Fairness-Aware Grouping for Continuous Sensitive Variables: Application for Debiasing Face Analysis with respect to Skin Tone

Veronika Shilova[A,B], Emmanuel Malherbe[A], Giovanni Palma[C], Laurent Risser[B] and Jean-Michel Loubes[B,D]
[A]Artefact Research Center, Paris, France
[B]Institut de Mathématiques de Toulouse (UMR 5219), CNRS, Université de Toulouse, F-31062 Toulouse, France
[C]L'Oréal Research and Innovation, Paris, France
[D]INRIA, ANITI, France

Within a legal framework, fairness in datasets and models is typically assessed by dividing observations into predefined groups and then computing fairness measures (e.g., Disparate Impact or Equality of Odds with respect to gender). However, when sensitive attributes such as skin color are continuous, dividing into default groups may overlook or obscure the discrimination experienced by certain minority subpopulations. To address this limitation, we propose a fairness-based grouping approach for continuous (possibly multidimensional) sensitive attributes. By grouping data according to observed levels of discrimination, our method identifies the partition that maximizes a novel criterion based on inter-group variance in discrimination, thereby isolating the most critical subgroups.
We validate the proposed approach using multiple synthetic datasets and demonstrate its robustness under changing population distributions—revealing how discrimination is manifested within the space of sensitive attributes. Furthermore, we examine a specialized setting of monotonic fairness for the case of skin color. Our empirical results on both CelebA and FFHQ, leveraging the skin tone as predicted by an industrial proprietary algorithm, show that the proposed segmentation uncovers more nuanced patterns of discrimination than previously reported, and that these findings remain stable across datasets for a given model. Finally, we leverage our grouping model for debiasing purpose, aiming at predicting fair scores with group-by-group post-processing. The results demonstrate that our approach improves fairness while having minimal impact on accuracy, thus confirming our partition method and opening the door for industrial deployment.

Keywords: *Algorithmic Fairness, Continuous Sensitive Attributes, Skin Tone, Debiasing*

1. Introduction

Group fairness measures inherently depend on the underlying distribution of machine learning data, such as specific input variables or the predicted scores for a given task. These measures are becoming increasingly impactful in light of recent AI regulations, such as the EU AI Act, which emphasize the need for reporting algorithmic performance across demographic groups. Beyond reporting, addressing disparities generally involves mitigation strategies applied on a group-by-group basis, based on qualitative sensitive variables. We refer, for instance, to Besse et al. (2022); Chouldechova and Roth (2020); Del Barrio et al. (2020); Dwork et al. (2012); Gordaliza et al. (2019); Kamishima et al. (2011); Risser et al. (2022); Wang et al. (2022) and references therein. The sensitive attribute can also be natively continuous, which is of primary interest in our paper. In this case, defining groups in the attribute space remains relevant, but raises the question on how to choose the partition and on which criteria.

A major field impacted by this problem is Computer Vision, for which widespread applications in high-stakes tasks have raised many ethical questions in recent years. A very sensitive scenario

Published at European Conference on Artificial Intelligence (ECAI) 2025
Corresponding author(s): veronika.shilova@artefact.com

is when models are applied to facial images, since the risk of bias with respect to (w.r.t.) skin tone might be related to ethnicity. This risk may occur for models that address transversal problems, such as face verification and attribute prediction, as first highlighted by Buolamwini and Gebru (2018) in which automated commercial facial analysis algorithms were shown to be biased against women with darker skin tones. It was later confirmed by Thong et al. (2023), as well as for text-to-image generation models, including DALL-E Cho et al. (2023). Such bias has been observed for many commercial applications, such as pedestrian detection Li et al. (2023), online ad delivery service Sweeney (2013), or online cosmetic product recommendation Malherbe et al. (2022). In this paper, we focus on the industrial scenario of a make-up recommendation algorithm based on selfie pictures. The recommended products and make-up strategy should indeed only depend on user's face morphology and not on its skin tone nor ethnicity.

Although being central for these fairness topics, the effective way to measure skin tone remains an open question. Many studies Buolamwini and Gebru (2018); Schumann et al. (2024) rely on manual annotators grouping individuals into predefined skin tone categories, such as the Fitzpatrick scale Fitzpatrick (1988) or Monk scale Schumann et al. (2024). However, this approach presents three major drawbacks. First, a simplistic division of tones risks oversimplifying the skin color and failing to capture nuanced biases embedded in datasets and models, which is obvious in the binary extreme of "darker" and "lighter" categories, but also occurs with the six groups of the commonly used Fitzpatrick scale. Second, these categories are predefined, while fairness is always related to a given task. The discriminated groups thus differ from one task to another, and hence cannot be captured by a default grouping of individuals. Third, manual annotations can introduce perception biases. Annotators might perceive skin tones differently depending on their cultural background Garcia and Abascal (2016); Gevers et al. (2012), which introduces annotator bias in the data. To overcome the limitations described above, an alternative line of research Robin et al. (2020); Thong et al. (2023) focuses on the estimated skin tone as a continuous quantity expressed in the color space, thus capturing a broader spectrum of skin tones. However, in the context of fairness, the necessity to consider groups raises the question on how to segment such skin color space.

In this paper, we introduce a paradigm that tailors the partitioning process to a given classification task. Our contributions are:

- We propose novel data-driven approaches to segment a continuous, possibly multidimensional sensitive variable into groups being homogeneous w.r.t. the discrimination present in a dataset or a model;
- We show through extensive experiments on skin tone that data-driven segmentation reveals more discrimination patterns than predefined groups and remains robust to population shifts, allowing us to better isolate how discrimination manifests within the sensitive attribute space. We also introduce and study the particular case of monotonic fairness, which may occur with skin tone;
- We use the groups identified by our data-driven method to mitigate the bias of a predictive model. Compared to other partitions, we obtain the best trade-off between model fairness and accuracy.

2. Related work

Measures of fairness for groups and for continuous attributes

Measuring and ensuring fairness in machine learning systems have become a central focus in recent years, leading to the development of various metrics assessing group-level disparities Caton and Haas (2024). Group fairness seeks to guarantee equitable treatment across predefined demographic groups,

and several standard measures have been proposed to evaluate and mitigate biases in algorithmic decision-making.

The most commonly used measure of algorithmic bias is Statistical Parity Corbett-Davies et al. (2017), frequently quantified in the fair learning literature through Disparate Impact (DI) Feldman et al. (2015). Statistical Parity requires that individuals from different demographic groups have equal chances of receiving a positive outcome. However, its simplicity comes at a cost, as it does not account for the true distributions of target outcomes between groups. For instance, if one group is historically underrepresented in the positive class, targeting equal outcomes across groups reduces overall accuracy. On the other hand, the Equalized Odds Berk et al. (2021) metric focuses on ensuring that the rates of true positives and false positives are comparable across all groups. By considering the underlying differences between groups, this measure highlights error rate equality, making it particularly relevant in contexts where fairness in outcomes is critical, such as criminal justice and healthcare. Despite its advantages, achieving Equalized Odds in practice can be challenging without sacrificing predictive accuracy.

In the case of a continuous sensitive attribute, fairness is often quantified through a measure of statistical correlation. The most common measures are Hirschfeld-Gebelein-Rényi (HGR) maximum correlation coefficient Mary et al. (2019) and Generalized Disparate Impact Giuliani et al. (2023), which is a modified formulation of HGR. These fairness metrics primarily evaluate the overall dependence between a continuous sensitive variable and a target variable, without specifying which groups within the population are more or less affected by discrimination. As a result, these metrics are not directly applicable to detect and locate biases in models. This limitation is particularly relevant in the context of emerging legal frameworks, such as the EU AI Act, which mandates that AI systems ensure they do not disadvantage specific groups based on attributes such as skin tone, age, or income, which, in fact, are inherently continuous.

Skin color as a sensitive variable in Computer Vision

While attributes representing skin tone is often provided in Computer Vision datasets, it is represented as a discrete attribute in the most commonly used ones. For example, the skin color attribute has been annotated in the two groups *Pale/Not Pale* in the CelebA dataset Liu et al. (2015), in the 6 groups of the Fitzpatrick scale Fitzpatrick (1988) in the IJB-C Maze et al. (2018) and Facet Gustafson et al. (2023) datasets, and in the 10 groups of the Monk scale Schumann et al. (2024) in the MST-E dataset Garcia and Abascal (2016). These commonly used groups for skin tone are of a limited number, when compared to the full range of skin colors. Besides, they are generally given by a human annotator (as in the datasets mentioned above), which has been observed to be biased with respect to its geographical region or its culture Gevers et al. (2012), or to the inferred race/ethnicity of the individual in the picture Garcia and Abascal (2016). To mitigate annotator bias and limited skin groups granularity, the recent line of work intended to consider skin tone as a measured color, thus a continuous quantity. The study Thong et al. (2023) proposed to extract skin color directly from pixels instead of relying on human perception and converted it in the CIELAB color space Weatherall and Coombs (1992). This color space, unlike RGB, correlates with the response of the human eye by covering its entire range of color perception and is a golden standard to measure skin color in clinical dermatology Wilkes et al. (2015) and aesthetic cosmetology Ly et al. (2020). However, estimation of skin color directly from pixels' color may be noisy as it depends on lightning conditions of a given picture, especially for in-the-wild pictures Thong et al. (2023). As a more reliable estimate, using a spectrophotometer to measure panelists' skin reflectance provides a ground truth value for training a supervised model Ly et al. (2020); Robin et al. (2020).

The previous approaches to quantifying fairness w.r.t. skin tone in datasets and models Buolamwini and Gebru (2018); Cho et al. (2023); Li et al. (2023); Thong et al. (2023) predefine the sensitive

attribute into binary or discrete groups. For instance, Buolamwini and Gebru (2018); Cho et al. (2023); Li et al. (2023), divide the population into two broad categories: *Lighter* and *Darker* skin tones. In contrast, Thong et al. (2023) proposes a more granular characterization with four subgroups that combine two dimensions: *Lighter* and *Darker* skin tones representing skin lightness, and *Red* and *Yellow* skin hues reflecting variations in perceived skin color gradation from reddish to yellowish tones. Further improvements can however be made, since these discretizations overlook both the underlying data distribution and the target variable. Another limit of these predefined groups is that samples inside a group are assumed to be homogeneous in terms of fairness, which is not guaranteed in reality. Besides, as already shown in Kearns et al. (2018), overly simplistic group definitions may overlook fairness nuances and underestimate the level of discrimination present in the data or/and introduced by the model.

