*Article*

SM+S
social media + society

Social Media + Society
April-June 2024: 1–15

DOI: 10.1177/20563051241239295
journals.sagepub.com/home/sms
Sage

# Mirror, Mirror on the Wall, Who Is the *Whitest* of All? Racial Biases in Social Media Beauty Filters

**Piera Riccio[1], Julien Colin[1], Shirley Ogolla[2], and Nuria Oliver[1]**

**Abstract**

Digital beauty filters are pervasive in social media platforms. Despite their popularity and relevance in the selfies culture, there is little research on their characteristics and potential biases. In this article, we study the existence of racial biases on the set of aesthetic canons embedded in social media beauty filters, which we refer to as the *Beautyverse*. First, we provide a historic contextualization of racial biases in beauty practices, followed by an extensive empirical study of racial biases in beauty filters through state-of-the-art face processing algorithms. We show that beauty filters embed Eurocentric or *white* canons of beauty, not only by brightening the skin color, but also by modifying facial features.

**Keywords**

beauty filters, social media, racial bias

## Introduction

Technological development is a socially entangled process that reflects the values and biases of the society where it takes place (Ash et al., 2018). Social media platforms, with billions of users worldwide, are a clear example of such a process. In less than three decades of existence, they have emerged as a key element that conforms the social fabric of human communities, allowing their members to connect, interact, and share information. They have created new opportunities for personal and professional networking, learning, entertainment, activism, and self-expression.

Historically, the marginalization of women from the use of technology has led to the inclusion of gendered notions in technological design (Cockburn, 1983; Wajcman, 2004). In the case of social media, many of the functionalities and algorithms used in these platforms emphasize physical beauty as a valuable attribute for women, to the point that female users tend to self-objectify in search of social validation (Winch, 2013; D. Zheng et al., 2019). Self-objectification influences self-presentation practices in many ways, such as posting edited selfies on social media to appear more attractive (Hong et al., 2020). Among the available digital beauty enhancement tools for photos and videos, social media platforms favor beauty filters, mostly designed by their users (Riccio et al., 2022). These filters leverage computer vision algorithms for face and facial feature detection and augmented reality (AR) to overlay in real-time digital elements that modify the features of the detected face, as depicted in Figure 1. The changes are typically applied to the skin, the eyes and eyelashes, the nose, the chin, the cheekbones, and the lips, creating a visually enhanced or *beautified* version of the user. The filters often reflect non-realistic beauty standards, making users believe that a *better* version of themselves is not only possible, but even needed and desirable, ultimately impacting self-perception and self-esteem (Eshiet, 2020; Grossman, 2017).

In this article, we investigate the existence of racial biases in social media beauty filters and their potential negative impact not only on the well-being of social media users—particularly women and people of color—but also on society at large. We refer to the *Beautyverse* as the set of aesthetic canons embedded in today's beauty filters. The often unreachable beauty ideals reflected in the *Beautyverse* may be internalized by users, who actively aspire to look like their beautified digital versions, reinforcing those standards even further through systematic social comparison (Lamp et al., 2019; Myers & Crowther, 2009). In such a complex scenario, it is of utmost importance to investigate the multiple facets of

[1]ELLIS Alicante, Spain
[2]Alexander von Humboldt Institute for Internet and Society, Germany

**Corresponding Author:**
Piera Riccio, ELLIS Alicante, Alicante 03001, Spain.
Email: piera@ellisalicante.org

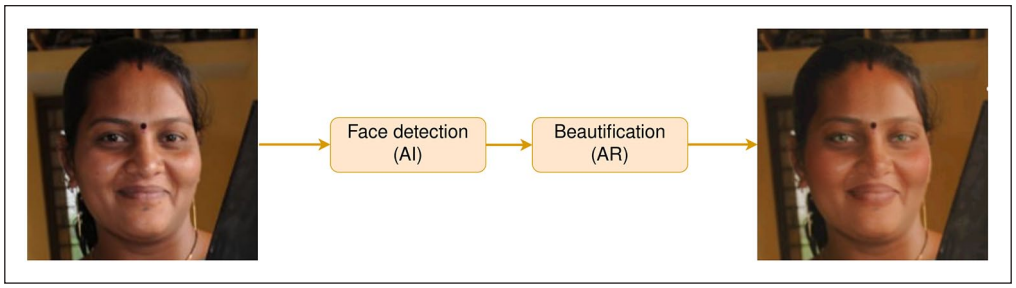


**Figure 1.** Flow of a beauty filter applied to an input image (from left to right). First, computer vision algorithms are applied for face and facial feature detection, followed by the use of augmented reality (AR) methods to overlay in real-time digital content that modifies (beautifies) the face. Possible changes include lightening and correcting imperfections in the skin, making the eyelashes longer, the cheekbones more prominent, the lips fuller and pinker, and the eyes bigger and of lighter color. The original (input) image on the left is from the FAIRFACE data set (Karkkainen & Joo, 2021); the (beautified) image on the right is from the FAIRBEAUTY data set (Riccio et al., 2022). The images depict the same individual.

the *Beautyverse*, especially its potential negative impact. In qualitative studies, scholars have argued that beauty filters perpetuate racism (Mulaudzi, 2017) and reinforce Eurocentered ethnic features (S. Li, 2020). In other words, the facial aesthetics embedded in such filters are inherently *white* (Jagota, 2016; Shein, 2021). Note that in this article, we use the terms Eurocentric and *white* indistinguishably. Furthermore, our definition of *whiteness* does not simply refer to the skin tone, but also includes other facial features, such as "nose and eyes shape, lips and hair type" (Dyer, 2017).

Despite the wide adoption of beauty filters and their impact on millions of users, there is little quantitative research to date on their characteristics and particularly on their biases. In this article, we aim to unravel and empirically evaluate the existence of racial biases in the *Beautyverse* by means of state-of-the-art machine learning-based computer vision algorithms. The expected impact of this research is to initiate an *interdisciplinary* discussion on the technical and social implications of this phenomenon.

## Related Work

In recent years, different research communities have investigated the increasingly popular use of digital beauty filters. In this section, we provide an overview of the most relevant previous work in Computer Science, Psychology, and Sociology.

Machine learning-based methods in computer vision are the main technical tool to study beauty filters from a Computer Science perspective. This field relies on the use of publicly available, standardized data sets of faces to enable the comparison of different approaches and the reproducibility of the results. However, to date, there are few public data sets of *beautified* faces (Hedman et al., 2022) and the use of faces downloaded from social media platforms is not possible unless there is explicit, informed consent from each of the individuals whose faces would be analyzed.

Bharati et al. (2017) created a data set of beautified faces from 600 different individuals belonging to three different ethnicities (Indian, Chinese, and Caucasian) using three commercial tools for beautification: Fotor,[1] BeautyPlus,[2] and PortraitPro Studio Max.[3] In this case, the beautification techniques modified the skin, facial structure, eyes, and lips of the original faces. In addition to sharing the data set, the authors proposed a novel semi-supervised autoencoder to detect whether the images had been retouched. Hedman et al. (2022) generated a beautified version of the Labeled Faces in the Wild (LFW) (Huang et al., 2008) faces data set. They performed an analysis of the impact of beauty filters on face recognition models. However, the beautification process only involved the superimposition of simple AR elements that create occlusions on the face. Mirabet Herranz et al. (2022) studied the impact of beauty filters on both face recognition and estimation of biometric features by beautifying the CALWF (T. Zheng et al., 2017) and VIP_attribute (Dantcheva et al., 2018) data sets. Riccio et al. (2022) proposed OPENFILTER, a framework to apply any filter available on social media platforms to existing collections of faces, sharing the beautified version of two large and publicly available face data sets (FAIRBEAUTY and B-LFW). In their work, they highlight that beauty filters increase the visual similarity among individuals without significantly impacting the performance of state-of-the-art face recognition systems. In addition, preliminary work on these data sets has hinted that racial biases might be present in beauty filters (Riccio & Oliver, 2023).

Beyond Computer Science, related work in Psychology and Sociology serves as an inspiration and provides a deeper understanding of the beautification phenomenon. Early work by Felisberti and Musholt (2014) focused on the impact of beauty filters on self-perception and self-esteem. The authors carried out a user study with 33 participants (23 females), finding that low self-esteem impacts the desirability of certain physical features, in particular, smaller nose and bigger eyes. Fribourg et al. (2021) analyzed the impact of beauty filters on the perception of attractiveness, intelligence, and personality through a user study with 20 males and 20 females. They reported that the perception of others

is often transferable to self-perception and that AR beauty filters seemed to decrease self-acceptance. Bakker (2022) presented a study with 103 female participants of the internalization of beauty ideals from beauty filters, highlighting that women using these filters internalize these ideals more easily, hence suffering from body dissatisfaction.

Shifting the focus to the subject matter of our work, we report on previous research that emphasizes the connection between beauty filters and racism/colorism from different cultural perspectives. Siddiqui (2021) studied the relationship between social media beauty filters and the deeply rooted colorism in Indian society. As in other countries in Asia, Africa, and South America,[4] a fairer skin is considered more attractive and an enabler of social opportunities in India. The author interviewed 26 young women, and concluded that beauty filters imitate hyper-realistic and fair-skinned beauty ideals, allegedly *emancipating* women but strongly impacting their self-esteem. Peng Peng (2021) provided a techno-feminist analysis of the development of beauty filters applications and the so-called *wanghong* beauty ideal in China, which is "characterized by big eyes, double eyelids, white skin, high-bridged nose, and pointed chin" (A. K. Li, 2019). Through a case study of the *BeautyCam*[5] application, the author suggested that the driving force for the development of such applications in Chinese society is a wave of pseudo-feminism. These applications are designed to target female users, with the argument that improving the physical appearance is a means to obtain social empowerment and emancipation. Such a claim implicitly embeds and propagates a gendered approach in the design of technology, and the need for women to adhere to an "ultra-feminine" physical representation.

