pointed out that balanced dataset are actually *not enough* to mitigate bias, as illustrated by (Albiero et al., 2020) for gender bias, (Gwilliam et al., 2021) for racial bias and (Wang et al., 2019b) for gender bias in face detection. (Gong et al., 2019), (Alasadi et al., 2019) and (Dhar et al., 2021) rely on adversarial methods that can reduce bias but are also known to be unstable and computationally expensive. All of the previously mentioned methods try to learn fair representations. In contrast, some other works do not affect the latent space but modify the decision rule instead: (Terhörst et al., 2020) act on the score function whereas (Salvador et al., 2021) rely on calibration methods. Despite encouraging results, these approaches do not solve the source of the problem which is the bias incurred by the embeddings used.

2. Fairness in Face Recognition

In this section, we first briefly recall the main principles of deep Face Recognition and introduce some notations. The interested reader may consult (Masi et al., 2018) or (Wang & Deng, 2018) for a detailed exposition. Then, we present the fairness metrics we adopt and argue of their relevance in our framework.

2.1. Overview of Face Recognition

Framework. A typical FR dataset consists of face images of individuals from which we wish to predict the identities. Assuming that the images are of size $h \times w$ and that there are K identities among the images, this can be modeled by i.i.d. realizations of a random variable $(X, y) \in \mathbb{R}^{h\times w\times c} \times \{1, \dots, K\}$, where c corresponds to the color channel dimension. In the following, we denote by $\mathbb{P}$ the corresponding probability law.

Objective. The usual goal of FR is to learn an encoder function $f_\theta : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$ that embeds the images in a way to bring same identities closer together. The resulting latent representation $Z := f_\theta(X)$ is the *face embedding* of X. Since the advent of deep learning, the encoder is a deep Convolutional Neural Network (CNN) whose parameters θ are learned on a huge FR dataset $(\boldsymbol{x}_i, y_i)_{1\leq i\leq N}$ made of N i.i.d. realizations of the random variables (X, y). There are generally two FR use cases: *identification*, which consists in finding the specific identity of a probe face among several previously enrolled identities, and *verification* (which we focus on throughout this paper), which aims at deciding whether two face images correspond to the same identity or not. To do so, the closeness between two embeddings is usually quantified with the cosine similarity measure $s(\boldsymbol{z}_i, \boldsymbol{z}_j) := \boldsymbol{z}_i^\intercal \boldsymbol{z}_j/(||\boldsymbol{z}_i|| \cdot ||\boldsymbol{z}_j||)$, where $||\cdot||$ stands for the usual Euclidean norm (the Euclidean metric $||\boldsymbol{z}_i - \boldsymbol{z}_j||$ is also used in some early works *e.g.* (Schroff et al., 2015)). Therefore, an operating point $t \in [-1, 1]$ (threshold of acceptance) has to be chosen to classify a pair $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ as *genuine* (same identity) if $s \geq t$ and *impostor* (distinct identities) otherwise.

Training. For the training phase only, a fully-connected layer is added on top of the deep embeddings so that the output is a K-dimensional vector, predicting the identity of each image within the training set. The full model (CNN + fully-connected layer) is trained as an identity classification task. Until 2018, most of the popular FR loss functions were of the form:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{\kappa \boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} e^{\kappa \boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}\right), \tag{1}$$

where the $\boldsymbol{\mu}_k$'s are the fully-connected layer's parameters, $\kappa > 0$ is the inverse temperature of the softmax function used in brackets and n is the batch size. Early works (Taigman et al., 2014; Sun et al., 2014) took $\kappa = 1$ and used a bias term in the fully-connected layer but (Wang et al., 2017) showed that the bias term degrades the performance of the model. It was thus quickly discarded in later works. Since the canonical similarity measure at the test stage is the cosine similarity, the decision rule only depends on the angle between two embeddings, whereas it could depend on the norms of $\boldsymbol{\mu}_k$ and $\boldsymbol{z}_i$ during training. This has led (Wang et al., 2017) and (Hasnat et al., 2017) to add a normalization step during training and take $\boldsymbol{\mu}_k, \boldsymbol{z}_i \in \mathbb{S}^{d-1} := \{z \in \mathbb{R}^d : ||z|| = 1\}$ as well as introducing the re-scaling parameter κ in Eq. 1: these ideas significantly improved upon former models and are now widely adopted. The hypersphere $\mathbb{S}^{d-1}$ to which the embeddings belong is commonly called face hypersphere. Denoting by θ_i the angle between $\boldsymbol{\mu}_{y_i}$ and $\boldsymbol{z}_i$, the major advance over the loss of Eq. 1 (with normalization of $\boldsymbol{\mu}_k, \boldsymbol{z}_i$) in recent years was to consider large-margin losses which replace

$\boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i = \cos(\theta_i)$ by a function that reduces intra-class angle variations, such as the $\cos(m\theta_i)$ of (Liu et al., 2017) or the $\cos(\theta_i) - m$ of (Wang et al., 2018). The most efficient choice is $\cos(\theta_i + m)$ and is due to (Deng et al., 2019a) who called their model ArcFace, on which we build our methodology. A fine training should result in the alignment of each embedding $\boldsymbol{z}_i$ with the vector $\boldsymbol{\mu}_{y_i}$. The aim is to bring together embeddings with the same identity. Indeed, during the test phase, the learned algorithm will have to decide whether two face images are related to the same, potentially unseen, individual (one refers to an *open set* framework).

Evaluation metrics. Let (X_1, y_1) and (X_2, y_2) be two independent random variables with law $\mathbb{P}$. We distinguish between the False Acceptance and False Rejection Rates, respectively defined by

$$\begin{aligned}\mathrm{FAR}(t) &:= \mathbb{P}(s(Z_1, Z_2) \geq t \mid y_1 \neq y_2)\\ \mathrm{FRR}(t) &:= \mathbb{P}(s(Z_1, Z_2) < t \mid y_1 = y_2)\end{aligned}$$

These quantities are crucial to evaluate a given algorithm in our context: Face Recognition is intrinsically linked to biometric applications, where the usual accuracy evaluation metric is not sufficient to assess the quality of a learned decision rule. For instance, security automation in an airport requires a very low FAR while keeping a reasonable FRR to ensure a pleasant user experience. As a result, the most widely used metric consists in first fixing a threshold t so that the FAR is equal to a pre-defined value $\alpha \in [0, 1]$, and then computing the FRR at this threshold. We use the *canonical FR notation* to denote the resulting quantity:

$$\mathrm{FRR@(FAR} = \alpha) := \mathrm{FRR}(t) \text{ with } \mathrm{FAR}(t) = \alpha.$$

The FAR level α determines the operational point of the FR system and corresponds to the security risk one is ready to take. According to the use case, it is typically set to 10^{-i} with $i \in \{1, \ldots, 6\}$.

2.2. Incorporating Fairness

While the FRR@FAR metric is the standard choice for measuring the performance of a FR algorithm, it does not take into account its variability among different subgroups of the population. In order to assess and correct for potential discriminatory biases, the practitioner must rely on suitable fairness metrics.

Framework. In order to incorporate fairness with respect to a given discrete sensitive attribute that can take $A > 1$ different values, we enrich our previous model and consider a random variable (X, y, a) where $a \in \{0, 1, \ldots, A-1\}$. With a slight abuse of notations, we still denote by $\mathbb{P}$ the corresponding probability law and, for every fixed value a, we can further define

$$\begin{aligned}\mathrm{FAR}_a(t) &:= \mathbb{P}(s(Z_1, Z_2) \geq t \mid y_1 \neq y_2,\ a_1 = a_2 = a)\\ \mathrm{FRR}_a(t) &:= \mathbb{P}(s(Z_1, Z_2) < t \mid y_1 = y_2,\ a_1 = a_2 = a).\end{aligned}$$

In our case, we focus on gender bias so we take $A = 2$ with the convention that $a = 0$ stands for male, $a = 1$ for female.

Existing fairness metrics. Before specifying our choice for the fairness metric used here, let us review some existing ones (Mehrabi et al., 2019) that derive from fairness in the context of binary classification (in FR, one classifies pairs in two groups: genuines or impostors). The *Demographic Parity* criterion requires the prediction to be independent of the sensitive attribute, which amounts to equalizing the likelihood of being genuine conditional to $a = 0$ and $a = 1$. Besides heavily depending on the number and quality of impostors and genuines pairs among subgroups, this criterion does not take into account the FARs and FRRs, which are instrumental in FR as previously mentioned. An attempt to incorporate those criteria could be to compare the intra-group performances: $\mathrm{FRR}_0\mathrm{@(FAR}_0 = \alpha)$ v.s. $\mathrm{FRR}_1\mathrm{@(FAR}_1 = \alpha)$. However, the operational points t_0 and t_1 satisfying $\mathrm{FAR}_0(\mathrm{t}_0) = \alpha$ and $\mathrm{FAR}_1(\mathrm{t}_1) = \alpha$ generically differ as pointed out by (Krishnapriya et al., 2020). To fairly assess the equity of an algorithm, one needs to compare intra-groups FARs and FRRs at the same threshold. Two such criteria exist in the fairness literature: the *Equal Opportunity* fairness criterion which requires $\mathrm{FRR}_0(t) = \mathrm{FRR}_1(t)$ and the *Equalized Odds* criterion which additionally requires $\mathrm{FAR}_0(t) = \mathrm{FAR}_1(t)$. Nevertheless, working at an arbitrary threshold t does not really make sense since, as previously mentioned, FR systems typically choose an operational point achieving a predefined FAR level so as to limit security breaches. This is why most current papers consider a fixed operational point t such that the global population False Acceptance Rate equals a fixed value α. For instance, (Dhar et al., 2021) computes

$$|\mathrm{FRR}_1(t) - \mathrm{FRR}_0(t)| \quad \text{with } \mathrm{FAR}(t) = \alpha. \tag{2}$$

However, we think that the choice of a threshold achieving a global FAR is not entirely relevant for it depends on the relative proportions of females and males of the considered dataset together with the relative proportion of intra-group impostors. For instance, at fixed images quality, if females represent a small proportion of the evaluation dataset, the threshold t of Eq. 2 is close to the male threshold t_0 satisfying $\mathrm{FAR}_0(t_0) = \alpha$ and away from the female threshold t_1 satisfying $\mathrm{FAR}_1(t_1) = \alpha$. Such a variability among datasets could lead to incorrect conclusions.