3. Color space for skin tone

In view of the litterature for skin tone estimation, in this work we consider the skin tone as a color in the three-dimensional CIELAB color space. We present below this color space and the color attributes commonly derived from it.

- **CIELAB color space for measures.** CIE $L^*a^*b^*$ (CIELAB) color space Sève (2009) was defined by the International Commission on Illumination in 1976 with the aim of approximating human perception of color differences. A color is expressed by a vector $(L^*, a^*, b^*) \in \mathbb{R}^3$ with the property that the Euclidean distance between two colors is proportional to their difference as perceived by humans, which is not verified in the RGB space. The $L^* \in [0, 100]$ component corresponds to the perceptual lightness and ranges from black to white. The $a^* \in [-128, 127]$ component refers to the green-red opponent, with negative values corresponding to green and positive values to red. The $b^* \in [-128, 127]$ component describes the blue-yellow opponent colors, with negative values corresponding to blue and positive values to yellow.
- **Individual typology angle.** The individual typology angle (*ITA*) is a one dimensional measure of skin tone defined by Chardon et al. (1991) in the CIE $L^*a^*b^*$ color space as follows:

$$ITA = arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180^\circ}{\pi}. \tag{1}$$

  It is widely used for skin color measurements in dermatology Wilkes et al. (2015) and aesthetic cosmetology Ly et al. (2020). In fairness field, *ITA* is commonly applied to facial images Karkkainen and Joo (2021); Merler et al. (2019) and skin analysis Kinyanjui et al. (2020). The Fitzpatrick scale is derived from *ITA* by predefined segmentation into 6 groups.
- **Hue angle.** Hue angle h^* measures the perceived gradation of colors, and is defined as the angular coordinate in the $L^*a^*b^*$ space:

$$h^* = arctan\left(\frac{b^*}{a^*}\right) \times \frac{180^\circ}{\pi}, \tag{2}$$

  where h^* ranges from 0° to 360°. However, we are primarily interested in the values between 0° to 90°, since they correspond to the red and yellow colors in which skin undertone is expressed Chardon et al. (1991). The hue angle was originally proposed for skin color measurements by Weatherall and Coombs (1992), and is now widely adapted for skin color estimation in cosmetology De Rigal et al. (2007).

In our experiments, to ensure that these quantities are not biased, we extract the skin tone values from face pictures leveraging a proprietary model implementing the methodology detailed in

Malherbe et al. (2022); Robin et al. (2020). Since the underlying algorithm aims to approximate physical measurements from a device, it is not susceptible to human-induced annotation bias. The detailed methodology is beyond the scope of this paper.

4. From continuous to discrete: enhancing fairness in sensitive variable segmentation

In this paper, we place ourselves in the case of a dataset with a binary classification task as target outcome and a continuous sensitive attribute. Formally, we consider a dataset consisting of i.i.d observations $(X, L, Y) \sim \mathcal{D}$, tuples of random variables $X \in \mathbb{R}^d$ ($d \in \mathbb{N}$), $L \in \mathcal{L} \subset \mathbb{R}^p$ ($p = 1$ or 2) and $Y \in \{0, 1\}$ drawn from a distribution $\mathcal{D}$. In our application, X contains images of human faces in the RGB color space, L is a continuous (multidimensional) sensitive attribute corresponding to skin color, such as L^*, *ITA* or (L^*, h^*) from CIE $L^*a^*b^*$ color space defined in previous section, and Y is an observed outcome variable. The parameters θ of a binary classifier $\hat{Y} = g_\theta(X) \in \{0, 1\}$ are trained using the observations from the dataset $\mathcal{D}$. In this setting, our objectives are two-fold:

1. Showcase the relevance of using continuous (possibly, multidimensional) skin color score to measure bias in Computer Vision datasets and models;
2. Propose a fairness-based segmentation of continuous (multidimensional) skin color. In other words, instead of measuring the fairness of a dataset or a model using predefined groups (e.g. Pale vs. Not Pale, Fitzpatrick groups), we propose to divide the observation according to groups in the skin color space $\mathcal{L}$. Such groups would need (i) to be connected and (ii) to have homogeneous levels of discrimination for the target problem Y or model $\hat{Y}$.

For the second objective, we formally write a partition $\mathcal{P} = \{\mathcal{P}_k\}_{k=1}^K$ as a set of connected subsets $\mathcal{P}_k \subset \mathcal{L}$ with the two following properties: there is no overlap between elements of the partition: $\forall k_1, k_2 \in \{1, \dots, K\}$, $\mathcal{P}_{k_1} \cap \mathcal{P}_{k_2} = \emptyset$; and $\mathcal{P}$ totally covers $\mathcal{L}$: $\cup_{k=1}^K \mathcal{P}_k = \mathcal{L}$. We consider that the number of groups K is given, as a user-defined variable.

In the remainder of the paper, we omit * in the notation of L^* and h^* for readability purposes. Although we will keep the example of skin tone, our methods apply to any continuous sensitive attribute.

4.1. Inter-group variability as criteria for a fairness based segmentation

We now focus on dividing $\mathcal{L}$ into K groups $\mathcal{P}_k$ representing different levels of fairness, given a classification problem. Contrary to predefined groups, these groups should be based on the level of discrimination introduced by either the ground truth Y or the predicted variable $\hat{Y}$.

Formally, for a partition $\mathcal{P}$ we define $S^{\mathcal{P}}$ to be a multinomial random variable associating the observations of L to their group $\mathcal{P}_k$:

$$S^{\mathcal{P}} = k \iff L \in \mathcal{P}_k, \text{ where } k = 1, \dots, K, \tag{3}$$

which is commonly called the group membership. Note that while it is generally a provided variable in the dataset, it depends here on the partition $\mathcal{P}$ and is thus not fixed given an observation. Then, we define $\Phi(S^{\mathcal{P}})$ as a real function of this multinomial random variable $S^{\mathcal{P}}$. Φ is an extension of a fairness measure m computed on each group $\mathcal{P}_k$ and takes K values $\Phi(k)$, $k = 1, \dots, K$:

$$\Phi(k) = m(\mathcal{D}, \mathcal{P}_k) \in \mathbb{R}. \tag{4}$$

For instance, a typical measure m is the disparate impact, computed for samples in group $\mathcal{P}_k$ versus all other samples:

$$m(\mathcal{D}, \mathcal{P}_k) = \mathbb{P}(Y = 1 | L \in \mathcal{P}_k) - \mathbb{P}(Y = 1 | L \notin \mathcal{P}_k). \tag{5}$$

We propose to find the partition $\mathcal{P}$ of the continuous sensitive attribute L, such that its components $\mathcal{P}_k$ have the most different levels of discrimination according to the fairness measured by Φ. In other terms, such partition is the most expressive one in terms of the measure of fairness Φ. Our objective is thus to find the partition $\mathcal{P}$ over all possible partitions that maximizes:

$$\arg\max_{\mathcal{P}} \mathbb{V}ar\Big(\Phi\big(S^{\mathcal{P}}\big)\Big). \tag{6}$$

where the variance is calculated over the observations of L in the dataset. The variance of $\Phi(S^{\mathcal{P}})$ can thus be rewritten as:

$$\begin{aligned} \mathbb{V}ar\Big(\Phi\big(S^{\mathcal{P}}\big)\Big) &= \mathbb{V}ar_{L\sim\mathcal{D}}\Big(\Phi\big(S^{\mathcal{P}}\big)\Big) = \\ &= \int_{l\in\mathcal{L}} \Big(\Phi\big(S^{\mathcal{P}}\big) - \mathbb{E}\big[\Phi\big(S^{\mathcal{P}}\big)\big]\Big)^2 p_L(l)dl = \end{aligned} \tag{7}$$

$$\begin{aligned} &= \sum_{k=1}^{K} \int_{l\in\mathcal{P}_k} \Big(\Phi\big(k\big) - \mathbb{E}\big[\Phi\big(S^{\mathcal{P}}\big)\big]\Big)^2 p_L(l)dl = \\ &= \sum_{k=1}^{K} \mathbb{P}\big(S^{\mathcal{P}} = k\big)\Big(\Phi\big(k\big) - \mathbb{E}\Big[\Phi\big(S^{\mathcal{P}}\big)\Big]\Big)^2, \end{aligned} \tag{8}$$

where $p_L(l)$ is the probability density function of L and

$$\mathbb{E}\Big[\Phi\big(S^{\mathcal{P}}\big)\Big] = \int_{l\in\mathcal{L}} \Phi\big(S^{\mathcal{P}}\big) p_L(l)dl = \sum_{k=1}^{K} \mathbb{P}\big(S^{\mathcal{P}} = k\big)\Phi\big(k\big). \tag{9}$$

The interpretation behind this maximization is two-fold. First, from the expression over $\mathcal{D}$ of Equation (7), we argue that the exact partition maximizes the variability of fairness outcomes across observations, so that it obtains on average the most "unfair" measures for each person of $\mathcal{D}$. Indeed, the variance in question captures the mean deviation between each observation's group fairness measure $\Phi\big(S^{\mathcal{P}}\big)$ and the population-wide expected fairness value. Second, when expressed as a sum over the groups (Equation (8)), this same variance can be viewed as the inter-group variance—weighted by each group's probability $\mathbb{P}\big(S^{\mathcal{P}} = k\big)$. Hence, maximizing it enhances the "expressivity" of our partition by increasing the variability of fairness measures among groups.

4.2. One-group-versus-all as fairness measure

We propose to choose the fairness measure $m(\mathcal{D}, \mathcal{P}_k)$ (Equation (4)) based on the following. A model g is considered fair w.r.t. the sensitive attribute $S^{\mathcal{P}}$ if the model's output does not depend on the value of the sensitive attribute. This is usually modeled as the expected output $\hat{Y}$ being independent of the sensitive attribute $S^{\mathcal{P}}$:

$$\mathbb{E}[\hat{Y}] = \mathbb{E}[\hat{Y} | S^{\mathcal{P}} = k], \forall k = 1, \ldots, K. \tag{10}$$

Hence, we choose as fairness measure:

$$\Phi(k) = m(\mathcal{D}, \mathcal{P}_k) = \mathbb{P}(Y = 1 | S^{\mathcal{P}} = k) - \mathbb{P}(Y = 1). \tag{11}$$

This measure is a natural extension of the well-known disparate impact for two classes (see Section 4.3) and is related to the definition of subgroup fairness defined in Kearns et al. (2018). Under this choice, Equation (6) can be rewritten in the following way

$$\begin{aligned} \mathbb{V}ar\Big(\Phi\big(S^{\mathcal{P}}\big)\Big) &= \mathbb{V}ar\Big(\mathbb{P}(Y=1|S^{\mathcal{P}}) - \mathbb{P}(Y=1)\Big) = \\ &= \sum_{k=1}^{K} \mathbb{P}\big(L \in \mathcal{P}_k\big)\Big(\mathbb{P}(Y=1|S^{\mathcal{P}}=k) - \mathbb{P}(Y=1)\Big)^2. \end{aligned} \tag{12}$$

Note that $\mathbb{E}\Big[\Phi(S^{\mathcal{P}})\Big] = \mathbb{E}\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}) - \mathbb{P}(Y=1)\Big] = 0$.