While previous work has tackled this matter from a qualitative perspective, this article contributes with an extensive quantitative and interdisciplinary study of racial biases in beauty filters. In particular, our contributions are four-fold:

1. We contextualize our technical work with a historical overview of racial biases in the social understanding of beauty, emphasizing how they impact other dimensions of collective and individual affirmation in society.
2. We empirically study the existence of racial biases in the *Beautyverse* by applying machine learning-based race classification algorithms to images of beautified and non-beautified faces.
3. We investigate the characteristics of such racial biases through a state-of-the-art explainable artificial intelligence (AI) method.
4. We draw six insights and implications regarding the use of beauty filters in social media.

## Beauty: A *White* Social Opportunity

In this section, we contextualize and motivate our work through a summary of the biases that have characterized beauty practices in human history, and their direct consequences for the members of marginalized communities.

Research has shown that beauty matters: people who are perceived as more beautiful are more likely to be successful in life by, for example, achieving better grades in school (Talamas et al., 2016), promotions and higher income at work (Morrow et al., 1990), more lenient criminal sentences (Stewart, 1980), and a better social status overall (Frieze et al., 1991). In parallel with the presumption that beauty standards are determined by culture and personal biases (Sartwell, 2012), studies have demonstrated that symmetry, averageness, and sexual dimorphism are important evolutionary factors in determining attractiveness across cultures (Rhodes, 2006). In particular, physical appearance is important especially for teenagers: female adolescents tend to have the highest rates of mental health issues, and particularly anxiety and depression related to body dissatisfaction (Alm & Låftman, 2018; McLean et al., 2022; Pivnick et al., 2022).

Social media has become an indispensable component in young people's lives (Boyd, 2008), with both positive and negative effects, particularly on mental health (Richards et al., 2015). We know that our digital self and its perception impact our analog self. For instance, having a highly sexualized virtual reality avatar affects how females act both online and offline, increasing their sense of self-objectification (Fox & Bailenson, 2009; Maloney & Robb, 2019). Moreover, selfie dysmorphia has led to an increase in plastic surgery to look like the beautified social media self which, in many cases, reflects an unattainable ideal of beauty (Cristel et al., 2021; Othman et al., 2021; Perrotta, 2020).

In this context, *white* beauty standards predominate in our society and current advancements in computer vision and AR, combined with the massive adoption of increasingly powerful smartphones and the ubiquitous use of social media platforms, threaten to amplify the predominance of such standards. Historically, structural systems privilege White people in every conceivable social, political, and economic opportunity (Fanon, 2008; Kilomba, 2021). Since Europeans colonized the world—occupying land, appropriating resources, and establishing slave trades—descendants of the colonized countries have relied on migrating to places where White people come from to find better life opportunities.

The social advantage conferred to White(r) individuals manifests itself in the two closely related concepts of colorism and racism,[6] which imply a hierarchical positioning of people according to their skin color, ethnicity, and other physical features. Colorism occurs within a particular racial or ethnic group, based on skin tone, such that, lighter-skinned individuals are preferred over darker-skinned ones. Racism takes place across different racial and ethnic groups, based on perceived differences in physical features, cultural practices, and social customs. Racism plays a role in shaping beauty standards, as it often involves a preference for Eurocentric features, such as straight hair, light skin color, light-colored, and large eyes, over features that are more commonly associated with non-European cultures.

As a consequence, lighter-skinned individuals—both from the same racial group and across racial groups—have been awarded privileges past and present (Mire, 2001). After being subjected to White people's privileges and their corresponding beauty standards for so long, it is no surprise that people of color might despise the color of their own skin, eyes, and hair, aiming for a *whiter* look (Tate & Fink, 2019). Today, most countries have banned skin-whitening products because of their toxic ingredients and damaging impact on mental health. Still, people—and particularly women—in the Global South are willing to take great health risks to change their appearance, so that, it conforms to *white* beauty canons and hence increases their chance to achieve higher socio-economic power (Adawe & Oberg, 2013).

Beyond colonization, globalization also plays a key role in influencing beauty standards around the world, which results in Western European and American beauty ideals being globally embraced (Dimitrov & Kroumpouzos, 2023). The fashion, media, cosmetics, and movie industries significantly contribute to the global culture and the definition of canons of beauty (Yan & Bissell, 2014). This globalization process is also reflected on how social media impacts the perception of beauty worldwide through systematic comparison with beauty influencers from the Western world (Ward & Paskhover, 2019).

While colonization and globalization are determining factors in establishing beauty standards worldwide, additional factors need to be considered as every cultural context is unique. For example, scholars have argued that the *shadeism* existing in the Indian sub-continent is not only related to the need of mimicking "colonial whiteness" (Fischer-Tiné, 2009) but also has a locally pre-colonial rooted history (Kullrich, 2022) as fair-skin tones were associated with upper castes: lightening the skin in India is not necessarily a matter of changing "color" but a matter of changing "shade" to hide the social and working status (Kullrich, 2022). In Africa, researchers have highlighted how the dominant homogenized representation of beauty in African magazines promotes "western" femininity. As a consequence, it is expected that Black women feel the need to adhere to *white* beauty ideals to feel beautiful (Akinro & Mbunyuza-Memani, 2019). At the same time, research has shown that within racial minorities in the United States, Asian women tend to idealize and follow mainstream *white* beauty standards more than Black women (Chin Evans & McConnell, 2003). With respect to Asia, the influence of Western canons of beauty is combined with their own traditional views on beauty, reflected in their art, literature and philosophy (Samizadeh, 2022). For example, a fair skin with smooth texture—so-called *porcelain* or *milk-like* skin—has been revered for centuries as illustrated in Asian poetry and literature. Furthermore, the change of facial features is no longer perceived as a disrespect to the ancestors due to globalization and the wide availability of non-surgical and surgical cosmetic procedures (Kim, 2003) to the point that South Korea is referred to as "the plastic surgery capital of the world," representing a 25% of the global beauty market[7] and China's cosmetic surgery industry is one of the largest and fastest-growing in the world. Finally, scholars have recently reported on the under-studied beauty and body image ideals in postcolonial Latin American countries and US Latinx women (Gruber et al., 2022), finding that beauty is primarily rooted in a Westernized and *white* ideology (Figueroa, 2021) (light skin tone and hair color, small noses) combined with a culturally rooted curvaceous figure (Lloréns, 2013).

In summary, while acknowledging that different cultural contexts follow diverse and unique trajectories to shape their beauty standards, we also highlight that the widespread use of beauty filters is turning beauty into a globally shared experience, which prompts research efforts like ours. In this article, we study whether *white* beauty standards are indeed present in today's beauty filters through the lens of state-of-the-art face processing algorithms. Such a computational approach enables us to study this phenomenon at scale, with thousands of images, and in a consistent, systematic and potentially more objective manner. We articulate our work on this topic by means of the following research questions (RQs), which we address in two different experiments, described in the next section:

**RQ1**: Do beauty filters make people conform with Eurocentric (*white*) beauty standards?

**RQ2**: If so, how do beauty filters embed *white* canons of beauty?

## Experiments

In this section, we describe the experiments that we carried out to tackle RQ1 and RQ2. The results can be reproduced through our GitHub repository.[8]

### *Data Sets and Data Pre-processing*

We perform our analyses on the widely used FAIRFACE (Karkkainen & Joo, 2021) data set and its beautified version, FAIRBEAUTY (Riccio et al., 2022). FAIRFACE is a collection of 108,501 face images, designed to represent a diverse set of faces. It contains examples of face images from seven different race groups and in a variety of resolutions, poses, and expressions as the faces were captured in-the-wild from Flickr, Twitter (now X), newspapers, and the web. In addition to the faces, the data set contains their attributes as metadata, including the label *race*, for which seven different categorical values are available, namely, Black, East Asian, Indian, Latino Hispanic, Middle Eastern, Southeast Asian, and White. Regarding gender, the data set provides a binary variable (male/female) for each image, such that, only two genders are available. FAIRBEAUTY is a beautified version of FAIRFACE, using eight Instagram beauty filters applied to the FAIRBEAUTY images. As explained in the work of Riccio et al.

**Figure 2.** Example the four versions of the same image considered in RQ1. From left to right, original image ($x$), beautified image ($x_b$), non-beautified image in Blur Case 1 ($x_g^{\sigma=2}$), and non-beautified image in Blur Case 2 ($x_g^{\sigma=3}$).

(2022), the filters were selected by popularity. All the filters were created by users with thousands of followers who define themselves as filters creators. Note that we perform our experiments without dividing the images according to the filters they are beautified with, as all filters perform similar facial transformations and we do not intend to compare them. Instead, our goal is to assess the presence of racial biases in a similar setting to that of social media where different beauty filters co-exist.

Beyond demographic diversity, the images in the two data sets contain one or more individuals in different poses, scenarios and with a variety of facial expressions. As the beauty filters are typically applied to selfies, we selected a subset of the examples in the FAIRFACE/FAIRBEAUTY data sets that satisfied the following conditions: (a) they had a similar resolution above a minimum level; (b) the faces were in a frontal pose, as similar as possible to a selfie; and (c) there would yield a gender and race-balanced set with roughly the same number of images per gender and race.