New fairness metrics. In this paper, we go one step further and work at a threshold achieving $\max_a \mathrm{FAR}_a = \alpha$ instead of $\mathrm{FAR} = \alpha$. This alleviates the previous proportion dependence. Besides, this allows to monitor the risk one is willing to take among each subgroup: for a pre-definite rate α deemed acceptable, one typically would like to compare the performance among subgroups for a threshold where *each* subgroup satisfies $\mathrm{FAR}_a \leq \alpha$. Our two resulting

metrics are thus:

$$\mathrm{BFRR}(\alpha) := \frac{\max_{a\in\{0,1\}} \mathrm{FRR}_a(t)}{\min_{a\in\{0,1\}} \mathrm{FRR}_a(t)} \tag{3}$$

and

$$\mathrm{BFAR}(\alpha) := \frac{\max_{a\in\{0,1\}} \mathrm{FAR}_a(t)}{\min_{a\in\{0,1\}} \mathrm{FAR}_a(t)}, \tag{4}$$

where t is taken such that $\max_{a\in\{0,1\}} \mathrm{FAR}_a(t) = \alpha$.

One can read the above acronyms "Bias in FRR/FAR". In addition to being more security demanding than previous metrics, BFRR and BFAR are more amenable to interpretation: the ratios of FRRs or FARs correspond to the number of times the algorithm makes more mistakes on the discriminated subgroup. Those metrics generalize well for more than 2 distinct values of the sensitive attribute.

3. Geometric Mitigation of Biases

Contrary to a common thinking about the origin of bias, training a FR model on a balanced training set (i.e. with as much female identities/images than male identities/images) is not enough to mitigate gender bias in FR (Albiero et al., 2020). It is therefore necessary to intervene by designing a model to counteract the gender bias.

3.1. A Geometrical Embedding View on Fairness

In fact, impostor scores (cosine similarities of impostor pairs) are higher for females than for males while genuine scores are lower for females than for males (Grother et al., 2019; Robinson et al., 2020). This puts females at a disadvantage compared to males in terms of both FAR and FRR. Typically, this is due to *(i)* a smaller repulsion between female identities and/or *(ii)* a greater intra-class variance (spread of embeddings of each identity) for female identities, as illustrated in Figure 2. Thus, we present in the following a statistical model which enables to set the intra-class variance for each identity on the face hypersphere.

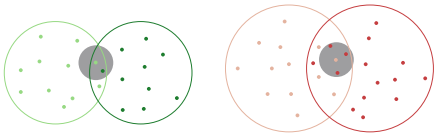


Figure 2: Illustration of the geometric nature of bias. Each point is the embedding of an image. In green: two male identities. In red: two female identities. The overlapping region between two identities is higher for females than for males. The grey circles are the acceptance zones, centered around an embedding of reference, associated to a constant threshold t of acceptance.

3.2. von-Mises Fisher Mixture Model

In order to mitigate the gender bias of deep FR systems, we set a statistical model on the latent representations of images. Recall that we assumed that each individual of a FR dataset is an i.i.d. realization of a random variable (X, y, a), where X is the image, y the identity and a the gender attribute. Also, recall that, both at the training and the testing stages, the embeddings are normalized on the hypersphere, meaning that $Z = f_\theta(X) \in \mathbb{S}^{d-1}$. As previously mentioned, a fine learning should result in an alignment of the embeddings $\{\boldsymbol{z}_i\}$ of a same identity y_i around their associated centroid $\boldsymbol{\mu}_{y_i} \in \mathbb{S}^{d-1}$. It is therefore reasonable to assume that the embeddings of a same identity are i.i.d. realizations of a radial distribution of gaussian-type on the hypersphere, centered at $\boldsymbol{\mu}_{y_i}$. A natural choice is thus to take the so-called von-Mises Fisher (vMF) distribution which is nothing but the law of a gaussian conditioned to live in the hypersphere. Before turning to the formal definition of the statistical model we put on the hypersphere, let us give the definition of this vMF distribution.

The von Mises-Fisher distribution. The vMF distribution in dimension d with mean direction $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and concentration parameter $\kappa > 0$ is a probability measure defined on the hypersphere $\mathbb{S}^{d-1}$ by the following density:

$$V_d(\boldsymbol{z}; \boldsymbol{\mu}, \kappa) := C_d(\kappa) e^{\kappa \boldsymbol{\mu}^\intercal \boldsymbol{z}},$$

with $C_d(\kappa) = \kappa^{\frac{d}{2}-1} / ((2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa))$. I_ν stands for the modified Bessel function of the first kind at order ν, whose logarithm can be computed with high precision (see supplementary material A.1). The vMF distribution corresponds to a gaussian distribution in dimension d with mean $\boldsymbol{\mu}$ and covariance matrix $(1/\kappa)I_d$, conditioned to live on $\mathbb{S}^{d-1}$. Figure 3 illustrates the influence of the concentration parameter κ on the vMF distribution.

Mixture model. Since the vMF distribution seems to reflect well the distribution of the embeddings of 1 identity around their centroid, we extend the model to include all the K identities from the training set by considering a mixture model where each component k ($1 \le k \le K$) is equiprobable and follows a vMF distribution $V_d(\boldsymbol{z}; \boldsymbol{\mu}_k, \kappa_k)$. Figure 4 provides an illustration of the mixture model.

Maximum likelihood. Let $N \ge 1$ and $(\boldsymbol{x}_i, y_i, a_i)_{1\le i\le N}$ be i.i.d. realizations of (X, y, a). Under the previous vMF mixture model assumption, the probability p_{ij} that a face embedding $\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i)$ belongs to identity j is given by

$$p_{ij} = \frac{V_d(\boldsymbol{z}_i|\boldsymbol{\mu}_j, \kappa_j)}{\sum_{k=1}^K V_d(\boldsymbol{z}_i|\boldsymbol{\mu}_k, \kappa_k)} = \frac{C_d(\kappa_j)\, e^{\kappa_j\, \boldsymbol{\mu}_j^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^K C_d(\kappa_k)\, e^{\kappa_k\, \boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}.$$

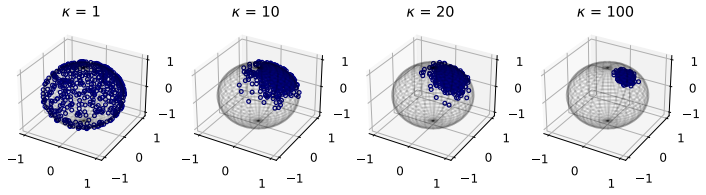


Figure 3: 500 samples from the vMF distribution in dimension 3 with parameters $\boldsymbol{\mu} = [0.5, 0, \sqrt{0.75}]$ and $\kappa > 0$.

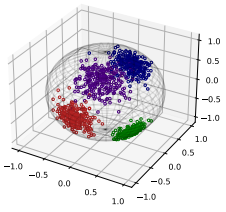


Figure 4: Illustration of a vMF mixture model.

Therefore, the negative log-likelihood of the model is

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{C_d(\kappa_{y_i})\, e^{\kappa_{y_i}\, \boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} C_d(\kappa_k)\, e^{\kappa_k\, \boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}\right]. \quad (5)$$

In that case, the above NLL is in fact the vMF loss function, firstly introduced in the context of FR by (Hasnat et al., 2017) who took a unique hyperparameter value κ. In this situation, the vMF loss reduces to the classical loss of Eq. 1 (when $\boldsymbol{z}_i$ and $\boldsymbol{\mu}_k$ are normalized). This makes the vMF loss a natural generalization of popular FR loss functions, before the advent of large-margin losses. (Zhang et al., 2019a) introduce a similar loss with 2 distinct κ values but they do not take into account the normalization constant $C_d(\kappa)$. (Zhe et al., 2019) use the vMF loss with unique concentration parameter κ for image classification and retrieval but the centroids $\boldsymbol{\mu_k}$ are not learned by gradient descent but rather by an approximate maximum likelihood estimation.

Training an Ethical Module with the vMF-loss. In order to correct for the gender bias contained within the learned latent representation, we train a shallow MultiLayer Perceptron (MLP) which is designed to give more representation power to females. To do so, we slightly modify Eq. 5 by replacing the concentration parameter κ_k of each identity k by a concentration parameter that only depends on the gender $a_k \in \{0, 1\}$ of the identity k. In other words, we replace κ_k by κ_{a_k} and we end up with only 2 concentration parameters ($\kappa_0, \kappa_1 > 0$) that we take as hyperparameters. To sum up, we train the MLP with the following *Fair von Mises-Fisher loss* (FvMF), on batches of size n:

$$\mathcal{L}_{\text{FvMF}} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{C_d(\kappa_{a_{y_i}}) e^{\kappa_{a_{y_i}} \boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} C_d(\kappa_{a_k}) e^{\kappa_{a_k} \boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}\right). \quad (6)$$

Notice that there are two ways of minimizing $\mathcal{L}_{\text{FvMF}}$: either by aligning normalized face embeddings $\boldsymbol{z}_i$ with associated ground-truth $\boldsymbol{\mu}_{y_i}$ (the intra-class variance is characterized by $\kappa_{a_{y_i}}$) or by pushing back wrong $\boldsymbol{\mu}_k$ (with $k \neq y_i$) from $\boldsymbol{z}_i$ (the repulsion strength is related to κ_{a_k}). This brings us back to the two geometric causes for bias in FR, presented in section 3.1. However, those two phenomena are in competition during the loss minimization, especially with two distinct values of concentration parameter, which makes it difficult to predict the optimal values of κ_0 and κ_1.