Therefore, maximizing the variance defined in Equation (12) when Y is binary is equivalent to maximizing the sum of squared distances between $\mathbb{E}[Y]$ and $\mathbb{E}[Y|S^{\mathcal{P}}=k]$ for each group $\mathcal{P}_k$. Each of the terms is multiplied by the "weight" of the group $\mathcal{P}_k$, which takes into account the impact of this group on the population. We see that with this fairness measure, the optimal partition $\mathcal{P}$ has the groups with the worst-case level of discrimination and with the major impact, or equivalently the most discriminating partition. In the following section, we provide a formal expression illustrating this in the case $K = 2$.

4.3. Relation with Disparate Impact in the binary case

For $K = 2$, we establish the relation between the $\Phi(k)$ chosen in the previous section, and the well-known Disparate Impact. The proof is in Appendix A. In the proposition below, we shift $S^{\mathcal{P}}$ values (Equation (3)) to 0 and 1 to align with usual binary notations.

Proposition 4.1. *Assume that $S^{\mathcal{P}}$ is a binary sensitive random variable and $\Phi(S^{\mathcal{P}}) = \mathbb{P}(Y=1|S^{\mathcal{P}}) - \mathbb{P}(Y=1)$. Then, the partition $\mathcal{P}$ maximizing $\mathbb{V}ar\Big(\Phi\big(S^{\mathcal{P}}\big)\Big)$ consists of the least and the most privileged groups with the maximal impact, i.e.*

$$\mathbb{V}ar\Big(\Phi\big(S^{\mathcal{P}}\big)\Big) = \pi(1-\pi)DI^2, \tag{13}$$

where $\pi = \mathbb{P}(S^{\mathcal{P}}=1)$ and $DI = \mathbb{P}(Y=1|S^{\mathcal{P}}=1) - \mathbb{P}(Y=1|S^{\mathcal{P}}=0)$.

The maximized quantity $\pi(1-\pi)DI^2$ is made of two terms. First, the usual DI which promotes a separation into groups exhibiting the most different behaviors, and thus promoting unfairness. Second, the impact of the weights $\pi(1-\pi)$ enforces the probability of occurrence of the two groups to be as similar as possible, with a maximum achieved for this part at $\pi = 1/2$. This provides another insight of the criterion we promote. We do not want to select a too small (or too large) group w.r.t. the other, exhibiting a particular behavior due to its size difference. Our objective is rather to identify two kind of behaviors which are intrinsically different without relying on the sampling properties of the individuals. This is coherent with the choice of so-called balance error groups when providing a mathematically sound definition for bias in the data as shown in Gordaliza et al. (2019).

5. Fairness aware segmentation algorithms

In the following section, we provide methods to find the partition from Equation (6) for one- and two-dimensional sensitive variable $L \in \mathcal{L} \subset \mathbb{R}^p$, $p = 1, 2$. We consider partitions $\mathcal{P}$ as follows:

- In the one-dimensional case, we consider $\mathcal{P}_k = [l_{k-1}, l_k]$ to be segments, with $l_{k-1} < l_k$.
- In the two-dimensional case, we consider $\mathcal{P}_k$ to be rectangles, i.e. of the shape $[l, l'] \times [h, h']$.

5.1. Leveraging K-Means in the case of monotonic fairness in 1D

Finding the exact solution to Equation (6) is not feasible. Hence an alternative is given by clustering algorithms Ezugwu et al. (2022); Xu and Wunsch (2005), which naturally align with a segmentation objective. Since we aim at maximizing the distance between different groups in terms of the fairness measure Φ, we naturally consider applying the usual K-Means algorithm MacQueen (1967) on a set of points $\psi_j \in \mathbb{R}$ whose values will aim at *reflecting* this fairness. This procedure is detailed hereafter.

Let us consider $M \in \mathbb{N}$ and a grid $\Lambda = \{\lambda_0, \dots, \lambda_M\}$ on $\mathcal{L}$, where $\lambda_j = \lambda_0 + j\delta$, $\delta = \frac{\lambda_M - \lambda_0}{M}$, $j = 0, \dots, M$. We compute on the M intervals the $\psi_j = \mathbb{P}(Y = 1 | L \in [\lambda_{j-1}, \lambda_j]) - \mathbb{P}(Y = 1)$, $j = 1, ..., M$, as shown in Figure 1. In this case, applying K-means on the ψ_j will compute a partition $\mathcal{P}$ that minimizes

$$\arg\min_{\mathcal{P}} \sum_{k=1}^{K} \sum_{j=1}^{M} (\psi_j - \mu_k)^2 \mathbb{1}_{\{\psi_j \in \mathcal{P}_k\}}, \tag{14}$$

where $\mu_k = \frac{\sum_{j=1}^{M} \psi_j \mathbb{1}_{\{\psi_j \in \mathcal{P}_k\}}}{\sum_{j=1}^{M} \mathbb{1}_{\{\psi_j \in \mathcal{P}_k\}}}$ is the centroid of the k-th group.

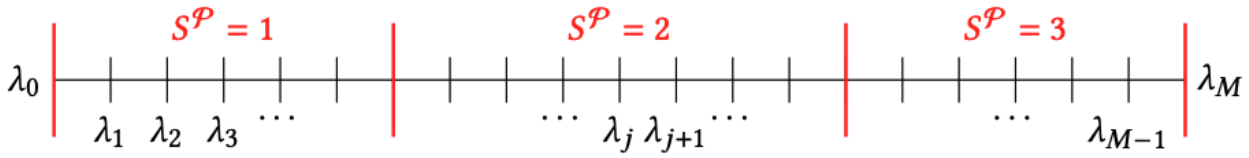


Figure 1 | Partition of $L \in [\lambda_0, \lambda_M]$ into intervals corresponding to $S^{\mathcal{P}} = 1, 2, 3$.

By applying K-Means algorithm, the k-th cluster will correspond to $\mathcal{P}_k$. The ψ_j values will be homogeneous within each cluster, while being as distinct as possible between clusters. The K-Means minimization is indeed equivalent to maximizing the sum of $(\mu_k - \mathbb{E}(\mu_k))^2$ over clusters, which relates to the maximization problem given by Equation (6). One notes that this is an heuristic approach and the obtained $\mathcal{P}$ is not the exact solution of Equation (6). Indeed, while groups $\mathcal{P}_k$ will be homogeneous in terms of fairness express by ψ_j, the exact Φ (Equation 4) is not computed on the groups $\mathcal{P}_k$, and in particular $\Phi(k) \neq \mu_k$.

Furthermore, in general, K-Means do not guarantee that groups will be connected in the $\mathcal{L}$ space. The underlying distance used in L dimension between λ_{j_1} and λ_{j_2} is indeed $|\psi_{j_1} - \psi_{j_2}|$, which can take low or null values for far $\lambda_{j_1}, \lambda_{j_2} \in \mathcal{L}$. We now define the property of monotonic fairness, that will guarantee that each groups from this method is connected.

Definition 5.1. We say that the fairness is monotonic w.r.t. to the continuous sensitive variable $L \in \mathcal{L} \subset \mathbb{R}$, if the distribution function $\mathbb{P}(Y = 1 | L = l)$ is monotonic w.r.t. $l \in \mathcal{L}$. Note that monotonicity w.r.t. L is equivalent to monotonicity w.r.t. $S^{\mathcal{P}}$ $\forall \mathcal{P}$.

Proposition 5.2. *Assume that $\mathbb{P}(Y = 1|L)$ is monotonic w.r.t. $L \in \mathcal{L} \subset \mathbb{R}$. Then $\mathbb{P}(Y = 1|L) - \mathbb{P}(Y = 1)$ is also monotonic w.r.t. L, and the partition $\mathcal{P}$ found by the K-Means algorithm always consists of segments of $\mathcal{L}$.*

According to Proposition 5.2, the assumption of monotonicity of fairness is crucial, since if it is not fulfilled, K-Means can produce groups that are disconnected (see Figure 6 in Appendix B for details). This notion of monotonic fairness can well occur in practice, for example in skin tone, as we'll illustrate in the following sections.

Note that Y may be replaced by the model output $\hat{Y}$ in Definition 5.1 and Proposition 5.2, to analyze the model forecasts' discrimination.

5.2. FairGroups: exhaustive search with optimal precalculation of fairness metric Φ

We now look for the solution of Equation (6) in the general case, without any assumption on function $P(Y=1|L=l)$. We again consider a grid Λ on L with M intervals. To find the partition $\mathcal{P}$ for a given K, we perform an exhaustive search over all possible combinations of intervals on the grid $[\lambda_{j_1}, \lambda_{j_2}]$, $j_1 < j_2$. To speed up the exhaustive search, we propose Algorithms 1 and 2 (Appendix C) to efficiently precalculates $\Psi_{j_1,j_2} := \mathbb{P}(Y=1|L \in [\lambda_{j_1}, \lambda_{j_2}]) - \mathbb{P}(Y=1)$, $\forall j_1 < j_2$ with a complexity $O(M^2)$ using a dynamic programming strategy.