Applying these conditions, we selected a total of 3,164 images, depicting the face of single individuals with frontal or nearly frontal poses, and having comparable resolution. Figure 1 exemplifies a canonical example of the selected images. The images are balanced across gender and racial categories: on average, we select 452 images per race (with a minimum of 420 and a maximum of 484, respectively, for Southeast Asian and Black). We address our RQs on this test set.

## *RQ1: Do Beauty Filters Make People Conform With Eurocentric (*White*) Beauty Standards?*

*Problem Formulation and Setup.* We consider a set of images $X \subset \mathcal{X}$, where $\mathcal{X}$ is the space of all possible images, and a set $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ of transformation functions applied to $\mathcal{X}$. We denote by $x_b$ the beautified version of $x$, that is, $x_b = t_b(x)$, with $t_b \in \mathcal{T}$ being the beautification filter applied to image $x$. In our experiments, $x_b$ is an image in the FAIRBEAUTY data set and $x$ is its corresponding, original image in the FAIRFACE data set. Following the procedure described in the work of Riccio et al. (2022), we also consider two sets of images according to two additional transformations, namely, blurring by means of Gaussian filters of different radius: $t_g^{\sigma=n} \in \mathcal{T}$ refers to the application of a blurring Gaussian filter of radius $n$ on image $x$ to generate $x_g^{\sigma=n}$. An example of the original image $x$, its beautified version $x_b$, and its blurred versions with filters of radius 2 ($x_g^{\sigma=2}$) and 3 ($x_g^{\sigma=3}$) are provided in Figure 2.

To address RQ1, we use two different state-of-the-art computer vision models $f_\theta : X \to \mathbb{R}^d$, namely, DEEPFACE (Serengil & Ozpinar, 2020) and FAIRFACE (Karkkainen & Joo, 2021), to predict the racial attribute on $x$, $x_b$, $x_g^{\sigma=2}$, and $x_g^{\sigma=3}$, and compare the class-wise performance. DEEPFACE is an ensemble method consisting of different state-of-the-art pre-trained models for facial analysis: VGG-Face (Parkhi et al., 2015), Google FaceNet (Schroff et al., 2015), OpenFace,[9] Facebook DEEPFACE (Taigman et al., 2014), DeepID (Sun et al., 2014), ArcFace (Deng et al., 2019), and Dlib.[10] FAIRFACE is the pre-trained race classification model used in the original article where the FAIRFACE data set was proposed, based on the ResNet34 model (Karkkainen & Joo, 2021). Note that both models simplify the concept of racial identity—a complex social and political construct—to a finite and distinct set of categorical labels. While we acknowledge the limitations of this approach, the use of categorical racial labels is the most widely adopted practice in machine learning research and the available data sets provide such categorical labels as ground truth to train and evaluate models (Benthall & Haynes, 2019).

*Results.* Table 1 depicts the confusion matrices obtained on race prediction. As seen in the Table, the beautified faces are more likely to be classified as White than the originals. As a consequence, the performance of both FAIRFACE and DEEPFACE *decreases* after beautification for all races except for the White race, where it *increases*. For example, before beautification, only 8.2% or 19.2% of the Latino Hispanic individuals were classified as White by FAIRFACE and DEEPFACE, respectively. After beautification, these figures increase to 34.1% (4.15x) and 35.0% (1.8x).

The use of blurred images serves as a reference to ensure that the obtained effect is not caused by an intrinsic artifact in the classification algorithms when facial features are blurred and harder to detect. We observe that the behavior on blurred images is also slightly biased toward predicting the White class, but to a much lower degree than on the

**Table 1.** Confusion matrices for the two race classification algorithms on four variations of the images (i.e., Original $x$, Beauty $x_b$, Blur1 $x_g^{\sigma=2}$, and Blur2 $x_g^{\sigma=3}$). In **green**, we highlight the highest classification percentage as White among the four variations of the images for each racial class. In **red**, we highlight the lowest class-wise classification performance.

(a) Confusion matrices for the FAIRFACE (Karkkainen & Joo, 2021) race classification algorithm. Columns and rows to be read as: White (W), Black (B), Latino Hispanic (L), East Asian (EA), Southeast Asian (SA), Indian (I), and Middle Eastern (ME). The vertical axis corresponds to the ground truth, and the horizontal to the predicted class.

| Original | W | B | L | EA | SA | I | ME | Beautified | W | B | L | EA | SA | I | ME |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W | 78.5 | 0.20 | 13.4 | 0.90 | 0.40 | 0.20 | 6.40 | | **84.4** | 0.40 | 8.10 | 1.30 | 0.90 | 0.20 | 4.70 |
| B | 0.40 | 91.5 | 4.80 | 0.00 | 0.40 | 2.90 | 0.00 | | 1.20 | 90.7 | 4.80 | 0.00 | 1.00 | 1.40 | 0.80 |
| L | 8.20 | 3.80 | 68.2 | 1.10 | 4.50 | 8.00 | 6.20 | | **34.1** | 3.30 | **43.2** | 2.40 | 3.80 | 4.50 | 8.70 |
| EA | 0.20 | 0.00 | 1.80 | 82.1 | 15.4 | 0.40 | 0.00 | | **2.70** | 0.40 | 0.40 | 80.8 | 15.4 | 0.20 | 0.00 |
| SA | 0.20 | 0.40 | 4.20 | 20.2 | 72.1 | 2.40 | 0.40 | | **2.20** | 1.50 | 4.20 | 29.7 | **58.7** | 2.40 | 1.30 |
| I | 0.90 | 3.60 | 10.0 | 0.20 | 5.20 | 76.8 | 3.20 | | **5.70** | 7.30 | 12.7 | 1.40 | 3.90 | **62.7** | 6.40 |
| ME | 16.0 | 0.70 | 15.0 | 0.20 | 1.00 | 4.50 | 62.6 | | **35.0** | 1.40 | 13.3 | 0.70 | 1.20 | 2.10 | **42.6** |
| Blur Case 1 | | | | | | | | Blur Case 2 | | | | | | | |
| W | 81.4 | 0.20 | 11.1 | 0.60 | 0.20 | 1.10 | 5.30 | | 80.6 | 0.20 | 10.4 | 1.30 | 0.20 | 0.60 | 6.70 |
| B | 0.80 | 88.8 | 5.60 | 0.00 | 0.80 | 3.70 | 0.20 | | **1.30** | **85.4** | 6.50 | 0.60 | 1.50 | 4.50 | 0.20 |
| L | 11.4 | 3.10 | 67.4 | 2.00 | 4.50 | 7.10 | 4.50 | | 14.4 | 2.50 | 62.0 | 2.50 | 5.40 | 7.90 | 5.40 |
| EA | 0.90 | 0.20 | 1.80 | 80.7 | 15.9 | 0.50 | 0.00 | | 2.10 | 0.20 | 2.30 | **79.2** | 15.7 | 0.50 | 0.00 |
| SA | 0.70 | 0.70 | 5.30 | 21.7 | 68.5 | 2.40 | 0.70 | | 1.10 | 0.50 | 6.50 | 24.3 | 64.4 | 2.50 | 0.70 |
| I | 1.40 | 3.20 | 11.0 | 0.20 | 5.30 | 77.1 | 1.80 | | 1.60 | 3.50 | 11.9 | 0.20 | 5.60 | 74.4 | 2.80 |
| ME | 20.6 | 1.00 | 17.9 | 0.50 | 1.00 | 5.00 | 54.1 | | 24.1 | 0.50 | 17.5 | 0.70 | 0.50 | 5.70 | 51.0 |

(b) Confusion matrices for the DEEPFACE (Serengil & Ozpinar, 2020) race classification algorithm. White (W), Black (B), Latino Hispanic (L), Asian (A), Indian (I), and Middle Eastern (ME). The vertical axis corresponds to the ground truth, and the horizontal to the predicted class.

| Original | W | B | L | A | I | ME | Beautified | W | B | L | A | I | ME |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W | 65.9 | 0.60 | 16.8 | 9.00 | 1.10 | 6.60 | | **72.7** | 1.10 | 10.7 | 8.50 | 1.10 | 6.00 |
| B | 1.20 | 87.4 | 2.50 | 5.80 | 2.10 | 1.00 | | 3.50 | 84.1 | 5.20 | 5.00 | 1.40 | 0.80 |
| L | 19.2 | 4.90 | 48.8 | 14.7 | 4.90 | 7.60 | | **35.0** | 7.10 | **34.5** | 10.0 | 4.50 | 8.90 |
| A | 7.90 | 3.00 | 4.10 | 82.50 | 1.20 | 1.30 | | 10.0 | 4.80 | 6.40 | 76.50 | 1.80 | 0.60 |
| I | 3.90 | 14.3 | 20.5 | 10.7 | 43.0 | 7.70 | | **10.9** | 17.0 | 23.4 | 9.50 | **31.4** | 7.70 |
| ME | 29.5 | 2.90 | 22.6 | 5.20 | 4.00 | 35.7 | | **45.0** | 4.00 | 15.5 | 3.80 | 4.50 | **27.1** |
| Blur Case 1 | | | | | | | Blur Case 2 | | | | | | |
| W | 67.2 | 0.40 | 13.2 | 10.4 | 1.30 | 7.50 | | 61.4 | 0.60 | 11.1 | 14.1 | 1.70 | 11.1 |
| B | **4.50** | 83.5 | 2.90 | 6.60 | 2.30 | 0.20 | | 3.30 | **80.6** | 3.50 | 9.90 | 1.90 | 0.80 |
| L | 24.7 | 4.20 | 40.8 | 16.3 | 3.80 | 10.2 | | 25.6 | 3.60 | 37.6 | 18.3 | 3.30 | 11.6 |
| A | 14.1 | 3.50 | 3.40 | **76.1** | 1.30 | 1.60 | | **14.4** | 2.40 | 2.50 | 77.2 | 2.00 | 1.40 |
| I | 9.10 | 14.3 | 16.8 | 12.5 | 39.1 | 8.20 | | 7.70 | 14.3 | 15.5 | 36.1 | 15.7 | 10.7 |
| ME | 32.9 | 2.40 | 16.0 | 8.30 | 4.30 | 36.2 | | 36.0 | 1.40 | 11.7 | 7.90 | 4.00 | 39.0 |

