DHS S&T Technical Paper Series

Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms

John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility,
Identity and Data Sciences Lab

Arun R. Vemury

The U.S. Department of Homeland Security
Science and Technology Directorate
Biometric and Identity Technology Center

Keywords: Face Recognition, Iris Recognition, Demographic Differentials, Performance Evaluations, Technology Social Factors

May 2021

Homeland Security

Science and Technology

Executive Summary

OVERVIEW: This study was sponsored by the U.S. Department of Homeland Security (DHS) and conducted at the Maryland Test Facility (MdTF) as part of ongoing evaluations of biometric performance across demographics groups. Using data gathered from the 2018 Biometric Technology Rally, we show that commercial face, but not iris, recognition algorithms use features associated with race and gender to establish individual identity. Here, we propose a first-of-its-kind method to quantify the extent to which different biometric algorithms exhibit this effect. We discuss the implications of these findings in the context of equitable performance of face recognition in verification and identification use cases.

WHAT WE DID: We showed that face recognition similarity scores between different people who were the same race and gender tended to be higher than scores between different people who did not share those groupings. We did this on five leading commercial face recognition algorithms. We then developed and applied a conceptual framework to understand how this property of face recognition may affect the performance of face verification and identification systems and proposed a method of quantifying this effect in black-box commercial algorithms. Finally, we performed analyses indicating that race and gender features can be removed from these systems without reducing performance below useful levels.

MOTIVATION: Human faces contain different kinds of features, such as nose width, distance between the eyes, brow length, etc. Combinations of these features have been shown to be effective at determining both individual identity and demographic information, such as race and gender. However, some (but not all) of these features tend to be shared by members of demographic groups. For example, male noses are, on average, shorter and broader than female noses. Face recognition algorithms that rely on features that are shared within a demographic group will be more likely to incorrectly match people within that group. Currently, the extent to which black-box commercial face recognition algorithms use gender and race features to determine identity requires further study to better understand impacts on the increasing number of deployments by government and industry.

MAJOR TAKEAWAYS: We found that all commercial face recognition algorithms in our test tended to assign higher similarity scores to different people that were the same race and/or gender. We believe, in concurrence with evidence from the U.S. National Institute of Standards

MAJOR TAKEAWAYS (CONT.):

and Technology, that this is a general property of all currently tested face recognition algorithms. However, we showed that only roughly 10% of face recognition similarity score variation could be attributed to race and gender sameness. Moreover, when this information was ignored, we observed a decrease in overall algorithm performance, but also that algorithms were less likely to confuse individuals based on race and gender. This suggests it is possible for face recognition algorithms to operate on face features that are unrelated to gender and race, albeit with somewhat lower recognition accuracy. However, this is not the current commercial practice. Using the conceptual framework developed by this research, we show why development of face recognition algorithms that ignore race and gender is necessary for equitable outcomes in large, one-to-many identification operations. We believe these findings have strong implications for the development, training, and deployment of more equitable face recognition algorithms.

Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms

John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury

Abstract—Human face features can be used to determine individual identity as well as demographic information like gender and race. However, the extent to which black-box commercial face recognition algorithms (CFRAs) use gender and race features to determine identity is poorly understood despite increasing deployments by government and industry. In this study, we quantified the degree to which gender and race features influenced face recognition similarity scores between different people, i.e. non-mated scores. We ran this study using five different CFRAs and a sample of 333 diverse test subjects. As a control, we compared the behavior of these non-mated distributions to a commercial iris recognition algorithm (CIRA). Confirming prior work, all CFRAs produced higher similarity scores for people of the same gender and race, an effect known as "broad homogeneity." No such effect was observed for the CIRA. Next, we applied principal components analysis (PCA) to similarity score matrices. We show that some principal components (PCs) of CFRAs cluster people by gender and race, but the majority do not. Demographic clustering in the PCs accounted for only 10% of the total CFRA score variance. No clustering was observed for the CIRA. This demonstrates that, although CFRAs use some gender and race features to establish identity, most features utilized by current CFRAs are unrelated to gender and race, similar to the iris texture patterns utilized by the CIRA. Finally, reconstruction of similarity score matrices using only PCs that showed no demographic clustering reduced broad homogeneity effects, but also decreased the separation between mated and non-mated scores. This suggests it is possible for CFRAs to operate on features unrelated to gender and race, albeit with somewhat lower recognition accuracy, but that this is not the current commercial practice.

Index Terms—Face Recognition, Iris Recognition, Performance Evaluations, Demographic Differentials, Technology Social Factors.