For this, we define an upper triangular matrix $U_{\Psi_\Lambda} \in \mathbb{R}^{M \times M}$ where we store all the precalculated values as follows

$$(U_{\Psi_\Lambda})_{i,j} = \begin{cases} \Psi_{i-1,j}, & \text{for } i \leq j, \text{ where } i,j = 1,\dots,M-1, \\ 0, & \text{for } i > j. \end{cases}$$

Each value Ψ_{j_1,j_2} of matrix U_{Ψ_Λ} is estimated using the ratio between $\sum_{j_1<j_2} \sum_{\mathcal{D}} \mathbb{1}_{\{L \in [\lambda_{j_1}, \lambda_{j_2}]\}} \mathbb{1}_{\{Y=1\}}$ and $\sum_{j_1<j_2} \sum_{\mathcal{D}} \mathbb{1}_{\{L \in [\lambda_{j_1}, \lambda_{j_2}]\}}$, that both can be computed by recurrence on λ_{j_1} and λ_{j_2} using Algorithm 1. Matrix U_{Ψ_Λ} is then efficiently precomputed with Algorithm 2 and is subsequently used for exhaustive search. This eliminates the need to recalculate the values of Ψ_{j_1,j_2}, when evaluating potential combination of intervals, thereby significantly speeding up the computational process. One notes that given the grid Λ, the ψ_j defined in the previous section are the diagonal of U_{Ψ_Λ}, since $\psi_j = \Psi_{j-1,j}$

Given this matrix U_{Ψ_Λ}, we find the partition $\mathcal{P}$ of segments $\mathcal{P}_k = [\lambda_{j_{k-1}}, \lambda_{j_k}]$ $(j_{k-1} < j_k)$ of Equation (6) as:

$$\underset{\substack{0<j_1<j_2<\dots \\ \dots<j_{K-1}<M}}{\arg\max} \quad \sum_{k=1}^{K} \mathbb{P}\left(L \in [\lambda_{j_{k-1}}, \lambda_{j_k}]\right) \left(\Psi_{j_{k-1},j_k} - \mathbb{E}\left[\Phi\left(S^{\mathcal{P}}\right)\right]\right)^2, \tag{15}$$

where $\mathbb{E}[\Phi(S^{\mathcal{P}})] = \sum_{k=1}^{K} \mathbb{P}(L \in [\lambda_{j_{k-1}}, \lambda_{j_k}]) \Psi_{j_{k-1},j_k}$. The $\mathbb{P}(L \in [\lambda_{j_{k-1}}, \lambda_{j_k}])$ values are precalculated using Algorithm 1. The solution is then exact, up to the approximation of the partitions search space induced by the grid Λ.

6. Experiments

The code for using our methods and reproducing these experiments is in open access [1].

Table 1 | Comparison of partitions of L for our synthetic datasets based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(Y=1|S^{\mathcal{P}}) - \mathbb{P}(Y=1)$, $K=5$.

Data distribution for computing $\mathcal{P}$	$L \sim \mathcal{U}(0,100)$		$L \sim \mathcal{N}(50,20)$ truncated on $[0,100]$	
Metric	$\mathbb{V}ar\left(\Phi\left(S^{\mathcal{P}}\right)\right)$	Rand Index (with ground truth)	$\mathbb{V}ar\left(\Phi\left(S^{\mathcal{P}}\right)\right)$	Rand Index (with ground truth)
$\mathcal{P}$: Ground Truth	**0.068**	1	**0.032**	1
$\mathcal{P}$: K-Means	0.067	0.972	0.025	0.804
$\mathcal{P}$: FairGroups	**0.068**	0.99	**0.032**	0.97

6.1. Synthetic datasets

We first evaluate our proposed methods on a series of synthetic data. To do so, we suppose that the discrimination is expressed in the dataset in terms of Statistical Parity, i.e. we have distinct intervals

[1] `https://github.com/artefactory/fairness_metrics`

Table 2 | Comparison of partitions of *ITA* based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(\hat{Y} = 1|S^{\mathcal{P}}) - \mathbb{P}(\hat{Y} = 1)$, $K = 6$.

Data for computing $\mathcal{P}$	CelebA		FFHQ	CelebA & FFHQ
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on CelebA	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	Rand Index between both partitions, on FFHQ
$\mathcal{P}$: Fitzpatrick	0.078	0.079	0.079	1
$\mathcal{P}$: K-Means	0.089	0.087	0.088	0.869
$\mathcal{P}$: FairGroups	**0.092**	**0.09**	**0.093**	0.921

with different probability of success $\mathbb{P}(Y = 1|L)$. We also choose to work in the scenario of monotonic fairness (Proposition 5.2).

We suppose that $L \sim \mathcal{U}(0, 100)$ and corresponding $Y \sim Bernoulli(p(L))$, where $p(L) = 0.1 \times \mathbb{1}_{\{L\leq 20\}} + 0.3 \times \mathbb{1}_{\{20<L\leq 30\}} + 0.5 \times \mathbb{1}_{\{30<L\leq 55\}} + 0.7 \times \mathbb{1}_{\{55<L\leq 88\}} + 0.9 \times \mathbb{1}_{\{88<L\leq 100\}}$ (see Figure 2). First, we sample $N = 50000$ pairs of observations (L, Y) from this distribution $\mathcal{D}$. Second, to illustrate a non-uniform scenario, we additionally sample a dataset in the case L follows a truncated normal distribution on $[0, 100]$ from $\sim \mathcal{N}(50, 20)$. The two empirical distributions are shown in Figure 7 (Appendix D).

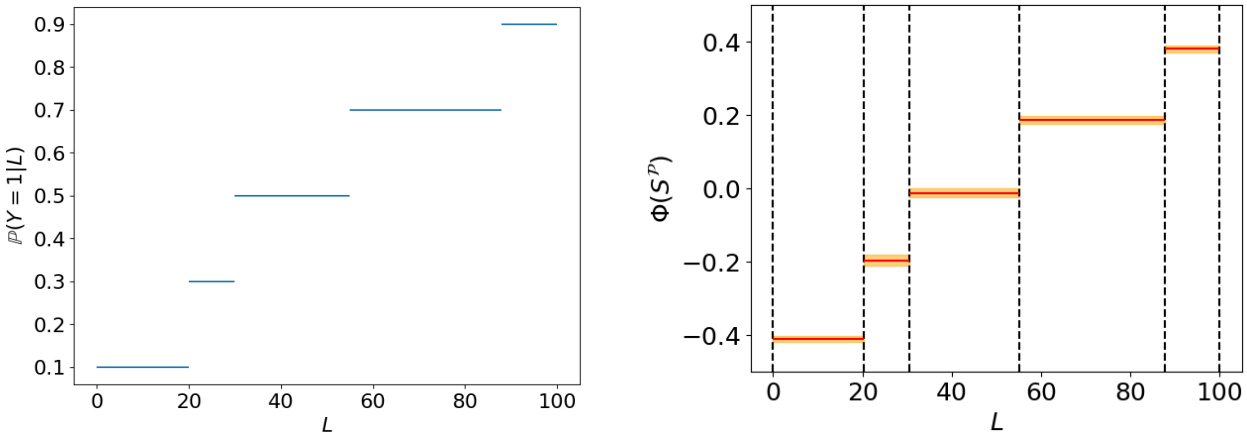


Figure 2 | Partitions on synthetic dataset with uniform distribution of L: $\mathbb{P}(Y = 1|L)$ on the left and FairGroups partition on the right.

On both datasets, we apply the K-Means and the FairGroups methods to find the partition of L, as shown in Figure 2. We compute $\Phi(S^{\mathcal{P}})$ values for each group $\mathcal{P}_k$ in y-axis, as well as their asymptotic confidence intervals using the delta method following the work Besse et al. (2022) (as red transparent areas in Figures 2 and 9 in Appendix D). To measure how close each found partition is to the ground truth partition, we measure the pairwise Rand index Rand (1971) with the ground truth partition, a common measure of the similarity between two data clusterings. One notes that the ground truth partition is independent of the L distribution and is thus the same for both synthetic datasets. Table 1 shows results alongside the variance $\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ values.

From Figure 2 and Table 1, we conclude that FairGroups approach managed to find the ground truth partition almost perfectly, i.e. the Rand Index is 0.99. Furthermore, we can see that FairGroups manages to find the ground truth partition of L regardless of its distribution, with very high Rand

Index for the non-uniform distribution. On the other hand, K-Means groups remains quite close to the ground truth in the case of uniform data, but is less accurate than FairGroups. For non-uniform data, we see that K-Means Rand Index is even lower.

6.2. Data and protocol for face analysis w.r.t skin tone

In this section, we use the women face pictures of CelebA dataset Liu et al. (2015) as $\mathcal{D}$, with the number of observations $N = 84602$. We predict skin tone L, a, b values for each image in $\mathcal{D}$ using the proprietary algorithm Robin et al. (2020), and derive from it L, *ITA* or (L, h) as the sensitive variable. As the target output variable Y, we use the attribute "Attractive". We fine-tune on 5 epochs a ResNet18 He et al. (2016) as model g_θ, without any bias mitigation strategy, with batch size equal to 64. The prediction $\hat{Y}$ for each CelebA image is obtained after a 5-fold cross-validation, stratified on 10 equal-sized bins of L. We also consider FFHQ[2] ($N = 26906$) for assessing the generalization of our partitioning algorithm, with skin tone extracted as for CelebA pictures. FFHQ does not have a ground truth Y for our task, so that we predict $\hat{Y}$ as the mean of our models g_θ, meaning we are looking at the discrimination of the same model. Please note that the sensitive attribute values being estimated by a proprietary algorithm, they cannot be publicly shared.

The "Attractive" target on CelebA dataset represents an interesting case of monotonic fairness (Definition 5.1), as shown in Figure 4. We indeed observe in the initial data as well as in the model's predictions that, on average, the darker a person's skin is, the more discriminated they are. Please note that this target is known to represent a biased annotation Risser et al. (2022), and was chosen purely for illustrative purpose.

6.3. Partitions of skin tone in 1D

We estimated partitions of skin tone given by *ITA* and compare them to the default partition of *ITA* with Fitzpatrick scale in 6 groups. Figure 3 and Table 2 show the results comparing the different groupings. Both FairGroups and K-Means partitions identify groups with varying levels of discrimination. The Fitzpatrick scale, as a default partition, does not consider the underlying data distribution, so that the confidence intervals for certain groups are relatively broad, so that it is difficult to draw conclusions regarding fairness between groups.

We also consider the partitions of FFHQ and compare them with the partitions of CelebA applied to FFHQ, as done in the previous subsection. From Table 2, we observe that both FairGroups and K-Means partitions remain stable when the data distribution changes. This suggests that the partitions effectively reveal how discrimination manifests within the sensitive attribute space.

Table 3 | Comparison of partitions of (L, h) in 2D based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(\hat{Y} = 1|S^{\mathcal{P}}) - \mathbb{P}(\hat{Y} = 1)$, $K = 4$.