beautified case. Interestingly, the Black and (East) Asian classes are the least impacted in terms of classification performance after beautification. In this case, the blurred images yield the worst classification accuracy for both in FAIRFACE and DEEPFACE. The decrease in performance obtained on beautified faces and the increase of their classification as White suggests a bias in the beautification process toward Eurocentric beauty standards that correspond to the White class. The loss in performance is particularly prominent for the Indian, Middle Eastern, Southeast Asian, and Latino Hispanic races: in the case of the Indian class, there is a loss in accuracy of 14.1 points or 18.3% (FAIRFACE) and 11.6 points or 27.0% (DEEPFACE); for Middle Eastern, 20 points or 31.9% (FAIRFACE) and 8.6 points or 24.0% (DEEPFACE); for Southeast Asian, 13.6 points or 18.5% (only available for DEEPFACE); and for Latino Hispanic, 25 points or 36.6% (FAIRFACE) and 14.3 points or 29.3% (DEEPFACE).

Furthermore, a comparison between the per gender race classification performance on the original $x$ and beautified $x_b$ images is depicted in Figure 3, where the performance on

female faces is shown with orange bars and the performance on male faces is depicted with purple bars. Figure 3 shows two different performance metrics.

The dark-colored bars correspond to the accuracy loss/gain (in percentage points) in classifying the race of the images after beautification, such that, a negative/positive value corresponds to a loss/gain in accuracy, respectively. The only race where the prediction performance consistently increases after the application of the beauty filters is the White race and hence bars show positive values. For the rest of the races, the race classification accuracy significantly decreases (negative values in the bars) after beautification, with the exception of East Asian and Black males, where the performance of the FAIRFACE model slightly improves after beautification.

The light-colored bars depict the percentage of images in each race category that are classified as White after the application of beauty filters but *were not classified as White before beautification*. This percentage is notably large in the case of the Latino Hispanic and Middle Eastern races, but it is present on all races, for both genders and with both race classification models. While the DEEPFACE model seems to be more sensitive to beautification than the FAIRFACE model, both methods are severely impacted by the beauty filters.

Regarding gender, we observe that both the images of male and female faces are more likely to be classified as White after beautification. Yet, there are some gender differences. We perform t-tests between the models' loss in performance for male and female faces after beautification and conclude that no gender bias is present in the case of the DEEPFACE model, that is, the loss in performance after beautification is similar for male and female faces across all racial categories. However, in the case of the FAIRFACE model, the difference in classification accuracy between male and female faces after beautification (dark-colored bars) is statistically significant in the case of the Southeast Asian ($p$-value $<.001$) and Indian ($p$-value $<.001$) races. In both cases, the loss in accuracy is larger for the male faces. Regarding the increase (in percentage points) in the number of individuals classified as White after beautification (light-colored bars), we observe a statistically significant difference only in the case of the Latino Hispanic ($p$-value $<.001$) class. In this case, female faces are more negatively affected than their male counterparts.

## RQ2: How Do Beauty Filters Embed Eurocentric Beauty Canons?

To address this RQ, we leverage attribution methods (Abhishek & Kamath, 2022), a popular tool within the explainable AI field (Fel et al., 2022). Attribution methods in computer vision are used to understand the contribution of different areas of an image to a specific output in the prediction of a model or algorithm. These methods are used to improve the interpretability and explainability of deep learning-based computer vision models (Colin et al., 2022).

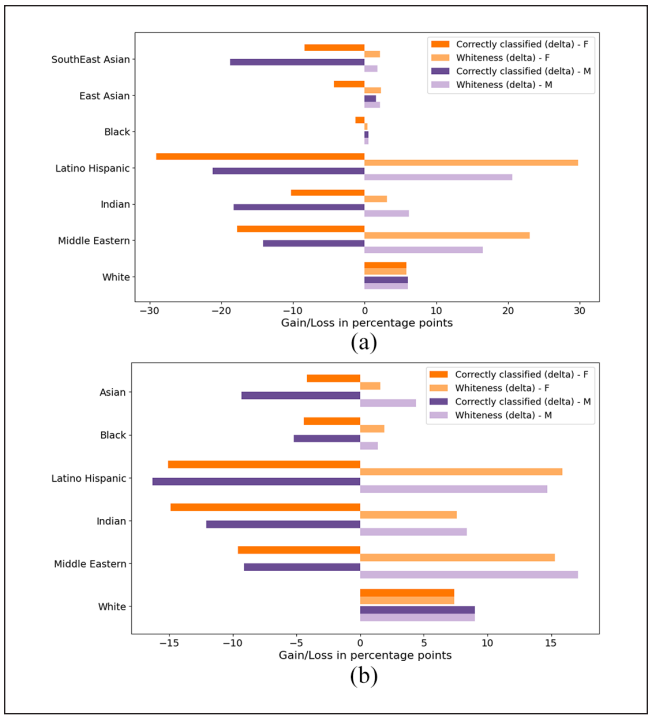


**Figure 3.** For every race and gender (F and M), the dark-colored bars represent the change in accuracy after beautification, while the light-colored bars depict the difference in the percentage of images that are classified as White after beautification. Note that in the case of DEEPFACE, the "East Asian" and "Southeast Asian" classes are labeled "Asian," as per the training process of the model. (a) Differences in classification performance after beautification for FAIRFACE (Karkkainen & Joo, 2021) on images of female (F, orange bars) and male (M, purple bars) individuals. (b) Differences in classification performance after beautification for DEEPFACE (Serengil & Ozpinar, 2020) on images of female (F, orange bars) and male (M, purple bars) individuals.

Attribution methods may be categorized as gradient-based (Simonyan et al., 2013; Sundararajan et al., 2017) or sensitivity-based (Fel et al., 2021; Zeiler & Fergus, 2014). Sensitivity-based attribution methods assign a numeric score to each pixel of the image according to how important it is for the classification by probing the model with $m$ occluded versions of the input and analyzing how each of them impacts the output score of the model. We use a sensitivity-based attribution method to shed light on the areas in the image that are the most informative to decide the race of the faces before and after beautification. By comparing these areas, we aim to pinpoint the factors that contribute to the decrease in performance of the race classification algorithms and the erroneous classification of non-White faces as White.

*Problem Formulation and Setup.* We define as $C \subset X$ the set of images for which $x$ and $x_b$ are classified correctly and as $F \subset X$ the set of images for which (1) $x$ is classified correctly as non-White but $x_b$ is classified incorrectly as White or (2) $x$ is classified incorrectly as non-White and $x_b$ is classified correctly as White.

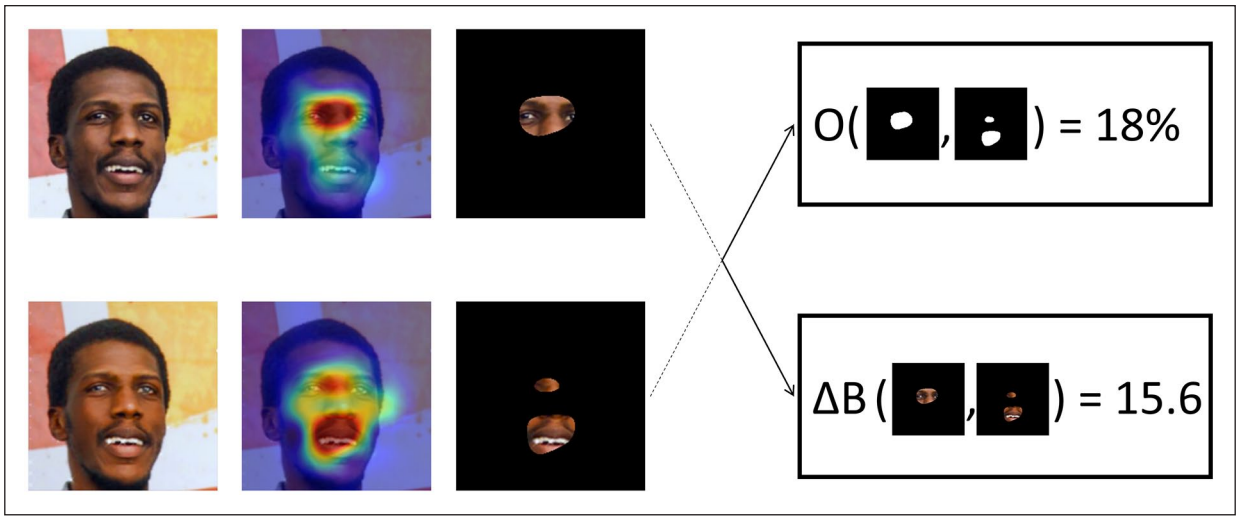


**Figure 4.** Explainability pipeline used to address RQ2. From left to right: input images ($x$ at the top and $x_b$ at the bottom), their respective heatmaps ($\psi(x)$ and $\psi(x_b)$) before thresholding, their masked images ($\tilde{x}$ and $\tilde{x}_b$), and an illustration of the *Overlap* and $\Delta B$ measures.