4. Numerical Experiments

Pre-trained models. We use the trained model ArcFace[3] whose CNN architecture is a ResNet100 (Han et al., 2017). As emphasized before, it achieves state-of-the-art performances in FR. It has been trained on the MS1M-RetinaFace dataset (also called MS1MV3), introduced by (Deng et al., 2019b) in the ICCV 2019 Lightweight Face Recognition Challenge. MS1MV3 is a cleaned version of the MS-Celeb1M dataset (Guo et al., 2016); all its face images have been pre-processed by the Retina-Face detector (Deng et al., 2019c) and are of size 112×112 pixels. It contains 5.1M images of 93k identities. We also consider other pre-trained models[4] (AdaCos (Zhang et al., 2019b), CosFace (Wang et al., 2018), CurricularFace (Huang et al., 2020)) whose backbone is a MobileFaceNet (Chen et al., 2018), trained on the MS-Celeb-1M-v1c-r dataset[5]. This dataset is another cleaned version of the MS-Celeb1M dataset and it contains 3.28M images of 73k identities. The images are also pre-processed by the Retina-Face detector and are of

[3] https://github.com/deepinsight/insightface/tree/master/recognition/arcface_torch.

[4] https://github.com/JDAI-CV/FaceX-Zoo/blob/main/training_mode/README.md.

[5] See footnote 4.

size 112×112 pixels.

Gender labels. For a fair comparison, we train our Ethical Module on the training set used to train the pre-trained models (MS1MV3 for ArcFace, MS-Celeb-1M-v1c-r for the models with MobileFaceNet backbone). However, ground-truth gender labels for MS1MV3/MS-Celeb-1M-v1c-r are not available. As the training of our Ethical Module needs the gender label of each face image within the training set, we use a private gender classifier to get those gender labels. Current gender classifiers achieve around 95% prediction accuracy on standard evaluation datasets and are widely used in FR to get gender annotations (Acien et al., 2018; Gong et al., 2020). Since some images from the same identity might be assigned different gender predictions, it is common practice to use a majority vote to decide the correct gender for each identity. We follow (Albiero et al., 2020) and only keep in our training sets the identities for which at least 75% of the same-identity face images are assigned the same gender. Doing so, we discard 25k images and 835 identities for MS1MV3, 10k images and 500 identities for MS-Celeb-1M-v1c-r.

Ethical Module. The face embeddings output by the pre-trained models are of dimension 512. Thus, the MLP within our Ethical Module has an input layer of 512 units. To emphasize the fact that our gender bias mitigation solution is much less costly than current solutions such as (Wang & Deng, 2020) and (Dhar et al., 2021), in terms of both training time and computation power (see supplementary material A), we choose a shallow MLP of size (512, 1024, 512) with a ReLU activation after the first layer, the output dimension being the same than for the pre-trained models. This MLP is trained with the fair version $\mathcal{L}_{\text{FvMF}}$ of the vMF loss, introduced in Eq. 6. For each experiment, we train the Ethical Module during 50 epochs with the Adam optimizer (Kingma & Ba, 2014). The batch size is set to 1024 and the learning rate to 0.01. The training is efficient as we first compute the face embeddings of the pre-trained models (on MS1MV3 for ArcFace, on MS-Celeb-1M-v1c-r for the models with MobileFaceNet backbone), store them, and then train a shallow MLP on those embeddings. Using one single GPU (NVIDIA RTX 3090), the computation of the embeddings takes 4 hours and each training takes 8 hours.

Reproducibility. We plan to release the code used to conduct our experiments.

4.1. Grid-Search on IJB-C

In order to select relevant pairs of gender-hyperparameters (κ_0, κ_1), we perform a grid-search and keep track of the canonical performance metric $\text{FRR@}(\text{FAR} = 10^{-3})$ together with our two fairness metrics $\text{BFRR}(10^{-3})$ and $\text{BFAR}(10^{-3})$ introduced in Eq. 3 and 4. To obtain reliable results, we need to compute the latter metrics on a sufficiently large FR dataset containing gender labels. We choose IJB-C (Maze et al., 2018), which contains about 3,5k identities for a total number of about 31k images and 117k unconstrained video frames. The 1:1 verification protocol[6] is performed on 19k genuine pairs and 15M impostor pairs. We choose ArcFace ResNet100 as the pre-trained model for this experiment. The results are displayed in Figure 5.

Several interesting trends emerge from Figure 5, suggesting an underlying regularity of the model with respect to the hyperparameters' space. More precisely, when κ_0 is fixed and κ_1 increases, BFAR tends to decrease, BFRR first increases and then decreases and FRR@FAR tends to increase. When κ_1 is fixed and κ_0 increases, BFAR first increases and then decreases, BFRR tends to decrease and FRR@FAR increases. In the supplementary material C, we give some explanations of the trends in Figure 5. Note that BFAR and BFRR have opposite behaviors, which reveals a trade-off between both fairness metrics.

Many (κ_0, κ_1) pairs could be considered as relevant and instead of defining an objective criterion, we select three of them in order to illustrate the trade-offs one needs to make between fairness metrics and pure performance. The selection is made based on Figure 5. We provide in Table 1 the (κ_0, κ_1) pairs which optimize each pair of the considered metrics and give them a name for what follows. The three versions of the Ethical Module presented in Table 1 are robust to a change of FAR level, when performing the grid-search, as illustrated in the supplementary material D.

Table 1: Hyperparameters selected to optimize each pair of metrics. We give a name to each of the (κ_0, κ_1) pairs. EM stands for Ethical Module.

NAME	BFRR	BFAR	FRR@FAR	κ_0	κ_1
EM-FAR	×	√	√	15	20
EM-FRR	√	×	√	25	20
EM-C	√	√	×	45	30

4.2. Fairness Evaluation on LFW

In this section, we evaluate the three versions of our Ethical Module (EM-FAR, EM-FRR, EM-C) and we compare them to the pre-trained model in terms of fairness and performance. All the models are evaluated on the LFW dataset (Huang et al., 2008). The official LFW protocol only considers a few matching pairs among all the possible pairs given the whole LFW dataset. The number of female images is typically not enough to get good estimates of our fairness metrics. To overcome this, we consider all possible same-gender matching pairs among the whole LFW dataset.

[6] https://github.com/deepinsight/insightface/tree/master/recognition/_evaluation_/ijb.

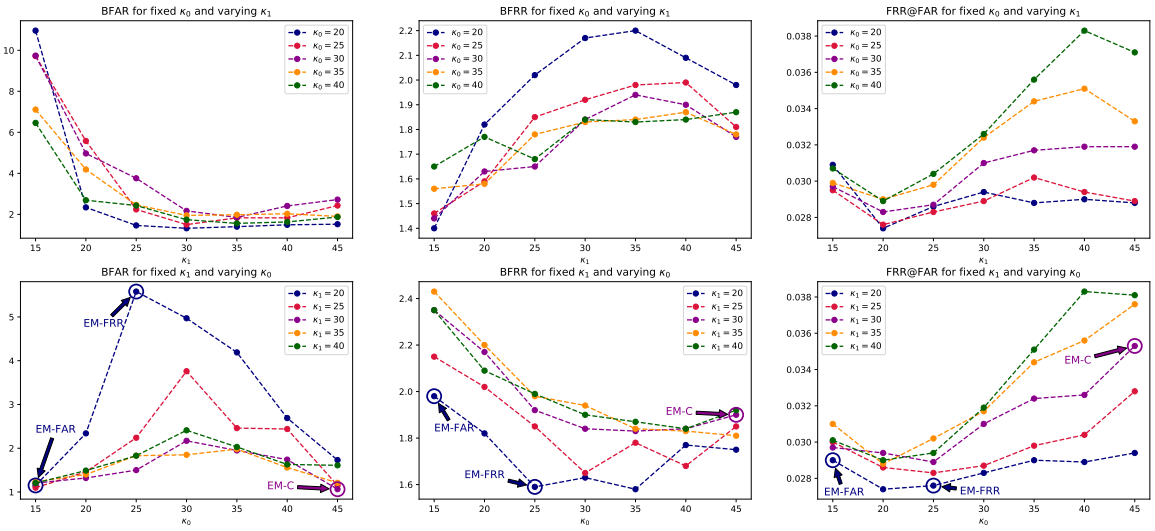


Figure 5: Fairness and evaluation metrics on IJB-C for the Ethical Module when one of the two hyperparameters is fixed. The FAR level defining the threshold t is set to 10^{-3}; the pre-trained model is ArcFace with a ResNet100 backbone. FRR@FAR is expressed as a percentage (%). The three versions of the Ethical Module presented in Table 1 are annotated with circles.

Table 2: Evaluation on LFW for ArcFace with ResNet50 backbone. FRR@FAR is expressed as a percentage (%). **Bold**=Best, Underlined=Second best.

FAR LEVEL:	10^{-4}			10^{-3}		
MODEL	FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
ARCFACE	**0.078**	10.27	4.72	0.059	4.17	1.81
ARCFACE + PASS-G	0.315	**4.54**	6.51	0.107	5.22	2.11
ARCFACE + EM-FAR	0.151	11.22	**2.11**	0.072	9.16	**1.19**
ARCFACE + EM-FRR	0.100	5.89	33.65	**0.058**	**4.11**	5.24
ARCFACE + EM-C	0.164	9.18	2.44	0.081	5.15	1.20

Doing so, we obtain 9.8k female genuine pairs, 232k male genuine pairs, 4.4M female impostor pairs and 52M male impostor pairs.

Baseline. The current state-of-the-art post-processing method for gender bias mitigation of FR models is achieved by PASS-g (Dhar et al., 2021). It also consists in transforming the embeddings output by the pre-trained model but it is trained in an adversarial way to classify identities and simultaneously reduce encoding of gender within the new embeddings. Although attempting to output embeddings that are independent from gender seems a good idea, we believe that the gender information contained within the embeddings helps any FR model a lot at the training stage (identity classification), and thus that such a training cannot be achieved without losing too much performance.