Data for computing $\mathcal{P}$	CelebA		FFHQ	CelebA & FFHQ
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on CelebA	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	Rand Index between both partitions, on FFHQ
$\mathcal{P}$: Default	0.054	0.082	0.082	1
$\mathcal{P}$: FairGroups	**0.084**	**0.102**	**0.105**	0.883

[2] `https://github.com/NVlabs/ffhq-dataset`

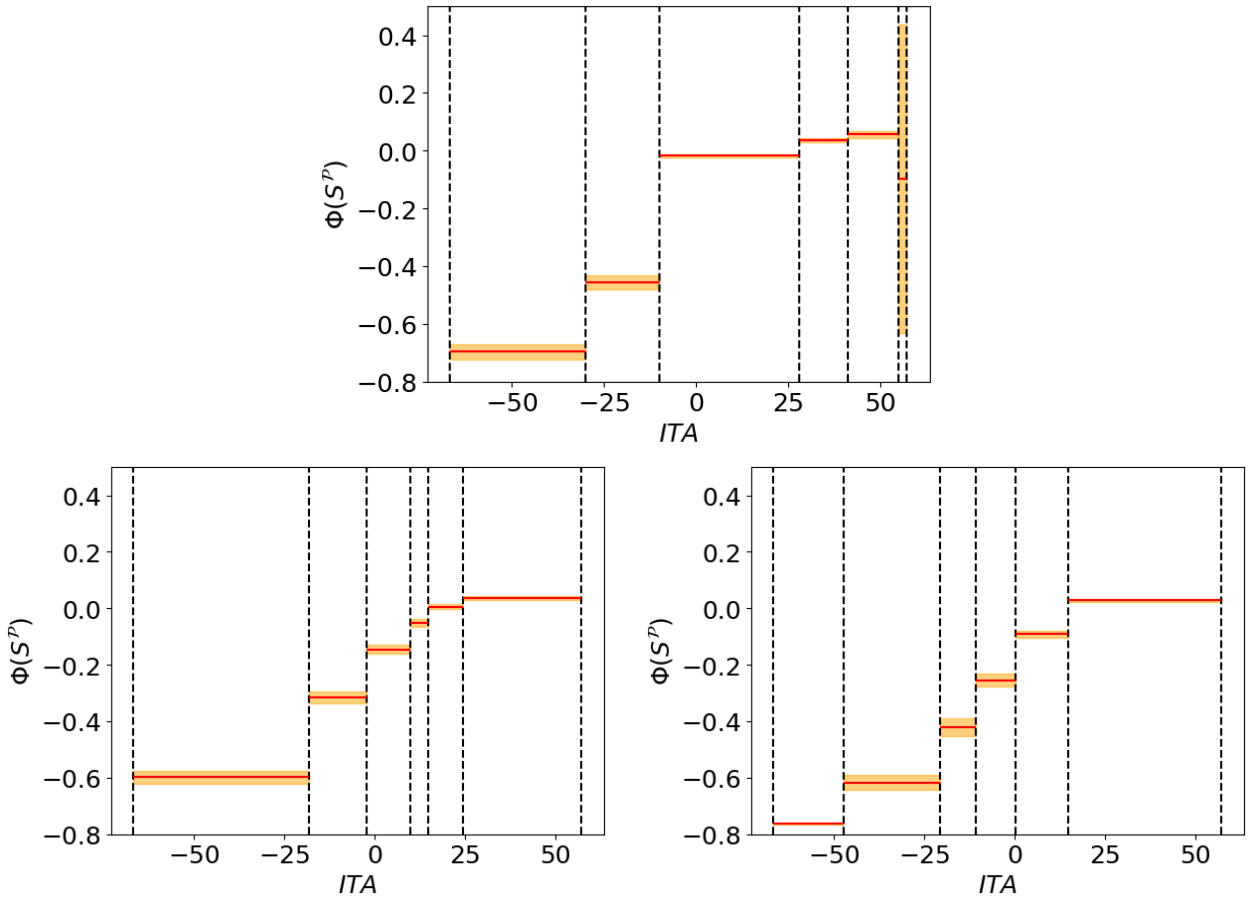


Figure 3 | Partitions of *ITA* using different methods. Fitzpatrick (default) partition of *ITA* (top). FairGroups partition of *ITA* (bottom left). K-Means partition of *ITA* (bottom right).

6.4. Amplification of skin color bias by the model

The proposed data-driven segmentation can also uncover whether discrimination is exacerbated when a model is trained on the target task. This is indeed the case if the differences between $\Phi(S^{\mathcal{P}})$ values for the ground truth Y and the model's output $\hat{Y}$ are significant.

To demonstrate that the model amplifies discrimination on CelebA, we first compute FairGroups partition $\mathcal{P}$ by using Y as a target variable. We then use this partition to recalculate the values of $\Phi(S^{\mathcal{P}})$ for $\hat{Y}$ as a target (Equation 11). Figure 4 illustrates that the discrimination is significantly amplified for the darkest skin tones ($L < 43$) in the CelebA dataset, since the fairness measure $\Phi(S^{\mathcal{P}})$ calculated on the predicted $\hat{Y}$ is significantly lower than when computed on the ground truth Y.

6.5. Extension to partitioning skin tone in 2D

We now compare different partitions in the two-dimensional space (L, h). In the literature, the default partition into 4 groups with the thresholds $L \leq 60$ and $h \leq 55$ is used to conclude about fairness Thong et al. (2023). In Figure 5, we compare the partition $\mathcal{P}$ computed by FairGroups with this default partition in the left, and the confidence intervals for $\Phi(S^{\mathcal{P}})$ on the right. We see that FairGroups partition identifies two discriminated groups (Groups I and II), while the default partition captures only one (Group I). Indeed, according to the default partition, Group II has a value $\Phi(S^{\mathcal{P}})$ close to

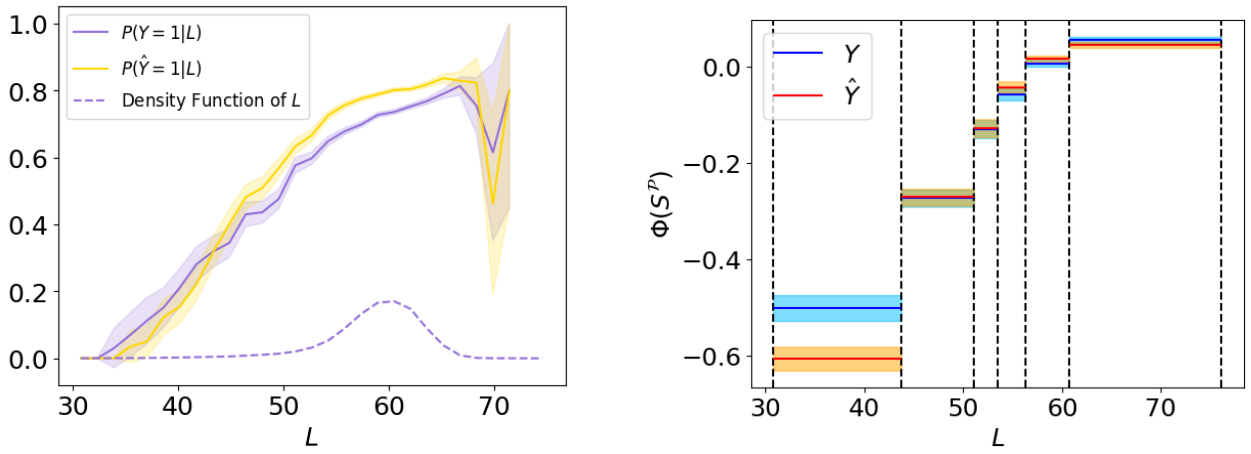


Figure 4 | Bias amplification for dark skin tones. Conditional distribution of "Attractive" $(Y/\hat{Y})$ w.r.t. L (left). FairGroups partition of L based on $\Phi(S^{\mathcal{P}})$ and Y (right).

0, suggesting that this group is not discriminated against. In contrast, the same group identified by FairGroups, located in the same quadrant, has a $\Phi(S^{\mathcal{P}})$ close to -0.2, indicating discrimination. Additionally, FairGroups partition finds a Group I facing a higher level of discrimination than the equivalent in default segmentation. Moreover, we note that the confidence intervals for $\Phi(S^{\mathcal{P}})$ in FairGroups partition are narrower compared to those of the default groups.

In Table 3, we compute the $\mathbb{V}ar\big(\Phi(\mathcal{S}^{\mathcal{P}})\big)$ values for the two partitions on CelebA and FFHQ, and observe significantly higher values for the one from FairGroups. We also apply the CelebA partition to FFHQ and demonstrate that it closely aligns with the skin tone partition of FFHQ (see Table 3), thus revealing how discrimination manifests in the skin tone space.

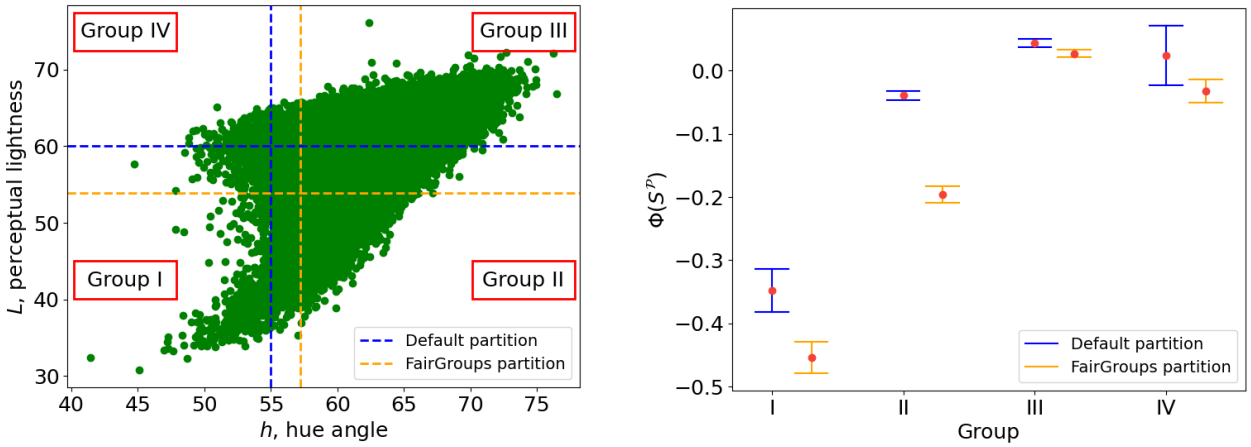


Figure 5 | Default and FairGroups Partition of (L, h) into 4 groups on CelebA (left), and their $\Phi(S^{\mathcal{P}})$ with Confidence intervals (right).