To gain an insight behind the reason for the classifications in both $F$ and $C$, we use a state-of-the-art Sobol-based sensitivity analysis attribution method (Fel et al., 2021; Sobol, 1993) to compute a heatmap $\psi(x)$ with the contribution $\psi(x^i)$ of each pixel $x^i$ of an input image $x$ to a given output of the model $f_\theta(x)$. The resulting heatmap $\psi(x)$ highlights the parts of the image that are the most important for the decision of the model.

This attribution method has been found to be effective in identifying a small number of important pixels that drive the prediction of the model. Typically, 5%–10% of the pixels account for more than 80% of the accuracy of the model (Petsiuk et al., 2018). Thus, to ease the comparison, we threshold $\psi$ and $\psi(x_b)$ to keep the 5% of the pixels contributing the most to the classification and put 0 everywhere else in the heatmap, creating the binary masks $\tilde{\psi}(x)$ and $\tilde{\psi}(x_b)$, *e.g.*, $\tilde{\psi}(x^i) = 1$ if $\psi(x^i) > 0$ else $\tilde{\psi}(x^i) = 0$. As a result, we obtain a binary mask where only the most relevant pixels for the race classification are marked as 1 and the remaining pixels are set to 0. We apply these binary masks on $x$ and $x_b$ to create the masked images $\tilde{x}$ and $\tilde{x}_b$, which are the original and beautified images but with non-zero values only for the pixels highly contributing to the classification.

Our goal is to determine whether the changes in the facial features caused by the beautification process lead to the algorithms paying attention to *different parts* of the face on the beautified images when compared to the original images, which might explain the classification errors. Therefore, to address RQ2, we postulate *two* hypotheses that we empirically evaluate by means of quantitative measurements (see Figure 4 for an illustration of the pipeline).

*$H_1$: When $x_b \in F$ is misclassified, the race detection algorithms focus on different parts of the images than when classifying $x$.*

The reason for this change of focus on $x_b$ might be due to the fact that beauty filters modify the original facial features, forcing the face processing algorithms to shift their attention to other facial elements in the beautified version of the images. To quantitatively evaluate this hypothesis, we compute the *overlap*, $O$, between the original $(\tilde{\psi}(x))$ and beautified $(\tilde{\psi}(x_b))$ heatmaps, defined as the number of pixels that are set to 1 in the heatmaps of both the original and the beautified images, normalized by the total number of non-zero pixels in the original heatmap. Formally, the overlap is thus given by the following expression:

$$O_{x,x_b} = \frac{\sum_i min(\tilde{\psi}(x^i), \tilde{\psi}(x_b^i))}{k} \quad (1)$$

with $k$ the number of non-zero pixels[11] in $\tilde{\psi}(x)$.

Our second hypothesis is formulated as follows:

*$H_2$: When the race detection algorithms misclassify $x_b \in F$ as White, they pay attention to parts of the image that are brighter than in the original image.*

The reasoning behind this hypothesis is that, in addition to the change of focus, the brightening of the faces that occurs after beautification might contribute to the misclassification. The quantitative measure that we propose to evaluate this hypothesis is $\Delta B$, defined as the normalized difference in brightness $B$ between the pixels in the masked original $(\tilde{x})$ and beautified $((\tilde{x}_b))$ images:

$$\Delta B_{x,x_b} = \frac{\sum_i B(\tilde{x}_b^{\,i}) - B(\tilde{x}^i)}{k}. \quad (2)$$

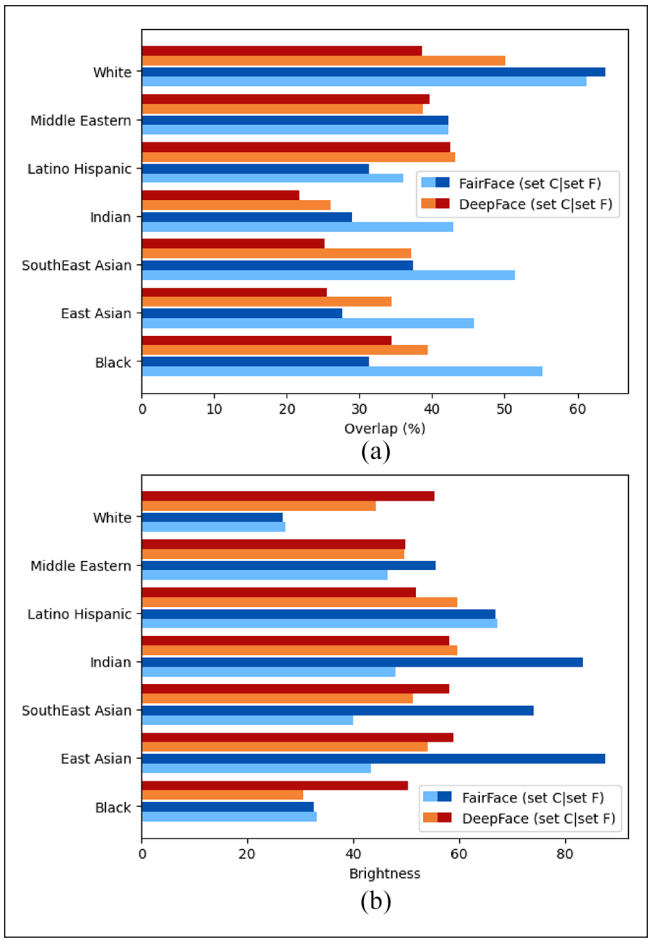


**Figure 5.** Overlap and $\Delta B$ results for the FAIRFACE and DEEPFACE algorithms. For every race, the cyan and orange bars depict the results for the images correctly classified before and after beautification (set $C$), and the blue and red bars show the results for the images that after beautification either get incorrectly classified as White or correctly classified as White when they were not before (set $F$). (a) Overlap for the FAIRFACE and DEEPFACE algorithms. (b) $\Delta B$ for the FAIRFACE and DEEPFACE algorithms.

Note that we compute the brightness $B$ by converting the image in RGB (red, green, and blue) to the HSV color space (also called HS**B** for hue, saturation, **brightness**).

*Results.* Figure 5 summarizes the per-race overlap (top graph) and $\Delta B$ (bottom graph) measurements for both the FAIRFACE (cyan-blue bars) and DEEPFACE (orange-red bars) algorithms.[12]

As seen in Figure 5a, the overlap in the heatmaps between the original image and its beautified version is smaller in the images that are misclassified (set $F$) when compared to the overlap in the images that are correctly classified (set $C$). The average overlap for all races is 47.7% in $C$ versus 42% in $F$ for FAIRFACE and 38.3% in $C$ versus 35.7% in $F$ for DEEPFACE. A $t$-test reveals that this difference is significant for FAIRFACE : $t(894) = 2.9, p = .004$ , but not for DEEPFACE: $t(792) = 1.35, p = .18$ . However, this difference is significant for some races even in the case of DEEPFACE, such as White ($t(122) = 2.65, p = .009$) and East Asian ($t(96) = 2.65, p = .01$). These results support our hypothesis $H_1$: as a result of the beautification process, the race detection algorithms—and especially FAIRFACE—focus on *different facial parts* than those used when analyzing the original image $x$.

Moreover, we observe in Figure 5b that the overall $\Delta B$ of the misclassified images is larger than that of the correctly classified images. The average $\Delta B$ for all races is 43.6 in $C$ versus 55 in $F$ for FAIRFACE and 52.4 in $C$ versus 54.2 in $F$ for DEEPFACE. Here again, a $t$-test reveals that the difference is significant for FAIRFACE: $t(894) = -4.2, p < .001$, but not for DEEPFACE: $t(792) = -0.64, p = .52$. In the case of DEEPFACE, this delta is significant for some races, such as Black ($t(22) = -2.78, p = .01$) and White ($t(122) = -1.83, p = .07$ ).

In other words, the parts of the images analyzed by the race classification algorithms to wrongly determine the race of the beautified faces (set $F$)—and most likely classify them as White according to the results reported in Figure 3—tend to be brighter than the parts used in the correctly classified images (set $C$), especially in the case of the FAIRFACE algorithm.

Interestingly, in Figure 5, we observe that for the two races with the largest misclassification rates (Latino Hispanic and Middle Eastern), these differences are less notable. For example, the loss in classification performance of DEEPFACE on the Middle Eastern class is of 24% as per our previous analysis. In this case, the overlap (39% vs. 39.61%) and $\Delta B$ (49.78% vs. 50%) are similar in the misclassified ($F$) than in the correctly classified ($C$) images. This result suggests that the changes made by beauty filters encompass complex modifications to the facial features and skin texture or color, beyond a simple brightening of the face. Figure 6 highlights two examples from the set $F$ where the FAIRFACE algorithm focuses on the *same* facial features both in $x$ and $x_b$, and the focus area is not brighter, yet $x_b$ is misclassified as White, whereas $x$ is correctly classified. This finding supports the hypothesis that the changes applied by the beauty filters to the facial features (e.g., changes in the eyes' shape and color, the mouth, and the nose) also play a role in explaining the racial bias.

## Discussion

The aim of this article is to spur a discussion toward a more equitable *Beautyverse* and, by extension, a healthier social media environment. Our work should be interpreted from this perspective, acknowledging that the pervasiveness of these filters has been found to impact their users (e.g., insecurities and body dissatisfaction), as previously discussed.