We first verify the effectiveness of our Ethical Module using the pre-trained model ArcFace ResNet50. For a fair comparison, we train PASS-g on the same training set than the Ethical Module (MS1MV3 in this case). In Table 2, we summarize the different metrics evaluated for the three versions of our Ethical Module on the LFW dataset and compare them with the pre-trained ArcFace and PASS-g baselines, at two FAR levels. EM-FAR achieves the best BFAR at both FAR levels while the best BFRR is obtained by EM-FRR at FAR $= 10^{-3}$ and by PASS-g at FAR $= 10^{-4}$. At the latter FAR level, the error rate FRR@FAR of PASS-g is slightly more than 4 times the error rate of the original pre-trained model. Finally, EM-C is the only model which succeeds in reducing both fairness metrics (BFRR and BFAR) of the pre-trained model at the same time for FAR $= 10^{-4}$.

In addition, we check the robustness of our method to a change of pre-trained model by considering competitive FR loss functions (AdaCos, CosFace, CurricularFace) with MobileFaceNet backbone. The results are displayed in Table 3. Additional results with other pre-trained models are available in the supplementary material G.

Table 3: Evaluation on LFW for different pre-trained models (AdaCos, CosFace, CurricularFace) with MobileFaceNet backbone. By "Original" we mean no Ethical Module is added to the pre-trained model. FRR@FAR is expressed as a percentage (%). **Bold**=Best, Underlined=Second best.

FAR LEVEL:	10^{-4}			10^{-3}		
MODEL	FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
ADACOS						
ORIGINAL	**2.97**	3.64	3.84	0.98	5.29	2.23
EM-FAR	4.56	4.42	**1.41**	1.33	6.34	**1.01**
EM-FRR	3.12	**2.71**	8.37	**0.91**	**4.23**	3.71
EM-C	4.05	4.51	1.57	1.26	7.28	1.08
COSFACE						
ORIGINAL	**1.73**	5.89	2.51	**0.58**	8.18	1.74
EM-FAR	3.69	5.76	**1.13**	1.05	8.41	**1.02**
EM-FRR	2.41	**3.03**	9.66	0.67	**5.09**	4.75
EM-C	2.60	4.30	3.69	0.82	6.81	1.87
CURRICULARFACE						
ORIGINAL	**2.52**	3.67	2.92	**0.81**	4.88	1.91
EM-FAR	3.86	5.26	**1.16**	1.17	6.35	**1.10**
EM-FRR	2.82	**2.58**	9.10	0.82	**3.89**	4.28
EM-C	3.61	3.40	2.30	1.02	5.63	1.27

5. Conclusion

In this paper, we introduce a novel method, the *Ethical Module*, to mitigate the gender bias of Face Recognition state-of-the-art models. It consists in learning a shallow MLP on top of a frozen pre-trained model, so as to correct the biases that exist in the embedding space. To achieve fairness, we rely on a fair version of the von Mises-Fisher loss that incorporates an hyperparameter per gender, related to the intra-class variance of each gender. Measuring the fairness of Face Recognition systems is a very challenging task and we introduce two new metrics, BFAR and BFRR, that both respond to the need for security and equity.

Besides being very simple, the resulting methodology is more stable and faster than most current methods of bias mitigation. It both leverages the strong accuracy of pre-trained models while correcting their bias. We illustrate the soundness of our methodology on several pre-trained models, and strongly believe it could also be used to alleviate other types of bias. Our work opens several lines of research: for instance, it would be interesting to extend our ideas to the context of multiclass sensitive attributes and of continuous sensitive attributes such as age. Another idea would be to somehow incorporate our fairness criteria during the training of the Ethical Module. Finally, we think that incorporating large-margin constraints into the loss used to train the Ethical Module would be a promising attempt to go beyond the trade-off between fairness and performance.

Acknowledgments

This research was partially supported by the French National Research Agency (ANR), under grant ANR-20-CE23-0028 (LIMPID project).

References

Abdurrahim, S. H., Samad, S. A., and Huddin, A. B. Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34(11): 1617–1630, 2018.

Acien, A., Morales, A., Vera-Rodriguez, R., Bartolome, I., and Fierrez, J. Measuring the gender and ethnicity bias in deep models for face recognition. In *Iberoamerican Congress on Pattern Recognition*, pp. 584–593. Springer, 2018.

Alasadi, J., Al Hilli, A., and Singh, V. K. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, pp. 19–25, 2019.

Albiero, V., Zhang, K., and Bowyer, K. W. How does gender balance in training data affect face recognition accuracy? *arXiv preprint arXiv:2002.02934*, 2020.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Castelnovo, A., Crupi, R., Greco, G., and Regoli, D. The zoo of fairness metrics in machine learning. *arXiv preprint arXiv:2106.00467*, 2021.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Chen, S., Liu, Y., Gao, X., and Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, 2018.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019a.

Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., and Shi, S. Lightweight face recognition challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2638–2646, 2019b. doi: 10.1109/ICCVW.2019.00322.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019c.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., and Chellappa, R. PASS: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15087–15096, October 2021.

Du, M., Yang, F., Zou, N., and Hu, X. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 2020.

Garg, P., Villasenor, J., and Foggo, V. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3662–3666. IEEE, 2020.

Gong, S., Liu, X., and Jain, A. K. Jointly de-biasing face recognition and demographic attribute estimation. *arXiv preprint arXiv:1911.08080*, 2019.

Gong, S., Liu, X., and Jain, A. K. Mitigating face recognition bias via group adaptive classifier. *arXiv preprint arXiv:2006.07576*, 2020.

Grother, P., Ngan, M., and Hanaoka, K. Ongoing face recognition vendor test (frvt) part 3: Demographic effects. *National Institute of Standards and Technology, Tech. Rep. NISTIR*, 8280, 2019.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pp. 87–102. Springer, 2016.

Gwilliam, M., Hegde, S., Tinubu, L., and Hanson, A. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4123–4132, 2021.

Han, D., Kim, J., and Kim, J. Deep pyramidal residual networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.668. URL http://dx.doi.org/10.1109/CVPR.2017.668.

Hasnat, M., Bohné, J., Milgram, J., Gentric, S., Chen, L., et al. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794, 2019.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2020.

Johansson, F. et al. *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 1.2.0)*, February 2021. http://mpmath.org/.

Kim, M. On pytorch implementation of density estimators for von mises-fisher and its mixture. *arXiv preprint arXiv:2102.05340*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krishnapriya, K., Albiero, V., Vangara, K., King, M. C., and Bowyer, K. W. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Liu, B., Deng, W., Zhong, Y., Wang, M., Hu, J., Tao, X., and Huang, Y. Fair loss: margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 10052–10061, 2019.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Masi, I., Wu, Y., Hassner, T., and Natarajan, P. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pp. 471–478. IEEE, 2018.

Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., Jain, A. K., Niggel, W. T., Anderson, J., Cheney, J., et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pp. 158–165. IEEE, 2018.

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

Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., and Bone, M. Face recognition vendor test 2002. In *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, pp. 44. IEEE, 2003.

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., and Timoner, S. Face recognition: Too bias, or not too bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–10, 2020. doi: 10.1109/CVPRW50498.2020.00008.

Salvador, T., Cairns, S., Voleti, V., Marshall, N., and Oberman, A. Bias mitigation of face recognition models through calibration. *arXiv preprint arXiv:2106.03761*, 2021.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Sun, Y., Wang, X., and Tang, X. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1891–1898, 2014.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F., and Kuijper, A. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020.

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Wang, M. and Deng, W. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

Wang, M. and Deng, W. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9322–9331, 2020.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 692–702, 2019a.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5310–5319, 2019b.

Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A. K., Duncan, J. A., Allen, K., et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 90–98, 2017.

Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. Feature transfer learning for face recognition with underrepresented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5704–5713, 2019.

Zhang, J., Guo, Q., Dong, Y., Xiong, F., and Bai, H. Adaptive parameters softmax loss for deep face recognition. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 1680–1684. IEEE, 2019a.

Zhang, X., Zhao, R., Qiao, Y., Wang, X., and Li, H. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10823–10832, 2019b.

Zhe, X., Chen, S., and Yan, H. Directional statistics-based deep metric learning for image classification and retrieval. *Pattern Recognition*, 93:113–123, 2019.

A. Numerical stability

A.1. von Mises-Fisher constants

Recall the loss defined in Equation 6:

$$\mathcal{L}_{\text{FvMF}} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{C_d(\kappa_{a_{y_i}})e^{\kappa_{a_{y_i}}\boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} C_d(\kappa_{a_k})e^{\kappa_{a_k}\boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}\right) \quad \text{with} \quad C_d(\kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)}.$$

I_ν stands for the modified Bessel function of the first kind at order ν, whose logarithm can be computed with high precision using a Python library for arbitrary-precision floating-point arithmetic such as `mpmath` (Johansson et al., 2021; Kim, 2021).

Once $\log(I_{\frac{d}{2}-1}(\kappa))$ is obtained, one is able to compute the logarithm of $C_d(\kappa)$ as:

$$\log(C_d(\kappa)) = (\frac{d}{2}-1)\log(\kappa) - \frac{d}{2}\log(2\pi) - \log(I_{\frac{d}{2}-1}(\kappa)).$$

Figure 6 displays $\log(I_{\frac{d}{2}-1}(\kappa))$ and $\log(C_d(\kappa))$ as functions of κ for $d = 512$.