7. Application to model debiasing

In this section, we leverage the proposed partitions to mitigate bias at inference time through a post-processing strategy. Specifically, we adopt the optimal transport-based approach introduced in Gouic et al. (2020), which maps the conditional distribution of model scores for each group to a fair target distribution. This approach is supported by a theoretical result related to Statistical Parity (SP): the minimum classification error achievable under SP constraints corresponds to the solution of a Wasserstein barycenter problem Gouic et al. (2020). In this framework, the barycenter serves as a common target distribution to which group-conditional score distributions are aligned, ensuring SP across groups while minimizing the loss in predictive accuracy.

In practice, our methodology closely follows the procedure outlined in Gouic et al. (2020) and Xian et al. (2023), where the predicted score distribution for a group $\mathcal{P}_k$ is adjusted via a transformation of its cumulative distribution function. Let $\eta_k(X) = \mathbb{E}(Y|X, L \in \mathcal{P}_k)$ denote the conditional expected score for group $\mathcal{P}_k$; the goal is to map this distribution to a group-specific target distribution g_k^* by solving:

$$\min_{\substack{\forall k, g_k \in \mathcal{G} \\ \forall k,k' \|g_k - g_{k'}\|_\infty < \alpha}} \sum_{k=1}^{K} \frac{\mathbb{P}(L \in \mathcal{P}_k)}{2} W_1(\eta_k, g_k), \tag{16}$$

where $\alpha \in [0, 1]$ is a user parameter, W_1 the Wasserstein-1 distance and $\mathcal{G}$ the class of scorers in X space (g_k^* representing $\mathbb{P}(\hat{Y} = 1|X, L \in \mathcal{P}_k)$). We select W_1 as our application focuses on binary classification; please note that for regression tasks or scenarios requiring Equality of Odds, W_2 is more appropriate, as detailed in Gouic et al. (2020). The details for this post-processing is beyond the scope of this paper (see Appendix F for experimental illustrations).

Table 4 | Comparison of model accuracy, PR-AUC and fairness (HGR) without and with post-processing applied on partitions $\mathcal{P}$ given by different grouping methods.

	Accuracy	PR-AUC	HGR($\hat{Y}$, *ITA*)
No post-processing	0.793	0.824	0.126
$\mathcal{P}$: Fitzpatrick	0.782	0.815	0.095
$\mathcal{P}$: K-Means	0.783	0.813	0.065
$\mathcal{P}$: FairGroups	0.781	0.811	**0.039**

We evaluated this post-processing on the model g_θ obtained using the protocol of Section 6.2. We calculated PR-AUC Zhu (2004) and HGR Mary et al. (2019) between predictions and *ITA* using the corrected $\mathbb{P}(\hat{Y} = 1|X)$ predicted on the successive test sets. As shown in Table 4, post-processing with FairGroups achieves the lowest HGR score, indicating a significant reduction in dependence between the predictions and the sensitive attribute, with only a minor reduction of accuracy and PR-AUC. This appears to be the best trade-off between predictive performance and fairness, which demonstrates the effectiveness of FairGroups partition to mitigate the bias of a model.

Note that given a sample X, this post-processing requires to know the sensitive attribute L in order to debias the predicted $\hat{Y}$, which is accessible at inference time for our industrial scenario Robin et al. (2020).

8. Conclusion

In this paper, we proposed FairGroups, a data-driven algorithm aiming to divide a continuous sensitive variable into groups that reflects the unfairness of a given task. We observed on synthetic and real-

world datasets that our method successfully captures the specificities of the target task while taking into account the group sizes, contrary to predefined groups. FairGroups partitions obtain better evaluation metrics, and appear stable across various data distributions.

We performed extensive experiments on the case of skin tone on CelebA and FFHQ face pictures, which is a typical sensitive attribute that we extracted as a continuous physical quantity using a proprietary algorithm. We observed the property of monotonic fairness w.r.t. skin tone, for which K-Means grouping provides a practical and effective approximation to the exact solution. We could also apply and validate FairGroups partitions in one and two-dimensional skin tone spaces, and obtain the best trade-off between fairness and accuracy when applying on it model debiasing based on an optimal-transport post-processing.

As future work, we will study the impact of the choice of K and investigate its optimal value, and improve the scalability of our algorithm in the case of multi-dimensional sensitive attribute.

Ethics Statement

Considering the potential subjectivity of any index based on physical appearance, our work falls into the general problem of ensuring that, in deployment, the model remains unbiased with respect to skin tone. This means that the quality of predictions should not be affected by a user's skin tone, and thus leading to an infringement of fundamental rights as defined in Article 21 of the charter of fundamental rights of the European Union. Hence to guarantee legal conformity of the AI system, appropriate bias measurement tools are required to assess whether the model exhibits discriminatory behavior.

References

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2), 2022.

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

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International journal of cosmetic science*, 1991.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Jean De Rigal, Marie-Laurence Abella, Franck Giron, Laurence Caisey, and Marc André Lefebvre. Development and validation of a new skin color chart®. *Skin Research and Technology*, 13(1): 101–109, 2007.

Eustasio Del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms:

State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

Denia Garcia and Maria Abascal. Colored perceptions: Racially distinctive names and assessments of skin color. *American Behavioral Scientist*, 60(4):420–441, 2016.

Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: fundamentals and applications*. John Wiley & Sons, 2012.

Luca Giuliani, Eleonora Misino, and Michele Lombardi. Generalized disparate impact for configurable fairness solutions in ml. In *International Conference on Machine Learning*, pages 11443–11458. PMLR, 2023.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR, 2019.

Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20370–20382, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th international conference on data mining workshops*, pages 643–650. IEEE, 2011.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2018.

Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

Xinyue Li, Zhenpeng Chen, Jie M Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. Dark-skin individuals are at more risk on the street: Unmasking fairness issues of autonomous driving systems. *arXiv preprint arXiv:2308.02935*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Bao Chau K Ly, Ethan B Dyer, Jessica L Feig, Anna L Chien, and Sandra Del Bino. Research techniques made simple: cutaneous colorimetry: a reliable technique for objective skin color measurement. *Journal of Investigative Dermatology*, 140(1):3–12, 2020.

J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

Emmanuel Malherbe, Michel Remise, Shuai Zhang, and Matthieu Perrot. Skin tone diagnosis in the wild: Towards more robust and inclusive user experience using oriented aleatoric uncertainty. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–19, 2022.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR, 2019.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 1971.

Laurent Risser, Alberto Gonzalez Sanz, Quentin Vincenot, and Jean-Michel Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *Journal of Mathematical Imaging and Vision*, 64(6), 2022.

KIPS Robin, TRAN Loïc, Emmanuel Malherbe, and Matthieu Perrot. Beyond color correction: Skin color estimation in the wild through deep learning. *Electronic Imaging*, 32:1–8, 2020.

Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert Sève. *Science de la couleur: Aspects physiques et perceptifs*. Chalagam éd., 2009.

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4903–4913, 2023.

Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022.

Ian L Weatherall and Bernard D Coombs. Skin color measurements in terms of cielab color space values. *Journal of investigative dermatology*, 1992.

Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology*, 151(8):902–903, 2015.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International conference on machine learning*, pages 37977–38012. PMLR, 2023.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6, 2004.

A. Relation with Disparate Impact in the Binary Case

Detailed Proof of Proposition 4.1.

Proof. When $S^{\mathcal{P}}$ is binary, that is, there are only two groups, the variance in Equation (12) rewrites

as follows.

$$
\begin{aligned}
&\mathbb{V}ar\Big(\mathbb{P}(Y=1|S^{\mathcal{P}})-\mathbb{P}(Y=1)\Big)= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=0)-\mathbb{P}(Y=1)\Big]^2+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1)\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=1)-\mathbb{P}(Y=1)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1)\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=1)- \\
&-\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\Big[(1-\mathbb{P}(S^{\mathcal{P}}=0))\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1)\Big[(1-\mathbb{P}(S^{\mathcal{P}}=1))\mathbb{P}(Y=1|S^{\mathcal{P}}=1)- \\
&-\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(Y=1|S^{\mathcal{P}}=0)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\Big[\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1)\Big[\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(Y=1|S^{\mathcal{P}}=1)- \\
&-\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(Y=1|S^{\mathcal{P}}=0)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(S^{\mathcal{P}}=1)^2\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1)\mathbb{P}(S^{\mathcal{P}}=0)^2\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=1)- \\
&-\mathbb{P}(Y=1|S^{\mathcal{P}}=0)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(S^{\mathcal{P}}=1)(\mathbb{P}(S^{\mathcal{P}}=0)+ \\
&+\mathbb{P}(S^{\mathcal{P}}=1))\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=0)-\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2 \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(S^{\mathcal{P}}=1)\Big[\mathbb{P}(Y=1|S^{\mathcal{P}}=0)- \\
&-\mathbb{P}(Y=1|S^{\mathcal{P}}=1)\Big]^2= \\
&=\mathbb{P}(S^{\mathcal{P}}=0)\mathbb{P}(S^{\mathcal{P}}=1)DI^2=\pi(1-\pi)DI^2.
\end{aligned}
\tag{17}
$$

□

B. K-Means in the case of non-monotonic fairness: disconnected clusters

If fairness is not monotonic (Definition 5.1), then K-Means may produce disconnected groups as illustrated in Figure 6. We observe that pink group is separated into two segments by blue group, since K-Means only use values of Ψ for grouping.

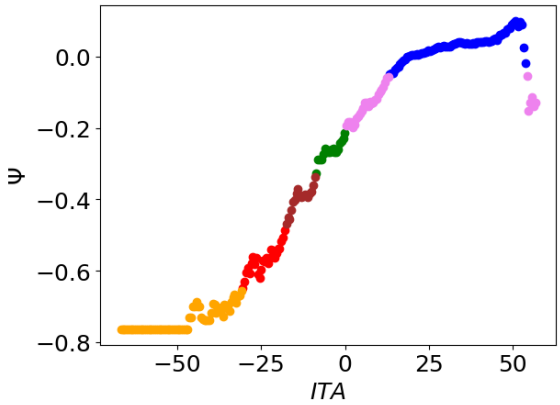


Figure 6 | K-Means produces disconnected groups in the case of non-monotonic fairness.