We have leveraged state-of-the-art computer vision techniques to study a complex social phenomenon, namely, the definition of beauty canons as reflected by beauty filters. Our work illustrates the need for all the stakeholders that contribute to the development of pervasive technologies on social

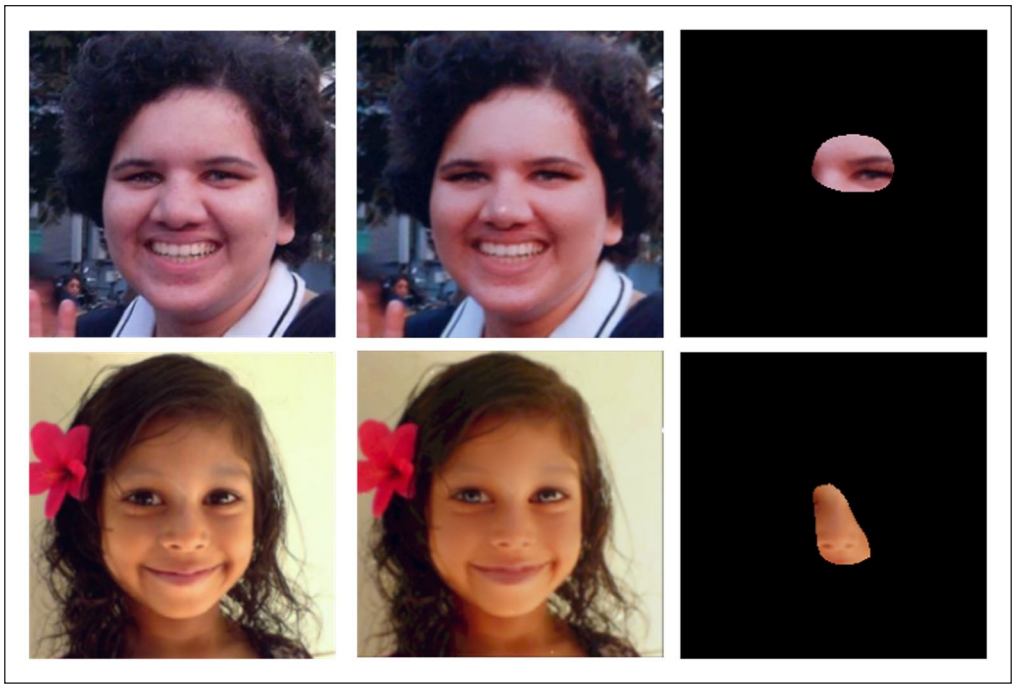

**Figure 6.** Examples of individuals that are misclassified as White by FAIRFACE after but not before beautification. In this case, the algorithm focuses on the eyes (top) and nose (bottom) regions. From left to right: image $x$, beautified image $x_b$, and shared pixels (overlap) driving the prediction both in $x$ (Latino Hispanic) and $x_b$ (White).

media—from scientists to developers, users and social media, and advertisement companies—to understand and embrace race, gender, body type, culture, beliefs, language and functional *diversity* (Ogolla & Gupta, 2018). In the case of social media platforms, they could review their community standards and guidelines to ensure they address the responsible use of beauty filters and other image-enhancing tools. This might include restrictions on filters that promote racial biases or misrepresent individuals' natural appearances.

From our experiments, we draw several insights and implications regarding the *Beautyverse*.

### *Beauty Filters Embed a Racial Bias*

We find that beauty filters transform the faces to conform with Eurocentric (*white*) beauty canons as perceived by state-of-the-art race classification algorithms (RQ1). Racial biases embedded in beauty filters had been previously hypothesized by researchers in humanity-related fields and by social media practitioners or users from marginalized communities. However, they had not been empirically validated to date until this study.

The fact that beauty filters reinforce and promote *white* beauty standards perpetuates the notion that Western features are the epitome of attractiveness. This finding suggests that beauty filters contribute to the perpetuation of racial stereotypes, reinforcing existing biases, contributing to the subconscious association of certain non-White racial traits with negative attributes or less beauty, and potentially further marginalizing and devaluing individuals with diverse racial backgrounds and features.

### *The Racial Bias Entails Changes Beyond Skin Whitening*

The reasons why race classification algorithms have a tendency to classify beautified faces—irrespective of their race—as White are complex. From our explainability experiment (RQ2), both a brightening of the skin color and changes in the facial features play a role in confusing the algorithms.

### *State-of-the-Art Face Processing Algorithms Are Sensitive to Beauty Filters*

According to our work, race classification algorithms are not robust to popular beauty filters from social media. Interestingly, while the FAIRFACE model exhibits the best classification performance on the original data sets, it is more impacted by the beauty filters than the DEEPFACE model, both in terms of absolute performance and gender bias. As we increasingly rely on face processing systems to automate or support human decisions—particularly in consequential areas, such as hiring, dating or college admissions—this fragility should be taken into account, especially given the ubiquity of beauty filters.

However, we do not intend this evidence to necessarily serve as an encouragement to develop more robust race classification algorithms. These models—along with other face

processing algorithms, including face recognition systems—pose significant legal and ethical challenges (Bu, 2021), which need to be taken into account before deciding to work on their development, deployment or technical improvement. Classifying humans through their visual characteristics may lead to the misuse of technology for oppression purposes, as we have witnessed in human history (Scheuerman et al., 2020). Should our readers decide to pursue such a research line, we strongly recommend performing a prior rigorous study of potentially unintended applications and the broad societal impact that these tools might have.

### *The Social Implications of This Phenomenon Should Be Further Studied*

The beauty filters considered in this study are designed by social media users. Therefore, our experiments may be seen as empirical evidence of the social influence of Eurocentric beauty standards in the definition of these filters and the choices that users make when designing them. A failure to acknowledge the existence of this systematic racial bias in our society will ultimately prevent achieving a more diverse, inclusive and equitable *Beautyverse*.

Given the prevalence of beauty filters on social media platforms, their biases contribute to a skewed perception of attractiveness and desirability, leading to implications for social interactions, dating apps, and even job opportunities in professions that heavily rely on virtual presence. Our work indeed emerges from important concerns for non-White individuals, and especially women. Not only are women worldwide subject to the pressure of a male-gazed (Mulvey, 1975) society that conceives them as objects of sexual desire that should satisfy the *pleasure in looking*, but they are also, once again in human history, subject to the idea that looking *beautiful* also means being *white*. In addition, recent advances in generative AI algorithms to automatically create images and videos could exacerbate the dangerous effects of representational biases for women and racial minorities even further (Luccioni et al., 2024). We leave to future work the analysis of potential racial biases in such algorithms.

### *Beauty Filters as a Colonial Symbol*

The popularity of beauty filters and the worldwide diffusion of the standardized—and biased—canons of beauty represented by these filters may be interpreted as a consequence of globalization, and globalization can be considered as a modern form of colonization (Banerjee & Linstead, 2001) that some authors define as "electronic colonization" (Zembylas & Vrasidas, 2005). Being a Western-driven process, it presents the Western world as attractive and beneficial, while appropriating, homogenizing, and standardizing the Global South (Akinro & Mbunyuza-Memani, 2019).

The research presented in this article contributes to a more nuanced, empirical and data-driven perspective on the standardization of beauty ideals that are defined, promoted and reinforced by this modern colonization phenomenon. Thus, a decolonization perspective regarding the use of beauty filters on social media is needed. Such a perspective underscores the need to critically examine and challenge the perpetuation of Eurocentric beauty standards in the digital space. By acknowledging historical colonial legacies, promoting cultural appreciation over appropriation, advocating for inclusive beauty standards, and empowering diverse communities to reclaim their narratives, our research aims to foster a more equitable, diverse, and respectful digital beauty culture that honors and celebrates the richness of global canons of beauty.

### *Beauty Filters as Equalizers*

At the same time, the goal of our research is not to denounce beauty filters per se, but rather investigate the biases and unintended negative consequences that these filters may have. Beauty filters could also be seen as tools for the democratization of beauty given the existence of the *attractiveness halo effect* (Dion, 1972; Gulati et al., 2022; Talamas et al., 2016). According to this cognitive bias, people that are considered to be attractive are also perceived as having a range of positive attributes, including higher morality and trustworthiness, better mental health, and superior intelligence. From this perspective, beauty filters could contribute to removing (or alleviating) physical appearance as a decisive factor in certain sensitive contexts (e.g., hiring processes or judicial sentences) and hence contribute to *fairer* decisions. This is a research direction that we plan to study in future work.

## Limitations

To the best of our knowledge, our work is the first extensive effort to quantitatively study racial biases in beauty filters. Our aim is to bring the attention toward this topic not only in the scientific community, but also among practitioners, developers, and industrial stakeholders that can effectively make a change in the status quo. Our aspiration is to contribute with our research to an ethical development of AI that would yield positive societal impact. However, our approach is not exempt from limitations.

First, in our data sets, users of social media platforms typically follow specific communication paradigms (e.g., adopt certain poses for selfies) (Qiu et al., 2015; Tifentale & Manovich, 2015) that might not be fully reflected in the data sets used in this research. As previously explained, to mitigate this issue, we selected a subset of the images in FAIRFACE to be as similar as possible to the kinds of images shared on social media. Moreover, the diversity by design in FAIRFACE might make this data set demographically more heterogeneous than the social media experience of most users. Despite this limitation, we believe that the findings of our experiments would apply to other face data sets. Working with a publicly available data set, such as FAIRFACE, is a choice driven by several factors. Directly scraping social media

platforms to collect face images is neither ethically nor legally acceptable, as this would entail processing faces of users (i.e., a sensitive attribute), without their explicit consent. In addition, our analyses require both non-beautified and beautified image pairs of the same individual, which might be difficult to obtain from social media data.