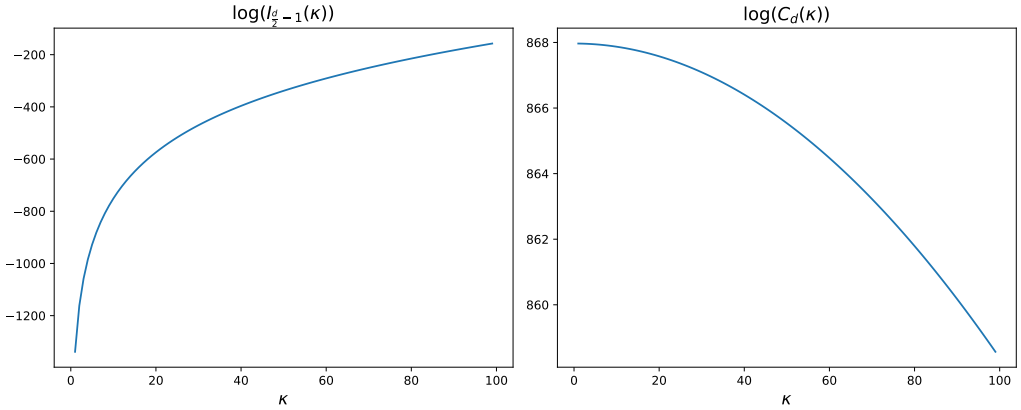


Figure 6: $\log(I_{\frac{d}{2}-1}(\kappa))$ and $\log(C_d(\kappa))$ as functions of κ for $d = 512$.

A.2. Loss stability

To make use of the numerical stability of the quantity $\log(C_d(\kappa))$, $\mathcal{L}_{\text{FvMF}}$ can be written as:

$$\mathcal{L}_{\text{FvMF}} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\log(C_d(\kappa_{a_{y_i}}))+\kappa_{a_{y_i}}\boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} e^{\log(C_d(\kappa_{a_k}))+\kappa_{a_k}\boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}}\right).$$

Recall the cross-entropy loss $\mathcal{L}_{CE}(\{q_{i,k}\},\{y_i\})$ with $1 \le i \le n$ and $1 \le k \le K$ defined as:

$$\mathcal{L}_{CE}(\{q_{i,k}\},\{y_i\}) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{q_{i,y_i}}}{\sum_{k=1}^{K} e^{q_{i,k}}}\right)$$

$\mathcal{L}_{\text{FvMF}}$ can be expressed as the cross-entropy loss:

$$\mathcal{L}_{\text{FvMF}} = \mathcal{L}_{CE}(\{q_{i,k}\},\{y_i\})$$

where the logits $q_{i,k} = \log(C_d(\kappa_{a_k})) + \kappa_{a_k}\boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i$ satisfy ($\boldsymbol{\mu}_k$, $\boldsymbol{z}_i \in \mathbb{S}^{d-1}$):

$$\log(C_d(\kappa_{a_k})) - \kappa_{a_k} \le q_{i,k} \le \log(C_d(\kappa_{a_k})) + \kappa_{a_k}$$

Those bounds are displayed in Figure 7. The fact that $\mathcal{L}_{\text{FvMF}}$ can be expressed as the cross-entropy loss makes it possible to use the logsoftmax trick and thus further increases its numerical stability.

We provide on Figure 8 the behavior of our $\mathcal{L}_{\text{FvMF}}$ training loss, used to train the Ethical Module on top of ArcFace ResNet50.

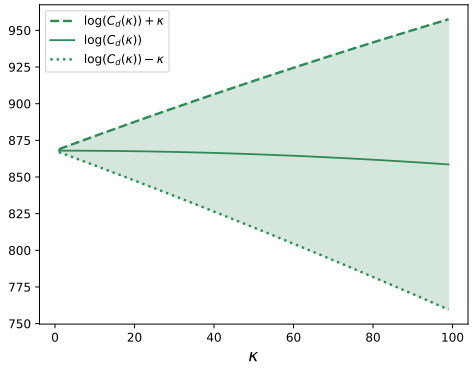


Figure 7: Range of values of the $\mathcal{L}_{\text{FvMF}}$ loss logits for $d = 512$.

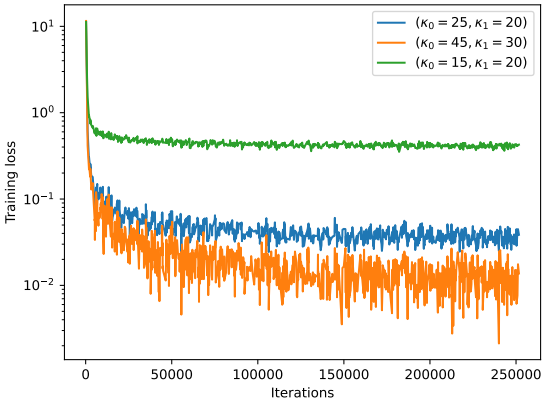


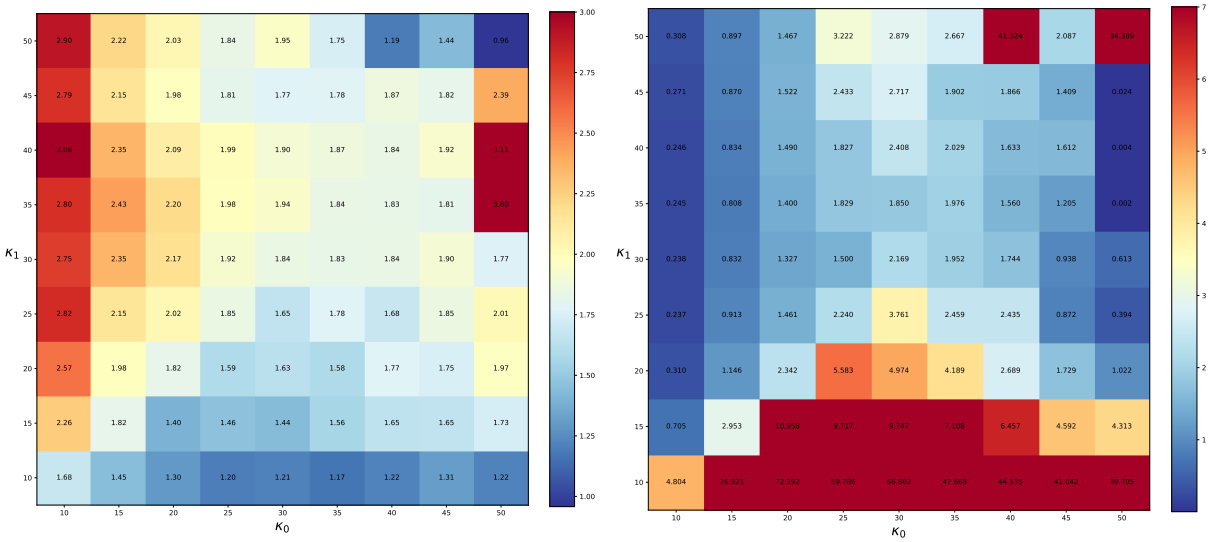
Figure 8: $\mathcal{L}_{\text{FvMF}}$ training loss for the pre-trained model ArcFace ResNet50.

B. Grid-search on IJB-C

In order to select relevant pairs of gender-hyperparameters (κ_0, κ_1), we perform a grid-search on a square of size 9×9 and keep track of the canonical performance metric $\text{FRR@}(\text{FAR} = 10^{-3})$ together with variants of our two fairness metrics $\text{BFRR}(10^{-3})$ and $\text{BFAR}(10^{-3})$ introduced in Eq. 3 and 4. These variants are respectively $\text{FRR}_1(t)/\text{FRR}_0(t)$ and $\text{FAR}_1(t)/\text{FAR}_0(t)$ computed at the threshold t satisfying $\max_{a\in\{0,1\}} \text{FAR}_a(t) = 10^{-3}$. In this way we can visualize the inversion of bias incurred by our model: in most settings, females are disadvantaged while some extreme values of the hyperparameters disadvantage males. The results displayed in Figure 9 contain the results of Figure 5 but they are more complete.

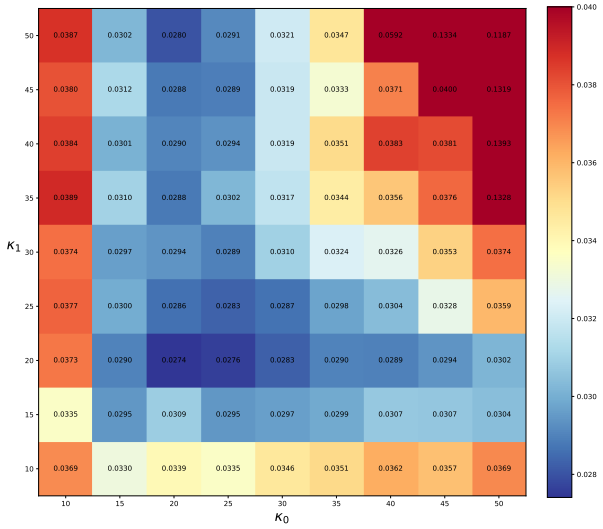
C. Trends in Figure 5

Recall that the vMF parametric interpretation of the model is that each identity is associated with a gaussian on the sphere with fixed mean and fixed concentration parameter. The images of a dataset are then seen as i.i.d. realization of the mixture of these gaussians and the loss consists in maximizing the log-likelihood. In order to control the representation power of



(a) $\mathrm{FRR}_1(t)/\mathrm{FRR}_0(t)$

(b) $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(t)$



(c) $\mathrm{FRR@}(\mathrm{FAR} = 10^{-3})$

Figure 9: Three metrics along the grid-search. Notice that 9(a) and 9(b) are computed at the threshold t satisfying $\max_{a\in\{0,1\}} \mathrm{FAR}_a(t) = 10^{-3}$. The pre-trained model is ArcFace with a ResNet100 backbone and the Ethical Module is evaluated on IJB-C.

males and females, we fix a concentration parameter κ_0 (resp. κ_1) for all males (resp. females). In Figure 5, we observe that the different metrics exhibit smooth behavior with respect to these hyperparameters. Let us give some insights on these phenomenons. In general, female are discriminated against so that the maximum is realized for FAR_1: we will always place ourselves in this situation for the following heuristic reasoning, meaning that we will always assume that

$$\max(\mathrm{FAR}_0(t), \mathrm{FAR}_1(t)) = \mathrm{FAR}_1(t). \tag{7}$$

Therefore, our heuristic will not take into account the observed empirical fact that, for some specific choices of hyperparmeters, male are discriminated against. We think one could push further the reasoning to include this case but restrict the scope of our explanations in order to focus on the underlying mechanisms of the vMF loss.