C. Exhaustive search with optimal precalculation of fairness metric Φ

We can approximate upper triangular matrix U_{Ψ_Λ}, defined in 5.2, as follows

$$
\begin{aligned}
U_{\Psi_\Lambda} &\approx \\
&\approx \begin{bmatrix} \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1, L_i \in [\lambda_0, \lambda_1]\}} & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1, L_i \in [\lambda_0, \lambda_{M-1}]\}} \\ 0 & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1, L_i \in [\lambda_1, \lambda_{M-1}]\}} \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1, L_i \in [\lambda_{M-2}, \lambda_{M-1}]\}} \end{bmatrix} \odot \\
&\odot \begin{bmatrix} \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_0, \lambda_1]\}} & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_0, \lambda_{M-1}]\}} \\ 0 & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_1, \lambda_{M-1}]\}} \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{M-2}, \lambda_{M-1}]\}} \end{bmatrix}^{\circ -1} - \\
&- \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}}}{N} \times U,
\end{aligned}
\tag{18}
$$

where $\odot$ is Hadamard (element-wise) product and $^{\circ -1}$ is Hadamard inverse, U is an upper triangular matrix of size $(M-1) \times (M-1)$ with ones on the main diagonal and in all elements above

it. Therefore, we have

$$\begin{aligned}\Psi_{j_1,j_2} \approx \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_2}]\}}}{\sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_2}]\}}} - \\ - \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}}}{\sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}}} - \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}}}{N} = \\ = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}} \left(\mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_2}]\}} - \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}} \right)}{\sum_{i=1}^{N} \left(\mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_2}]\}} - \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}} \right)} - \\ - \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}}}{N} = \\ = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}} \mathbb{1}_{\{L_i \in [\lambda_{j_1}, \lambda_{j_2}]\}}}{\sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{j_1}, \lambda_{j_2}]\}}} - \frac{\sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}}}{N}.\end{aligned} \tag{19}$$

From Equations (18) and (19), we deduce how the values of Ψ_{j_1,j_2} can be computed using dynamic programming strategy.

Algorithm 1: `count_on_all_ranges(sum_by_bin)`: Iteratively counting the number of observations $\forall\ [\lambda_{j_1}, \lambda_{j_2}]$, where $j_1 < j_2$, $j_1 = 0, \ldots, M-1$, $j_2 = 1, \ldots, M$

Data: *sum_by_bin* - 1D-array of size M, representing $\sum_{L \in [\lambda_{j_1}, \lambda_{j_1+1}]} Z$ for observations L, Z (user-defined values)
Result: *sum_inside_range* - 2D-array of size $M \times M$, representing $\sum_{L \in [\lambda_{j_1}, \lambda_{j_2}]} Z$ for observations L, Z

```
sum_inside_range ← [[...]] ;                          /* Initialize array*/
sum_inside_range[0][0] ← sum_by_bin[0];
for j2 ← 1 to M − 1 do
    sum_inside_range[0][j2] ← sum_inside_range[0][j2 − 1] + sum_by_bin[j2];
for j1 ← 1 to M − 1 do
    for j2 ← j1 to M − 1 do
        sum_inside_range[j1][j2] ← sum_inside_range[j1 − 1][j2] − sum_by_bin[j1 − 1];
return sum_inside_range
```

D. Partition of L on synthetic dataset

Figure 8 shows partitions of $L \sim \mathcal{U}(0, 100)$ on synthetic data.

Figure 9 shows partitions of $L \sim \mathcal{N}(50, 20)$ truncated on $[0, 100]$ on synthetic data.

E. Partitions of skin tone in two groups

We start by comparing partitions of a one-dimensional skin tone variable L with the "predefined" groups following Thong et al. (2023). In fairness analyses, L is usually divided into two groups using the default threshold from Ly et al. (2020). When $L > 60$, it corresponds to a light skin tone, and conversely, to a dark skin tone when $L \leq 60$.

Algorithm 2: Calculating values of Ψ_{j_1,j_2} for all possible intervals $[\lambda_{j_1}, \lambda_{j_2}]$, given the grid $\lambda_0 < \lambda_1 < ... < \lambda_M$

Data: Y, L - two 1D-arrays of size N representing observations of random variables
Grid of size M given by increasing values $\lambda_0, ..., \lambda_M$
Result: *phi_by_range* - 2D-array of size $M \times M$, representing U_{Ψ_Λ}

$N \leftarrow size(Y)$; `/* Total number of observations*/`
$N_positive \leftarrow \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=1\}}$; `/* Total number of positive observations*/`
$nb_obs_by_bin \leftarrow [\dots]$; `/* Initialize array*/`
$nb_pos_obs_by_bin \leftarrow [\dots]$; `/* Initialize array*/`
for $j_1 \leftarrow 1$ *to* M **do**
  $nb_obs_by_bin[j_1] \leftarrow \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}}$;
  $nb_pos_obs_by_bin[j_1] \leftarrow \sum_{i=1}^{N} \mathbb{1}_{\{L_i \in [\lambda_{j_1-1}, \lambda_{j_1}]\}} \mathbb{1}_{\{Y_i=1\}}$;
$nb_obs_inside_range \leftarrow$ `count_on_all_ranges(`$nb_obs_by_bin$`);`
$nb_pos_obs_by_range \leftarrow$ `count_on_all_ranges(`$nb_pos_obs_by_bin$`);`
`/* Array-wise operations below */`
$proba_by_range \leftarrow \frac{nb_pos_obs_inside_range}{nb_obs_inside_range}$;
$phi_by_range \leftarrow proba_by_range - \frac{N_{positive}}{N}$;
return *phi_by_range*

Figure 10 shows the three different partitions of L. K-Means and FairGroups partitions identify well the least and the most privileged group. However, K-Means does not account for the impact on the least privileged group. As a result, it produces a smaller least privileged group compared with FairGroups, with a larger confidence interval.

From Figure10, we conclude that when we default partition L with the threshold used in the literature, we cannot see the discrimination that is present in the model w.r.t. the skin tone. Both groups in the default partition have Φ close to 0, therefore one might erroneously conclude that the model does not discriminate w.r.t. skin color.

Furthermore, we examine how groups in each partition differ in terms of the level of discrimination. Table 5 summarizes the metrics for all partitions.

Table 5 | Comparison of partitions of L based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(\hat{Y} = 1|S^{\mathcal{P}}) - \mathbb{P}(\hat{Y} = 1)$, $K = 2$.

Data for computing $\mathcal{P}$	CelebA		FFHQ	CelebA & FFHQ
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on CelebA	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	Rand Index between both partitions, on FFHQ
$\mathcal{P}$: $L \leq 60$	0.045	0.071	0.071	1
$\mathcal{P}$: K-Means	0.067	0.084	0.084	0.938
$\mathcal{P}$: FairGroups	**0.079**	**0.097**	**0.101**	0.945

Moreover, we measure whether the partitions found for the skin color on CelebA dataset generalize well to the second dataset, FFHQ. To do so, we first compute the variance $\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ observations grouped with the partitions $\mathcal{P}$ computed on CelebA. These values can be compared among various partition methods and with the variance for partitions computed on FFHQ, thus more exact. For each partition method, we finally compare the partitions computed on both dataset using the Rand Index. From the results in Table 5, we observe a good generalization of each method

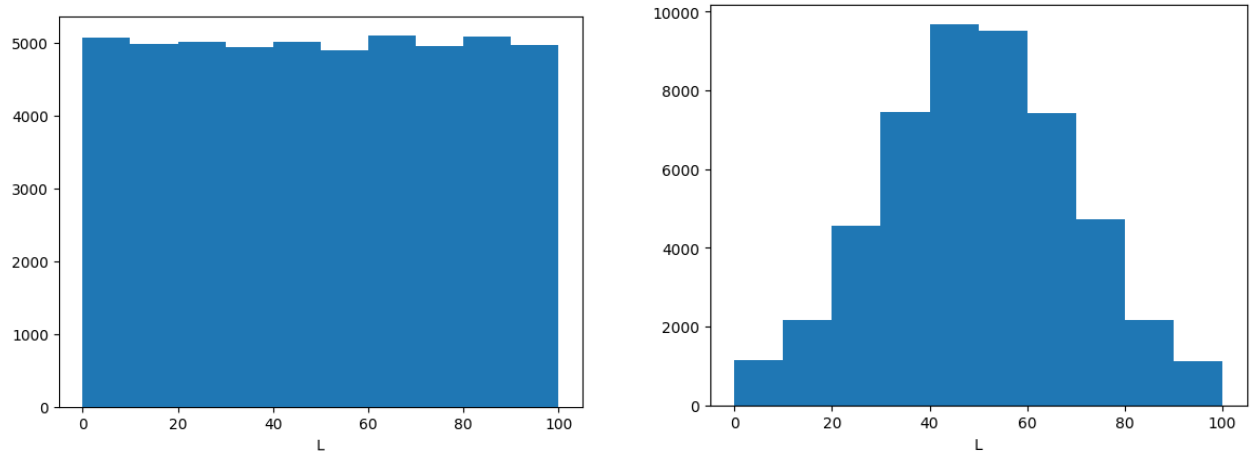


Figure 7 | Synthetic datasets. Generated observations, $L \sim \mathcal{U}(0, 100)$ (left). Generated observations, $L \sim \mathcal{N}(50, 20)$ truncated on $[0, 100]$ (right).

since values of $\mathbb{V}ar\Big(\Phi(S^{\mathcal{P}})\Big)$ are close on both datasets. In addition, the Rand Index is closer to 1 for FairGroups, which indicates a higher the stability of the skin tone partition. In these experiments, it means that our methods identify the partitions which reveal how the discrimination is manifested within the space of sensitive attributes, relatively independently of the data distribution.

F. Debiasing using optimal transport

To demonstrate the mechanism of optimal transport-based post-processing, we operate within the setting described in Section 6.3, which details the dataset and partitioning scheme used. In this context, we mitigate the bias of the model g_θ using the FairGroups partition. Figure 11 illustrates how the cumulative distribution functions (CDFs) of each group $\mathcal{P}_k$ in the partition $\mathcal{P}$ are affected by the post-processing step. Each solid line represents the CDF of a group $\mathcal{P}_k$ prior to post-processing, while the corresponding dashed line depicts the CDF after the application of the post-processing procedure. From Figure 11, we observe that the CDFs across groups become more aligned following post-processing, indicating a reduction in group-level disparities.