A second limitation stems from the fact that most of the algorithms used in this article are complex deep learning-based systems that combine different modules with opaque inner workings (e.g., DEEPFACE uses a pre-trained ensemble). The complexity of these systems may impact the results, as the different modules could be affected by the beauty filters differently, possibly leading to unexpected outcomes. To mitigate this limitation, we use two different methods in our experiments and obtain consistent results.

Third, we recognize the limitation of using categorical racial labels, which is a highly debated topic and an open RQ. This non-ideal choice was due to technical reasons. Given that the machine learning community is still not critical enough in its engagement with the socially constructed meaning of races and their political derivations (Benthall & Haynes, 2019), existing race classification algorithms model race as a categorical label (Guo et al., 2016; Parkhi et al., 2015). An interesting direction of future work in this area would be to develop systems that are able to move beyond categorical racial labels.

## Conclusion

We have investigated the racial biases embedded in today's social media beauty filters, contextualizing our work from a historical perspective. We have applied race classification algorithms to over 3,000 images from the FAIRFACE and FAIRBEAUTY data sets, corroborating the hypothesis that beauty filters transform their users' faces, so that, they conform with Eurocentric beauty standards, with varying impact on different racial groups. The most impacted racial groups are Latino Hispanic and Middle Eastern, with a drop in performance of up to 25 points (from 68.2% to 43.2% accuracy) and 20 points (from 62.6% to 42.6% accuracy), respectively, and a significant increase in their classification probability as White individuals. Furthermore, we have leveraged a state-of-the-art explainability framework to analyze the facial changes embedded in the beauty filters that contribute to the misclassification. We conclude that such a misclassification is not simply due to a brightening of the skin color, but also to a modification of the characteristic facial features of the different racial groups.

We hope that our research will spur further interdisciplinary work to build a more inclusive, equitable and diverse social media environment.

### Acknowledgements

The authors deeply thank all the colleagues that have read this work and provided insightful feedback and suggestions, which surely helped us improve the quality of our research. They also thank Núria Camps for her substantial help in data set pre-processing.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: P.R., J.C., and N.O. are supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación). PR and JC are also supported by a grant by the Bank Sabadell Foundation. JC and NO are also supported by Intel Corporation.

### ORCID iD

Piera Riccio https://orcid.org/0000-0001-8602-8271

### Notes

1. Fotor, accessed on January 13, 2023, https://www.fotor.com/es/.
2. BeautyPlus, accessed on January 13, 2023, https://play.google.com/store/apps/details?id=com.commsource.beautyplus.
3. Portrait Pro Studio Max, accessed on January 13, 2023, https://www.anthropics.com/portraitpro/
4. "Shadeism" is the dark side of discrimination we ignore, Global News, accessed on January 13, 2023, https://globalnews.ca/news/5302019/shadeism-colourism-racism-canada/
5. BeautyCam, accessed on May 3, 2023, https://play.google.com/store/apps/details?id=com.meitu.meiyancamerahl=itgl=USpli=1.
6. Colorism vs. Racism: What's the Difference?, accessed on April 25, 2023, https://www.rd.com/article/colorism/
7. Medical Korea, accessed on December 26, 2023, https://english.visitkorea.or.kr/svc/contents/contentsView.do?menuSn=612vcontsId=139792.
8. ELLIS Alicante GitHub Repository: https://github.com/ellisalicante/racialbias-beautyfilters.
9. OpenFace, accessed on April 26, 2023, https://cmusatyalab.github.io/openface/
10. Face Recognition with Dlib in Python, accessed on April 26, 2023, https://sefiks.com/2020/07/11/face-recognition-with-dlib-in-python/.
11. $k(x) = H(x) \times W(x) \times n$, with H and W, respectively, the height and the width of the image $x$, and $n = 0.05$ or 5% as previously explained.
12. Note that in the case of DEEPFACE, the East Asian and Southeast Asian classes are merged into Asian.

### References

Abhishek, K., & Kamath, D. (2022). Attribution-based XAI methods in computer vision: A review. *arXiv preprint arXiv:2211.14736*.

Adawe, A., & Oberg, C. (2013). Skin-lightening practices and mercury exposure in the Somali community. *Minnesota Medicine*, *96*(7), 48–49.

Akinro, N., & Mbunyuza-Memani, L. (2019). Black is not beautiful: Persistent messages and the globalization of "white" beauty in African women's magazines. *Journal of International and Intercultural Communication*, *12*(4), 308–324.

Alm, S., & Låftman, S. B. (2018). The gendered mirror on the wall: Satisfaction with physical appearance and its relationship to global self-esteem and psychosomatic complaints among adolescent boys and girls. *Young*, *26*(5), 525–541.

Ash, J., Anderson, B., Gordon, R., & Langley, P. (2018). Digital interface design and power: Friction, threshold, transition. *Environment and Planning D: Society and Space*, *36*(6), 1136–1153.

Bakker, M. (2022). *# nofilter how beauty filters affect the internalization of beauty ideals* [Master's thesis, Utrecht University].

Banerjee, S. B., & Linstead, S. (2001). Globalization, multiculturalism and other fictions: Colonialism for the new millennium? *Organization*, *8*(4), 683–722.

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency, FAT* '19* (pp. 289–298). Association for Computing Machinery.

Bharati, A., Vatsa, M., Singh, R., Bowyer, K. W., & Tong, X. (2017). Demography-based facial retouching detection using subclass supervised sparse autoencoder. In *IEEE international joint conference on biometrics (IJCB)* (pp. 474–482). IEEE.

Boyd, D. (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), *The John D. and Catherine T. MacArthur Foundation series on digital media and learning. Youth, identity, and digital media* (pp. 2007–2016). The MIT Press.

Bu, Q. (2021). The global governance on automated facial recognition (AFR): Ethical and legal opportunities and privacy challenges. *International Cybersecurity Law Review*, *2*(1), 113–145.

Chin Evans, P., & McConnell, A. R. (2003). Do racial minorities respond in the same way to mainstream beauty standards? Social comparison processes in Asian, black, and white women. *Self and Identity*, *2*(2), 153–167.

Cockburn, C. (1983). *Brothers: Male dominance and technological change*. Sage.

Colin, J., Fel, T., Cadène, R., & Serre, T. (2022). What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, *35*, 2832–2845.

Cristel, R. T., Dayan, S. H., Akinosun, M., & Russell, P. T. (2021). Evaluation of selfies and filtered selfies and effects on first impressions. *Aesthetic Surgery Journal*, *41*(1), 122–130.

Dantcheva, A., Bremond, F., & Bilinski, P. (2018). Show me your face and i will tell you your height, weight and body mass index. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 3555–3560). IEEE.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690–4699). IEEE.

Dimitrov, D., & Kroumpouzos, G. (2023). Beauty perception: A historical and contemporary review. *Clinics in Dermatology*, *41*(1), 33–40.

Dion, K. K. (1972). Physical attractiveness and evaluation of childrens transgressions. *Journal of Personality and Social Psychology*, *24*(2), 207–213.

Dyer, R. (2017). *White*. Routledge.

Eshiet, J. (2020). *"Real me versus social media me": Filters, Snapchat dysmorphia, and beauty perceptions among young women* [Master's thesis, California State University, San Bernardino].

Fanon, F. (2008). *Black skin, white masks*. Grove press.

Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., & Serre, T. (2021). Look at the variance! Efficient black-box explanations with sobol-based sensitivity analysis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 26005–26014). Curran Associates, Inc.

Fel, T., Hervier, L., Vigouroux, D., Poche, A., Plakoo, J., Cadene, R., Chalvidal, M., Colin, J., Boissin, T., Bethune, L., Picard, A., Nicodeme, C., Gardes, L., Flandin, G., & Serre, T. (2022). Xplique: A deep learning explainability toolbox. *arXiv preprint arXiv:2206.04394*.

Felisberti, F. M., & Musholt, K. (2014). Self-face perception: Individual differences and discrepancies associated with mental self-face representation, attractiveness and self-esteem. *Psychology & Neuroscience*, *7*(2), 65–72.

Figueroa, M. G. M. (2021). Picking your battles: Beauty, complacency, and the other life of racism. In M. L. Craig (Ed.), *The Routledge companion to beauty politics* (pp. 49–59). Routledge.

Fischer-Tiné, H. (2009). *Low and licentious Europeans: Race, class and "white subalternity" in colonial India* (Vol. 30). Orient Blackswan.

Fox, J., & Bailenson, J. N. (2009). Virtual self-modeling: The effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychology*, *12*(1), 1–25.

Fribourg, R., Peillard, E., & McDonnell, R. (2021). Mirror, mirror on my phone: Investigating dimensions of self-face perception induced by augmented reality filters. In *IEEE international symposium on mixed and augmented reality (ISMAR)* (pp. 470–478). IEEE.

Frieze, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management. *Journal of Applied Social Psychology*, *21*(13), 1039–1057.

Grossman, M. (2017). *Study of social media users: The relationship between online deception, machiavellian personality, self-esteem, and social desirability* [Doctoral thesis, California School of Professional Psychology].

Gruber, E., Kalkbrenner, M. T., & Hitter, T. L. (2022). A complex conceptualization of beauty in Latinx women: A mixed methods study. *Body Image*, *41*, 432–442.