Restriction to the study of $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$**.** We claim it is sufficient to focus on the evolution of $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$, from which the evolution of $\mathrm{FRR}_1(t)/\mathrm{FRR}_0(\mathrm{t})$ can be deduced, at least at the heuristic level developed here. Two cases may occur:

- If $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$ increases, it means that there are more False Acceptance among females. From a geometric viewpoint, this means that females are more spread around than males. Therefore, there will be more False Reject among males who are more concentrated. Thus, when $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$ increases, $\mathrm{FRR}_1(t)/\mathrm{FRR}_0(\mathrm{t})$ decreases.
- If $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$ decreases, it means that there are less False Acceptance among females. From a geometric viewpoint, this means that females are more concentrated than males. Therefore, there will be less False Reject among males who are less concentrated. Thus, when $\mathrm{FAR}_1(t)/\mathrm{FAR}_0(\mathrm{t})$ decreases, $\mathrm{FRR}_1(t)/\mathrm{FRR}_0(\mathrm{t})$ increases.

These two observations are confirmed by the graphical representations of Figure 5.

Suppose that κ_1 is increased by a small amount $\Delta\kappa_1$ while κ_0 remains fixed.
We will denote by $\mathrm{FAR}_a^{\kappa_1}$ the False Acceptance Rate curve of subgroup a for the hyperparameters choice (κ_0, κ_1) and by $\mathrm{FAR}_a^{\kappa_1+\Delta\kappa_1}$ the False Acceptance Rate curve of subgroup a for the hyperparameters choice $(\kappa_0, \kappa_1 + \Delta\kappa_1)$.
The representation with hyperparameters $(\kappa_0, \kappa_1 + \Delta\kappa_1)$ increases the concentration parameter of females. As a result, the images stemming from a same female identity should be closer from one another, leading to a better FAR performance. Therefore, one should have:

$$\forall t \in [0,1], \quad \mathrm{FAR}_1^{\kappa_1+\Delta\kappa_1}(t) < \mathrm{FAR}_1^{\kappa_1}(t). \tag{8}$$

Let us denote by t_{κ_1} and $t_{\kappa_1+\Delta\kappa_1}$ the points satisfying:

$$\mathrm{FAR}_1^{\kappa_1}(t_{\kappa_1}) = \alpha \quad \text{and} \quad \mathrm{FAR}_1^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1}) = \alpha.$$

Using Equation 7 and Equation 8, this implies that $t_{\kappa_1+\Delta\kappa_1} < t_{\kappa_1}$, as illustrated in Figure 10.

We can now distinguish two situations depending on the magnitude of κ_1.

- If κ_1 is small, its variation does not affect the representation of males at least at a first order approximation. In that case $\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1}) = \mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})$. Since $\mathrm{FAR}_0^{\kappa_1}$ is nonincreasing, we deduce that $\mathrm{FAR}_0(t_{\kappa_1+\Delta\kappa_1}) > \mathrm{FAR}_0(t_{\kappa_1})$, which finally implies that:

$$\frac{\mathrm{FAR}_1^{\kappa_1}(t_{\kappa_1})}{\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1})} = \frac{\alpha}{\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1})} > \frac{\alpha}{\mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})} = \frac{\mathrm{FAR}_1^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})}{\mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})}.$$

- If κ_1 is large enough, tightening the representations of females among themselves starts to affect the males representation. Indeed, they enjoy more space and can therefore be better spread, which implies that $\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1}) > \mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})$, as illustrated in Figure 10 (b). As a result:

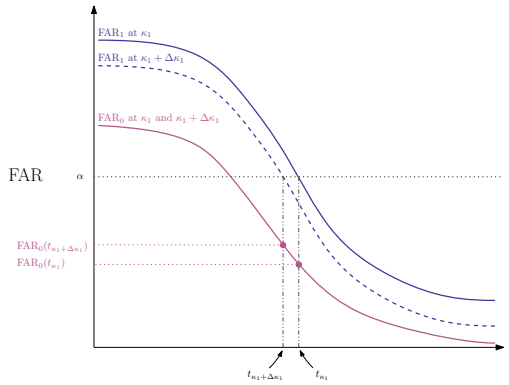
$$\frac{\mathrm{FAR}_1^{\kappa_1}(t_{\kappa_1})}{\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1})} = \frac{\alpha}{\mathrm{FAR}_0^{\kappa_1}(t_{\kappa_1})} < \frac{\alpha}{\mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})} = \frac{\mathrm{FAR}_1^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})}{\mathrm{FAR}_0^{\kappa_1+\Delta\kappa_1}(t_{\kappa_1+\Delta\kappa_1})}.$$

The two previous points are confirmed by the top left corner graphical representation of Figure 5: for all fixed values of κ_0, the curves start by decreasing when κ_1 increases, then begin an increasing phase when κ_1 becomes sufficiently large.

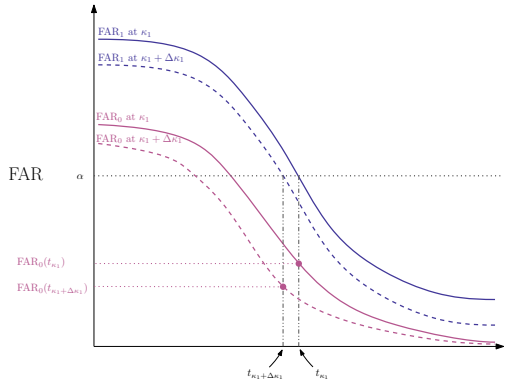
Suppose that κ_0 is increased by a small amount $\Delta\kappa_0$ while κ_0 remains fixed.
We will denote by $\mathrm{FAR}_a^{\kappa_0}$ the False Acceptance Rate curve of subgroup a for the hyperparameters choice (κ_0, κ_1) and by $\mathrm{FAR}_a^{\kappa_0+\Delta\kappa_0}$ the False Acceptance Rate curve of subgroup a for the hyperparameters choice $(\kappa_0 + \Delta\kappa_0, \kappa_1)$.
The representation with hyperparameters $(\kappa_0 + \Delta\kappa_0, \kappa_1)$ increases the concentration parameter of males. As a result, the images stemming from a same male identity should be closer from one another, leading to a better FAR performance. Therefore, one should have:

$$\forall t \in [0,1], \quad \mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t) < \mathrm{FAR}_0^{\kappa_0}(t). \tag{9}$$

As before, we can distinguish two situations depending on the magnitude of κ_0.



(a) When κ_1 is small, the FAR_0 curve is not perturbed.



(b) When κ_1 is large enough, the FAR_0 curve starts to be perturbed.

Figure 10: Heuristic explanation of the $\mathrm{FAR}_a(t)$ evolution at fixed κ_0 and when κ_1 increases.

- If κ_0 is small, one can suppose that females are not affected by its variation, meaning that $\mathrm{FAR}_1^{\kappa_0} = \mathrm{FAR}_1^{\kappa_0+\Delta\kappa_0}$ at a first order approximation (see (a) of Figure 11 for an illustration). In that case, $t_{\kappa_0} = t_{\kappa_0+\Delta\kappa_0}$, and Equation 9 implies that $\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0}) < \mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})$. As a result:

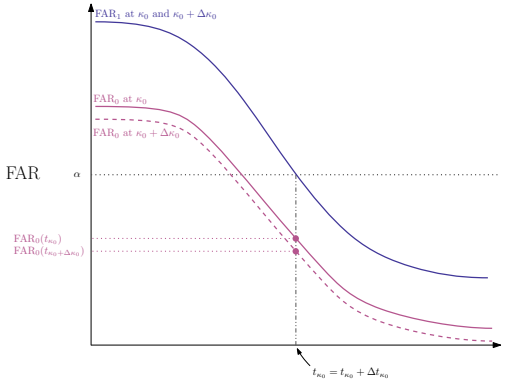
$$\frac{\mathrm{FAR}_1^{\kappa_0}(t_{\kappa_0})}{\mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})} = \frac{\alpha}{\mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})} < \frac{\alpha}{\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})} = \frac{\mathrm{FAR}_1^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})}{\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})}.$$

- If κ_0 is large enough, tightening the representations of males among themselves starts to affect the females representation: they have more space to spread around (see (b) of Figure 11). As a result, one can have $\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0}) > \mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})$

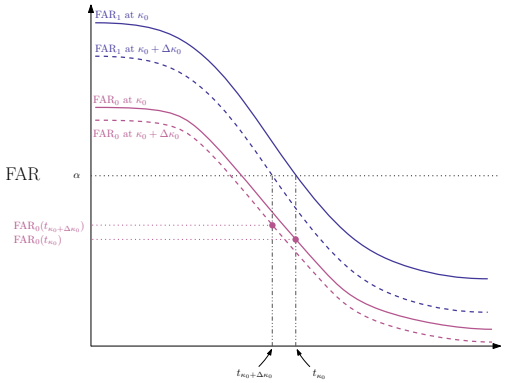
$$\frac{\mathrm{FAR}_1^{\kappa_0}(t_{\kappa_0})}{\mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})} = \frac{\alpha}{\mathrm{FAR}_0^{\kappa_0}(t_{\kappa_0})} > \frac{\alpha}{\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})} = \frac{\mathrm{FAR}_1^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})}{\mathrm{FAR}_0^{\kappa_0+\Delta\kappa_0}(t_{\kappa_0+\Delta\kappa_0})}.$$

D. Robustness of the selected hyperparameters

The grid-search, presented in Figure 5, is performed using the IJB-C dataset, at a FAR level equal to 10^{-3}. The three versions of the Ethical Module presented in Table 1 are found, based on this grid-search. One relevant issue about this method is the robustness of the three selected hyperparameters, typically when performing the grid-search at a different FAR level. Figure 12 displays the same grid-search on IJB-C than Figure 5, but at a FAR level equal to 10^{-4}. The three versions of the Ethical Module are robust to a change of FAR level on the validation set.