However, it is important to note that post-processing does not map all group distributions to a single, identical CDF. Rather, the residual discrepancies between the post-processed CDFs are modulated by the parameter α, which governs the strength of the debiasing and enables a controlled trade-off between predictive accuracy and group fairness.

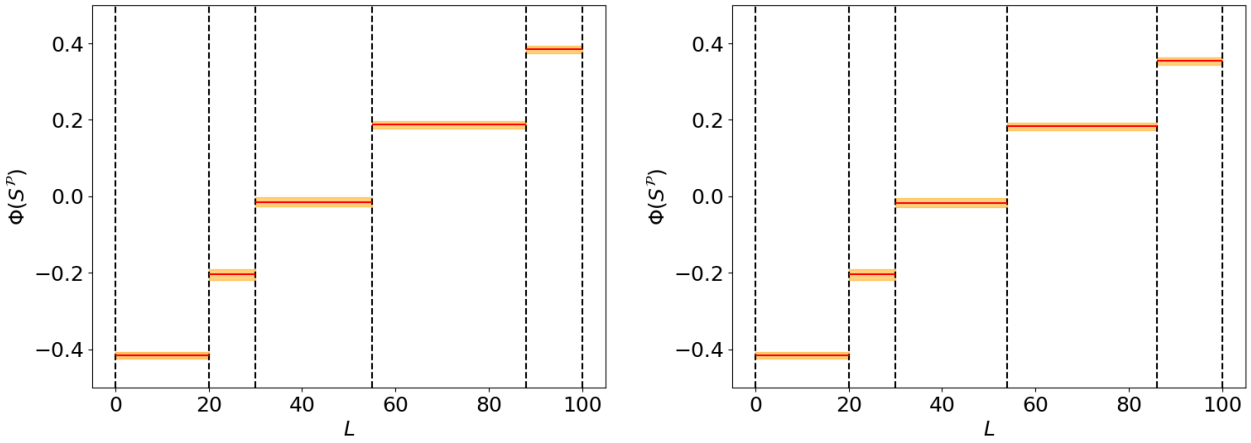


Figure 8 | Partitions on synthetic dataset with uniform distribution of L. Ground truth partition (left). K-Means partition (right).

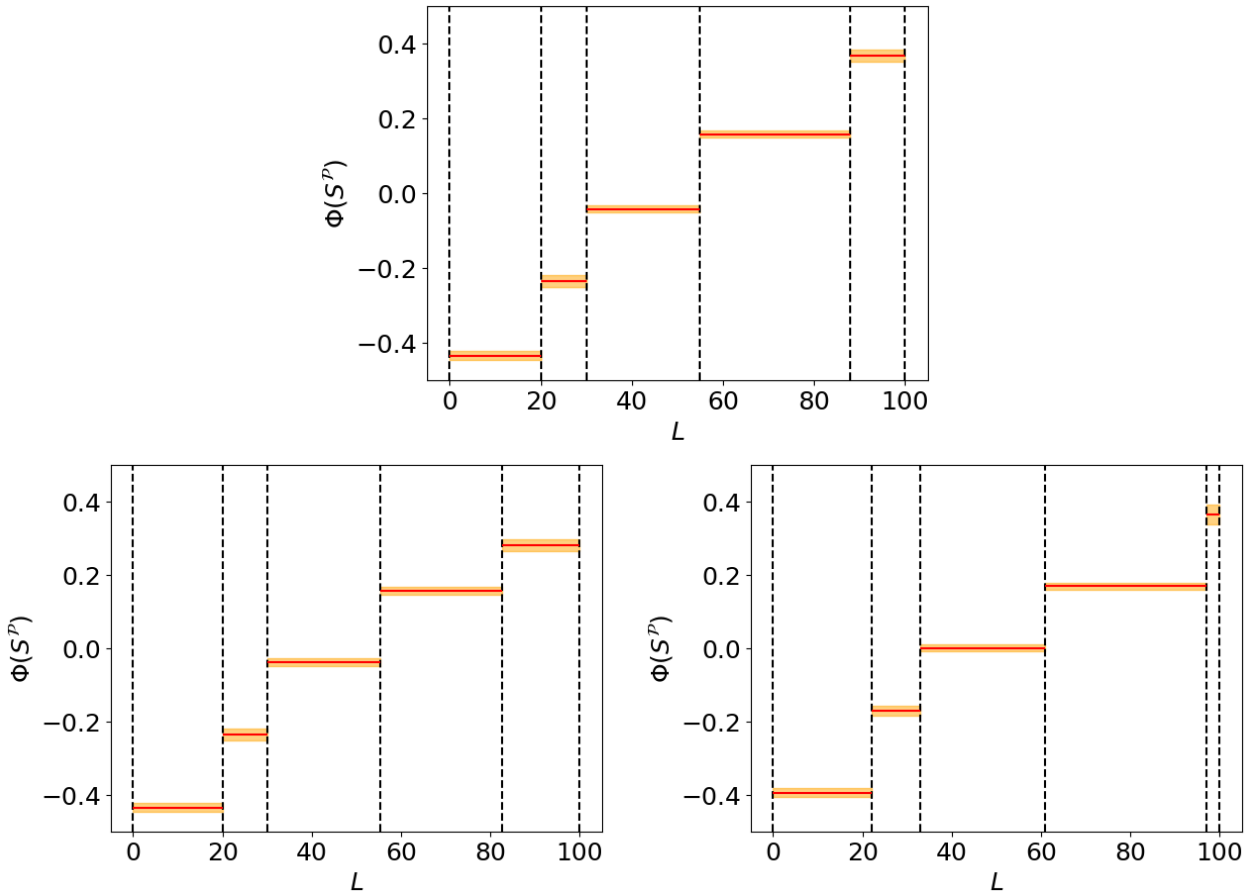


Figure 9 | Partitions on synthetic dataset with truncated Gaussian distribution of L. Ground truth partition of L (top). FairGroups partition of L (bottom left). K-Means partition of L (bottom right).

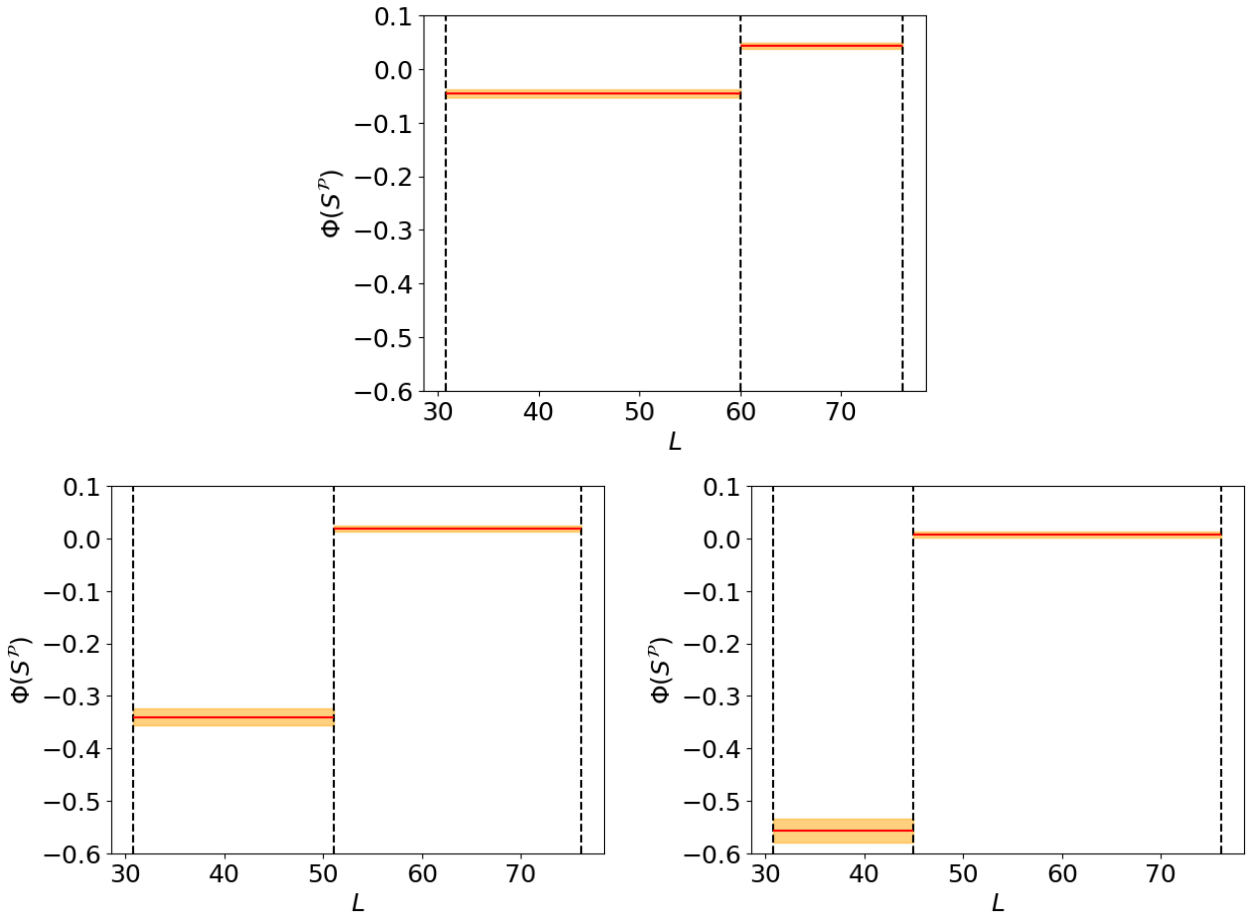


Figure 10 | Partitions of L into the least and the most privileged groups on CelebA. Default partition of L (top). FairGroups partition of L (bottom left). K-Means partition of L (bottom right).

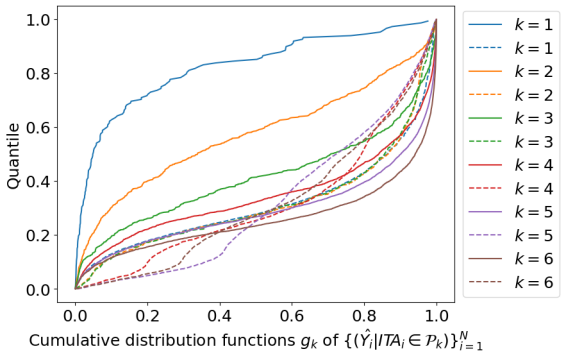


Figure 11 | Cumulative distribution functions g_k before (solid lines) and after (dashed lines) post-processing on FairGroups partition of *ITA*.