Gulati, A., Lozano, M. A., Lepri, B., & Oliver, N. (2022). Biased: Bringing irrationality into automated system design. *arXiv preprint arXiv:2210.01122*.

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

Hedman, P., Skepetzis, V., Hernandez-Diaz, K., Bigun, J., & Alonso-Fernandez, F. (2022). LFW-beautified: A dataset of face images with beautification and augmented reality filters. *arXiv:2203.06082*.

Hong, S., Jahng, M. R., Lee, N., & Wise, K. R. (2020). Do you filter who you are? Excessive self-presentation, social cues, and user evaluations of Instagram selfies. *Computers in Human Behavior*, *104*, 106159.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008, October). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in "real-life" images: Detection, alignment, and recognition*. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France.

Jagota, V. (2016). Why do all the Snapchat filters try to make you look white? https://www.complex.com/life/a/vrinda-jagota/implicit-racial-bias-tech

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 1548–1558). IEEE.

Kilomba, G. (2021). *Plantation memories: Episodes of everyday racism*. Between the Lines.

Kim, T. (2003). Neo-Confucian body techniques: Women's bodies in Korea's consumer society. *Body & Society*, *9*(2), 97–113.

Kullrich, N. (2022). In this country, beauty is defined by fairness of skin. In J. B. Metzler (Ed.), *Skin colour politics: Whiteness and beauty in India* (pp. 1–50). Springer.

Lamp, S. J., Cugle, A., Silverman, A. L., Thomas, M. T., Liss, M., & Erchull, M. J. (2019). Picture perfect: The relationship between selfie behaviors, self-objectification, and depressive symptoms. *Sex Roles*, *81*(11), 704–712.

Li, A. K. (2019). Papi Jiang and microcelebrity in China: A multilevel analysis. *International Journal of Communication*, *13*, 19.

Li, S. (2020). *The problems with Instagram's most popular beauty filters, from augmentation to Eurocentrism*. https://www.nylon.com/beauty/instagrams-beauty-filters-perpetuate-the-industrys-ongoing-racism

Lloréns, H. (2013). Latina bodies in the era of elective aesthetic surgery. *Latino Studies*, *11*, 547–569.

Luccioni, S., Akiki, C., Mitchell, M., & Jernite, Y. (2024). Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, *36*, 56338–56351.

Maloney, D., & Robb, A. (2019). An initial investigation into stereotypical influences on implicit racial bias and embodied avatars. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)* (pp. 1074–1075). IEEE.

McLean, S. A., Rodgers, R. F., Slater, A., Jarman, H. K., Gordon, C. S., & Paxton, S. J. (2022). Clinically significant body dissatisfaction: Prevalence and association with depressive symptoms in adolescent boys and girls. *European Child & Adolescent Psychiatry*, *31*(12), 1921–1932.

Mirabet Herranz, N., Galdi, C., & Dugelay, J.-L. (2022). Impact of digital face beautification in biometrics. In *2022 10th European workshop on visual information processing (EUVIP)* (pp. 1–6). IEEE.

Mire, A. (2001). Skin-bleaching: Poison, beauty, power, and the politics of the colour line. *Resources for Feminist Research*, *28*(3–4), 13–41.

Morrow, P. C., McElroy, J. C., Stamper, B. G., & Wilson, M. A. (1990). The effects of physical attractiveness and other demographic characteristics on promotion decisions. *Journal of Management*, *16*(4), 723–736.

Mulaudzi, S. (2017). Let's be honest: Snapchat filters are a little racist. https://www.huffingtonpost.co.uk/entry/snapchat-filters-are-harming-black-womens-self-image_uk_5c7e945ce4b078abc6c0f1e7

Mulvey, L. (1975). Visual pleasure and narrative cinema. *Screen*, *16*(3), 6–18.

Myers, T. A., & Crowther, J. H. (2009). Social comparison as a predictor of body dissatisfaction: A meta-analytic review. *Journal of Abnormal Psychology*, *118*(4), 683–698.

Ogolla, S., & Gupta, A. (2018). Inclusive design. Methods to ensure a high degree of participation in artificial intelligence (AI) systems. In *University of Oxford connected life 2018–conference proceedings* (Vol. 12, pp. 23–34). Oxford Internet Institute.

Othman, S., Lyons, T., Cohn, J. E., Shokri, T., & Bloom, J. D. (2021). The influence of photo editing applications on patients seeking facial plastic surgery services. *Aesthetic Surgery Journal*, *41*(3), NP101–NP110.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC 2015-proceedings of the British machine vision conference 2015* (pp. 1–12). British Machine Vision Association.

Peng, A. Y. (2021). A techno-feminist analysis of beauty app development in China's high-tech industry. *Journal of Gender Studies*, *30*(5), 596–608.

Perrotta, G. (2020). The concept of altered perception in "body dysmorphic disorder": The subtle border between the abuse of selfies in social networks and cosmetic surgery, between socially accepted dysfunctionality and the pathological condition. *Journal of Neurology, Neurological Science and Disorders*, *6*(1), 001–007.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. In *British machine vision conference (BMVC)* (p. 17). British Machine Vision Association (BMVA).

Pivnick, L. K., Gordon, R. A., & Crosnoe, R. (2022). The developmental significance of looks from middle childhood to early adolescence. *Journal of Research on Adolescence*, *32*(3), 1125–1139.

Qiu, L., Lu, J., Yang, S., Qu, W., & Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior*, *52*, 443–449.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226.

Riccio, P., & Oliver, N. (2023). Racial bias in the beautyverse: Evaluation of augmented-reality beauty filters. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer vision—ECCV 2022 workshops* (pp. 714–721). Springer Nature.

Riccio, P., Psomas, B., Galati, F., Escolano, F., Hofmann, T., & Oliver, N. (2022). Openfilter: A framework to democratize research access to social media AR filters. *Advances in Neural Information Processing Systems*, *35*, 12491–12503.

Richards, D., Caldwell, P. H., & Go, H. (2015). Impact of social media on the health of children and young people. *Journal of Paediatrics and Child Health*, *51*(12), 1152–1157.

Samizadeh, S. (2022). Beauty standards in Asia. In *Non-surgical rejuvenation of Asian faces* (pp. 21–32).

Sartwell, C. (2012). Beauty. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/beauty/

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human Computer Interaction*, *4*(CSCW1), 1–35.

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

Serengil, S. I., & Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In Innovations in intelligent systems and applications conference (ASYU) (pp. 1–5). IEEE.

Shein, E. (2021). Filtering for beauty. *Communications of the ACM*, *64*(11), 17–19.

Siddiqui, A. (2021). Social media and its role in amplifying a certain idea of beauty. *Infotheca—Journal for Digital Humanities*, *21*(1), 73–85.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sobol, I. M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, *1*, 407–414.

Stewart, J. E. (1980). Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study. *Journal of Applied Social Psychology*, *10*(4), 348–361.

Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems*, *27*, 1988–1996.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning. Vol. 70: Proceedings of the machine learning research* (pp. 3319–3328). PMLR.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). 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). IEEE.

Talamas, S. N., Mavor, K. I., & Perrett, D. I. (2016). Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *PLOS ONE*, *11*(2), Article e0148284.

Tate, S. A., & Fink, K. (2019). Skin colour politics and the white beauty standard. In C. Liebelt, S. Böllinger, & U. Vierke (Eds.), *Beauty and the norm* (pp. 283–297). Springer.

Tifentale, A., & Manovich, L. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In D. M. Berry, & M. Dieter (Eds.), *Postdigital aesthetics: Art, computation and design* (pp. 109–122).

Wajcman, J. (2004). *Technofeminism*. Polity.

Ward, B., & Paskhover, B. (2019). The influence of popular online beauty content creators on lip fillers. *Aesthetic Surgery Journal*, *39*(10), NP437–NP438.

Winch, A. (2013). *Girlfriends and postfeminist sisterhood*. Springer.

Yan, Y., & Bissell, K. (2014). The globalization of beauty: How is ideal beauty influenced by globally published fashion and beauty magazines? *Journal of Intercultural Communication Research*, *43*(3), 194–214.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision—ECCV 2014* (pp. 818–833). Springer.

Zembylas, M., & Vrasidas, C. (2005). Globalization, information and communication technologies, and the prospect of a 'global village': Promises of inclusion or electronic colonization? *Journal of Curriculum Studies*, *37*(1), 65–83.

Zheng, D., Ni, X.-l., & Luo, Y.-j. (2019). Selfie posting on social networking sites and female adolescents' self-objectification: The moderating role of imaginary audience ideation. *Sex Roles*, *80*(5), 325–331.

Zheng, T., Deng, W., & Hu, J. (2017). Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*.

## Author Biographies

Piera Riccio (MSc ICT for Smart Societies, Politecnico di Torino; MSc Data Science & Engineering, Télécom Paris, EURECOM) is a PhD student at ELLIS Alicante. Her research interests include the influence of artificial intelligence (AI) technologies on human aesthetics.

Julien Colin (MSc Cognitive Sciences, Natural and Artificial Cognition, Grenoble Institute of Technology) is a PhD student at ELLIS Alicante. His research interests focus on human-centric explainable AI.

Shirley Ogolla (MA Media Studies, Humboldt University of Berlin) is a doctoral researcher at the Humboldt Institute for Internet & Society, investigating the introduction of AI in knowledge work.

Nuria Oliver (PhD, Massachusetts Institute of Technology) is Scientific Director of ELLIS Alicante. Her research interests include human-centric AI, AI for social good, human–computer interaction, and mobile and ubiquitous computing.