(a) When κ_1 does not increase to much, the FAR_0 curve is not perturbed.



(b) When κ_1 is large enough, the FAR_0 curve starts to be perturbed.

Figure 11: Heuristic explanation of the $\mathrm{FAR}_a(t)$ evolution at fixed κ_0 and when κ_1 increases.

E. Comparison of the spread of embeddings between genders

The Ethical Module has hyperparameters which partly control the intra-class variance per gender. Our solution EM-FRR significantly reduces the BFRR metric: at any given operating point t, we should thus have $\mathrm{FRR}_0(t) \sim \mathrm{FRR}_1(t)$. This could be understood as follows: genuine male images are as spread around their centroid than genuine female images are around their own centroid. We would like to check whether this phenomenon occurs.

Since the training phase is an iterative process, the centroids might not represent the exact center of each identity within the hypersphere. We choose to compute the center of a given identity by the mean of the embeddings that form this identity, renormalized to lie on the hypersphere.

To measure the variability of the embeddings $z_1, \ldots, z_n$ of a given identity, we first compute the empirical mean $\overline{z} := (1/n)\sum_{1\leq i\leq n} z_i$, and renormalize it on the hypersphere: $\mathfrak{z} = \overline{z}/||\overline{z}||_2$. Then, we compute the inertia of $z_1, \ldots, z_n$ with respect to $\mathfrak{z}$:

$$I := \frac{1}{n}\sum_{i=1}^{n} ||z_i - \mathfrak{z}||_2^2.$$

We use the pre-trained model ArcFace ResNet100 for this experiment; trained on MS1MV3 as our EM-FRR. In order to have good estimates of this spread measure, we only considered identities having at least 100 images within the training set (MS1MV3). We end up with 3569 male identities and 5045 female identities. For each of those identities, we compute the spread measure and get an histogram of those values per gender (one histogram for male identities, another for female

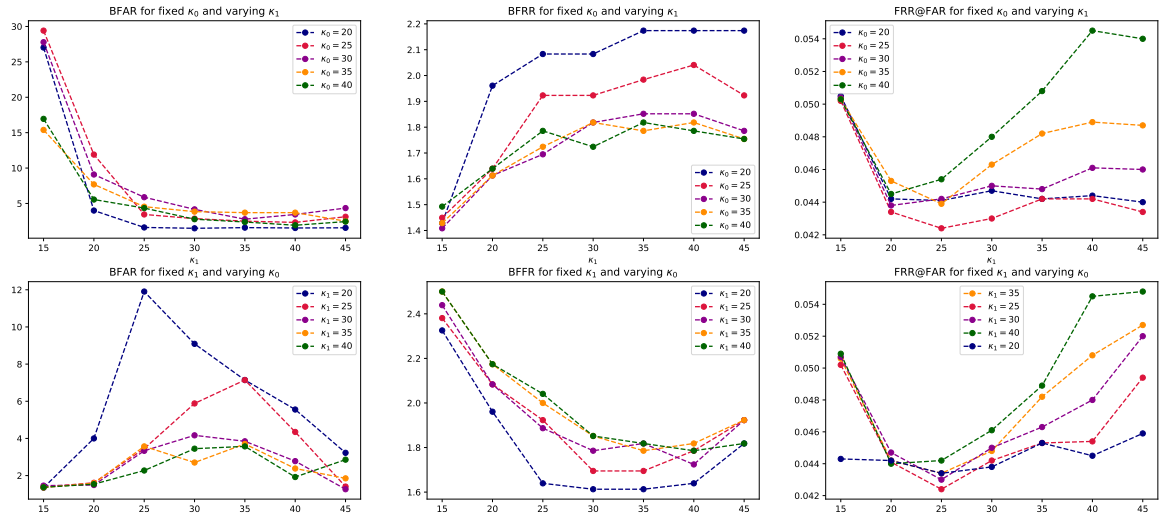
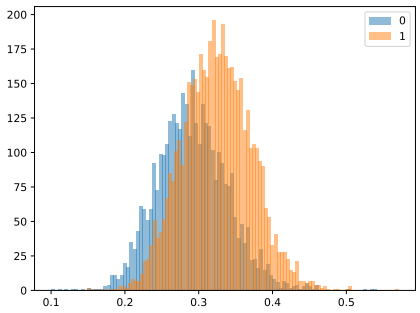


Figure 12: Fairness and evaluation metrics on IJB-C for the Ethical Module when one of the two hyperparameters is fixed. The FAR level defining the threshold t is set to 10^{-4}; the pre-trained model is ArcFace with a ResNet100 backbone. FRR@FAR is expressed as a percentage (%). The three versions of the Ethical Module presented in Table 1 are robust to a change of FAR level, when performing the grid-search.

identities).

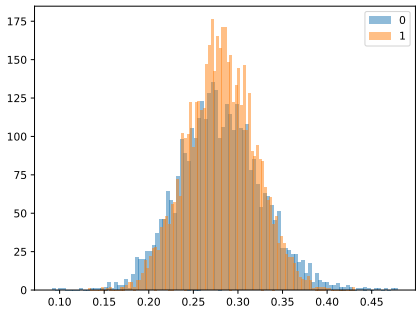
F. Influence of κ_1 on $\mathrm{FAR}_a(t)$ and $\mathrm{FRR}_a(t)$

After training our Ethical Module with ArcFace ResNet100 as the pre-trained model, we evaluate it on the LFW dataset and compute the quantities $\mathrm{FAR}_a(t)$ (Figure 15) and $\mathrm{FRR}_a(t)$ (Figure 16) with varying κ_1. This shows that our vMF mixture statistical model has a clear impact on the representation of deep embeddings.



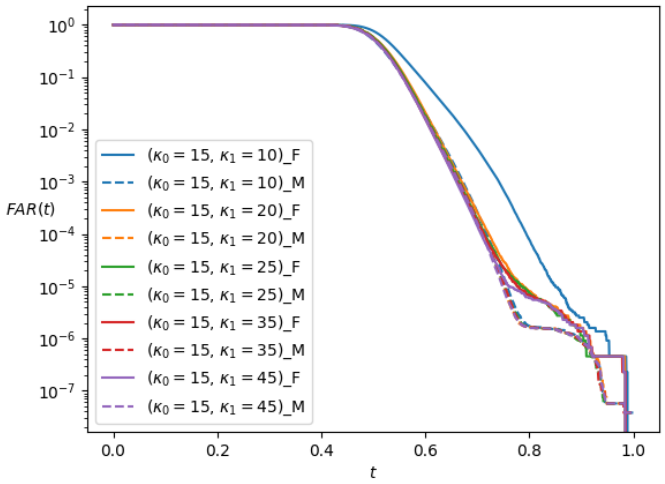
(a) ArcFace

Figure 13: Histograms of identities inertias. In orange: for females. In blue: for males. For the pretrained model, the two histograms are not aligned.

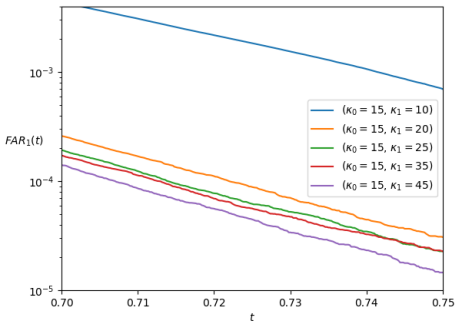


(a) EM-FRR

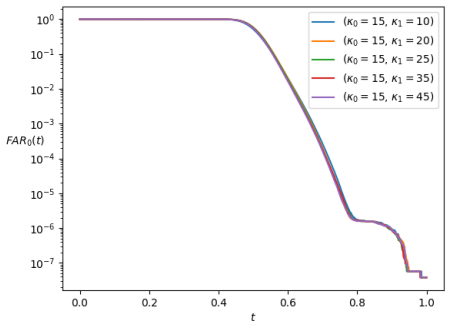
Figure 14: Histograms of identities inertias. In orange: for females. In blue: for males. For our EM-FRR model, the two histograms are aligned.



(a) Influence of κ_1 on $\mathrm{FAR}_a(t)$. Females are depicted with solid lines while males are depicted with dashed lines.

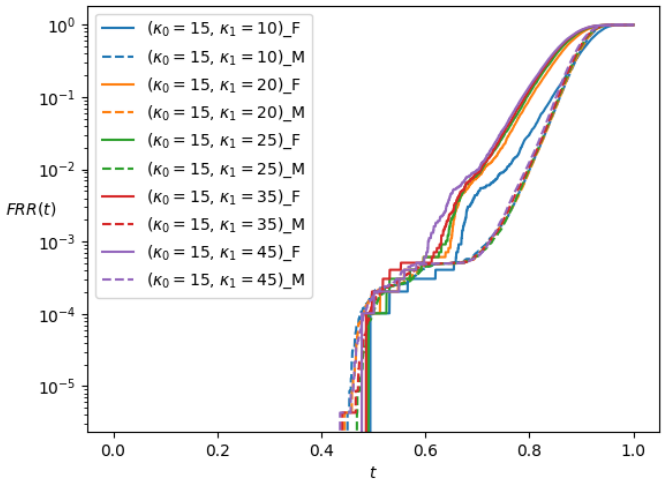


(b) Zoom on Figure 15(a) for females.

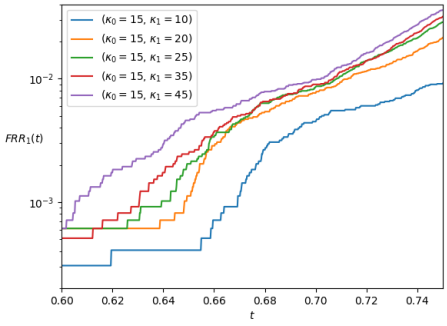


(c) Same as Figure 15(a) but only males are displayed. The female concentration parameter does not affect $\mathrm{FAR}_0(t)$.

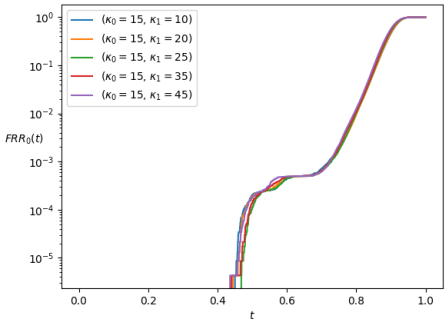
Figure 15: Influence of κ_1 on $\mathrm{FAR}_a(t)$.



(a) Influence of κ_1 on $\mathrm{FRR}_a(t)$. Females are depicted with solid lines while males are depicted with dashed lines.



(b) Zoom on Figure 16(a) for females.



(c) Same as Figure 16(a) but only males are displayed. The female concentration parameter barely affects $\mathrm{FRR}_0(t)$.

Figure 16: Influence of κ_1 on $\mathrm{FRR}_a(t)$.

Table 4: Evaluation on LFW for ArcFace with ResNet100 backbone and different pre-trained models (AdaCos, CosFace, CurricularFace) with MobileFaceNet backbone. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument). FRR@FAR is expressed as a percentage (%).

FAR level:		10^{-4}			10^{-3}		
model		FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
ArcFace	original	**0.063**	10.76	3.98	**0.052**	2.23	1.81
	(15,20)	0.119	12.73	**1.72**	0.067	8.43	**1.04**
	(25,20)	0.076	**5.35**	29.33	0.052	**1.94**	3.96
	(45,30)	0.129	13.47	2.99	0.067	6.02	1.24
AdaCos	original	**2.97**	3.64	3.84	0.98	5.29	2.23
	(15,20)	4.56	4.42	**1.41**	1.33	6.34	**1.01**
	(25,20)	3.12	**2.71**	8.37	**0.91**	**4.23**	3.71
	(45,30)	4.05	4.51	1.57	1.26	7.28	1.08
CosFace	original	**1.73**	5.89	2.51	**0.58**	8.18	1.74
	(15,20)	3.69	5.76	**1.13**	1.05	8.41	**1.02**
	(25,20)	2.41	**3.03**	9.66	0.67	**5.09**	4.75
	(45,30)	2.60	4.30	3.69	0.82	6.81	1.87
Curricular	original	**2.52**	3.67	2.92	**0.81**	4.88	1.91
	(15,20)	3.86	5.26	**1.16**	1.17	6.35	**1.10**
	(25,20)	2.82	**2.58**	9.10	0.82	**3.89**	4.28
	(45,30)	3.61	3.40	2.30	1.02	5.63	1.27

G. Additional numerical results

In this section, we provide more numerical experiments, varying the evaluation dataset (LFW, IJB-C, IJB-B) and different kinds of pre-trained models (ArcFace with several ResNet architectures, other pre-trained models with MobileFaceNet backbone).

G.1. Fairness evaluation on IJB-C and LFW

In Table 4, Table 5, Table 6, Table 7, we provide additional fairness evaluations on the IJB-C and LFW datasets, for ArcFace (with different ResNet architectures) and other pre-trained models with MobileFaceNet backbone (CosFace, CurricularFace, AdaCos (Zhang et al., 2019b)).

G.2. Verification evaluation on IJB-B

We finally investigate the FRR@FAR metric (Table 8, Table 9) of the three selected points (κ_0, κ_1) on IJB-B (Whitelam et al., 2017). In the verification setting, this dataset contains 10k genuine pairs and 8M impostor pairs. Notice that we do not lose too much in performance with respect to the original model.

Table 5: IJBC 1:1 protocol for ArcFace with ResNet100 backbone and different pre-trained models with MobileFaceNet backbone. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument).

FAR level:		10^{-4}			10^{-3}		
model		FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
ArcFace	original	**3.94**	1.97	4.06	**2.68**	1.79	2.04
	(15,20)	4.90	2.33	**1.17**	2.90	1.98	1.15
	(25,20)	4.34	**1.62**	11.86	2.76	**1.60**	5.58
	(45,30)	5.20	1.92	1.25	3.53	1.91	**1.07**
AdaCos	original	18.85	1.18	5.44	**9.74**	1.24	3.84
	(15,20)	20.75	1.30	**2.06**	10.31	1.42	2.20
	(25,20)	20.28	**1.02**	13.00	10.09	**1.06**	7.80
	(45,30)	**17.48**	1.28	2.86	9.85	1.33	**2.06**
CosFace	original	**15.67**	1.24	3.08	**8.55**	1.35	2.54
	(15,20)	19.52	1.35	**2.75**	10.24	1.41	**2.33**
	(25,20)	20.57	**1.03**	86.69	10.24	**1.04**	13.61
	(45,30)	17.27	1.12	8.67	9.69	1.11	4.29
Curricular	original	**17.69**	1.19	8.18	**9.26**	1.30	4.21
	(15,20)	19.97	1.33	**3.23**	10.37	1.42	**2.23**
	(25,20)	20.35	**1.04**	20.88	10.02	**1.03**	9.54
	(45,30)	18.07	1.18	5.29	9.99	1.22	3.33

Table 6: Evaluation on LFW for ArcFace on different ResNet architectures. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument).

FAR level:		10^{-4}			10^{-3}		
model		FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
R100	original	**0.063**	10.76	3.98	**0.052**	2.23	1.81
	(15,20)	0.119	12.73	**1.72**	0.067	8.43	**1.04**
	(25,20)	0.076	**5.35**	29.33	0.052	**1.94**	3.96
	(45,30)	0.129	13.47	2.99	0.067	6.02	1.24
R50	original	**0.078**	10.27	4.72	0.059	4.17	1.81
	(15,20)	0.151	11.22	**2.11**	0.072	9.16	**1.19**
	(25,20)	0.100	**5.89**	33.65	**0.058**	**4.11**	5.24
	(45,30)	0.164	9.18	2.44	0.081	5.15	1.20
R34	original	**0.104**	11.81	7.62	**0.063**	8.64	2.17
	(15,20)	0.204	14.27	**3.31**	0.087	17.56	1.59
	(25,20)	0.163	**5.63**	43.55	0.069	**8.09**	6.43
	(45,30)	0.226	8.85	4.42	0.095	8.80	**1.02**
R18	original	**0.214**	11.16	2.80	**0.116**	7.53	1.93
	(15,20)	0.465	11.15	**1.59**	0.197	10.60	**1.34**
	(25,20)	0.310	**4.44**	24.59	0.125	**6.53**	7.57
	(45,30)	0.349	6.69	4.21	0.162	6.92	1.76

Table 7: IJBC 1:1 protocol for ArcFace on different ResNet architectures. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument). Notice that PASS-g performs as well as our method on BFRR, but at the price of a poor performance / BFAR metrics compared to our method.

FAR level:		10^{-4}			10^{-3}		
model		FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
R100	original	**3.94**	1.97	4.06	**2.68**	1.79	2.04
	(15,20)	4.90	2.33	**1.17**	2.90	1.98	1.15
	(25,20)	4.34	**1.62**	11.86	2.76	**1.60**	5.58
	(45,30)	5.20	1.92	1.25	3.53	1.91	**1.07**
	PASS-g	9.00	1.70	4.49	6.27	1.79	2.97
R50	original	**4.29**	1.81	3.41	**3.00**	1.88	1.95
	(15,20)	5.56	2.18	1.28	3.40	2.18	**1.00**
	(25,20)	4.91	1.49	10.87	3.19	**1.50**	6.49
	(45,30)	5.41	1.73	**1.24**	3.71	1.77	1.09
	PASS-g	10.34	**1.45**	6.93	7.06	1.51	3.63
R34	original	**4.95**	1.72	2.83	**3.47**	1.77	1.88
	(15,20)	6.38	2.05	**1.17**	3.85	2.04	**1.06**
	(25,20)	5.67	1.45	13.69	3.60	**1.50**	5.86
	(45,30)	6.13	1.62	1.70	4.24	1.69	**1.06**
	PASS-g	12.03	**1.43**	4.00	8.36	**1.50**	2.79
R18	original	**6.64**	1.68	3.81	**4.41**	1.58	2.37
	(15,20)	8.64	1.83	**1.39**	4.96	1.88	**1.43**
	(25,20)	8.27	**1.19**	16.25	4.76	**1.26**	10.94
	(45,30)	7.46	1.50	3.16	4.97	1.56	1.85

Table 8: IJB-B 1:1 protocol for ArcFace on different ResNet architectures. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument).

		FRR@FAR (%)	
FAR level:		10^{-4}	10^{-3}
R100	original	5.38	3.78
	(15,20)	6.79	4.11
	(25,20)	6.00	3.84
	(45,30)	7.03	4.81
R50	original	5.95	4.20
	(15,20)	7.58	4.71
	(25,20)	6.71	4.26
	(45,30)	7.34	5.10
R34	original	6.72	4.63
	(15,20)	8.54	5.18
	(25,20)	7.62	4.60
	(45,30)	8.11	5.57
R18	original	8.59	5.76
	(15,20)	11.12	6.53
	(25,20)	10.94	6.01
	(45,30)	9.72	6.35

Table 9: IJB-B 1:1 protocol for ArcFace with ResNet100 backbone and different pre-trained models with MobileFaceNet backbone. By "original" we mean no Ethical Module is added to the pre-trained model. The tuples correspond to the choices of κ_0 (first argument) and κ_1 (second argument).

		FRR@FAR (%)	
FAR level:		10^{-4}	10^{-3}
ArcFace	original	5.38	3.78
	(15,20)	6.79	4.11
	(25,20)	6.00	3.84
	(45,30)	7.03	4.81
AdaCos	original	22.98	12.27
	(15,20)	24.06	12.78
	(25,20)	24.41	12.78
	(45,30)	21.25	12.44
CosFace	original	18.85	10.65
	(15,20)	23.38	12.51
	(25,20)	26.10	13.01
	(45,30)	21.22	12.27
Curricular	original	12.20	11.42
	(15,20)	24.50	12.56
	(25,20)	24.91	12.35
	(45,30)	21.88	11.97