1 Introduction

During the period from 2015 to 2020, face recognition experienced enormous increases in commercial investment, public interest, and public facing deployments. In 2014, deep convolutional neural nets (DCNNs) applied to face recognition achieved near human performance for the first time [1]. By 2016, at least two Fortune 500 companies began offering commercial facial recognition algorithms (CFRAs) via their cloud platforms to the general public [2], [3]. From 2015 to 2019, face recognition became the predominant method by which individuals access their personal devices, a list that now includes laptops, tablets, and smartphones [4], [5]. Government use of this technology also expanded during this period. In the travel environment, government face recognition services identify international travelers arriving to the United States [6] and also facilitate ticketless international departures [7]. Face recognition has also been adopted at the state and local level, particularly for policing [8], [9]. Broadly, these deployments have led to successes in identifying criminal suspects [10], [11] and detecting fraud [12], but have also resulted in notable false identifications [13], [14], [15].

- *J. Howard, Y. Sirotin, and J. Tipton work at the Maryland Test Facility in Upper Malboro, Maryland.*
- *A. Vemury works at the United States Department of Homeland Security, Science and Technology Directorate in Washington, DC.*
- *Authors listed alphabetically. E-mail correspondence should be sent to info@mdtf.org*

The increased use of face recognition in the public domain has also resulted in additional scrutiny, particularly around the topic of equitability and how this technology performs across race and gender categories. In 2016, the Georgetown Center for Privacy and Technology claimed that "facial recognition algorithms exhibit racial bias" in part due to over-representation of some demographic groups in law enforcement galleries [16]. This was followed by a 2018 report from the American Civil Liberties Union (ACLU), which claimed that publicly available face recognition software had incorrectly matched 28 members of Congress to a database of mugshot images [17] with false matches largely constrained along racial categories.

In the scientific community, there were early reports that CFRA performance varied for people based on their demographic group membership [18], [19], [20]. Even after the widespread application of neural nets to automated face recognition, these effects have continued to be documented by scientists [21], [22], [23], [24], [25], [26]. One type of demographic variation observed is the tendency of CFRAs to assign greater similarity scores to different individuals that share gender and race categories. For example, comparing images of women to images of other women produces higher scores relative to scores produced when images of women are compared to images of men [22], an effect termed "broad homogeneity" [24].

While intuitive based on human perception, this property of CFRAs can create undesirable behavior in many identification scenarios. For example, if an identification

gallery, such as a most-wanted list, skews predominantly male, then men who are not in the gallery are more likely to be misidentified when searched against that gallery than women, solely on the basis of their male facial features.

To some, the notion that a CFRA would find males more similar to other males as opposed to other females may be the expected, even desirable, outcome. However, others have pointed out that the combination of this effect and demographically homogeneous galleries will lead to unfair false positive identification rates [16]. In this study we present evidence that, while broad homogeneity effects appear to exist in nearly all currently tested CFRAs (a notion also supported by Annex 5 of [22]), it does not have to be so. We first document broad homogeneity effects in five CFRA's and one commercial iris recognition algorithm (CIRA). We then use a novel technique to quantify and compare these effects across black-box commercial algorithms. Finally, we provide evidence that ignoring facial components that cluster demographically similar imposter pairs, only results in a partial reduction in the d-prime metric between mated and non-mated score distributions. This suggests that it's possible for CFRAs to operate on features unrelated to gender and race, albeit with somewhat lower recognition accuracy, but that this is not the current commercial practice.

2 Background and Significance

2.1 Face Features

The human face has many features that can be measured to help establish identity. For example, intercanthal width is the distance between the inner portion of the eyes, and morphological nose width is the distance between the exterior nostrils [27]. The relative positions of *some* of these facial landmarks are shared by members of demographic groups. For example, the male nose is shorter, broader, and more projecting relative to females [28], [29] and people of Sub-Saharan African ancestry tend to have broader noses than people of European and East Asian ancestry [30], [31]. However, other face features and their combinations are unlikely to be associated with gender or race. For instance, genetic disorders can be associated to specific common changes in face shape [32], [33], [34]. Likewise, features thought to be formed stochastically during development, such as iris texture utilized by iris recognition (IR) algorithms are unique not only to specific individuals, but to each eye [35]. Indeed, recent work indicates that gender and race features in face images can be manipulated while identity information relevant for face recognition is maintained [36], [37].

2.2 Fairness Criteria with respect to False Match Rate in Face Recognition

All biometric samples inevitably share some common patterns. Biometric samples come from biological systems that may share some features due to common genetics, environment, or simply due to chance. When two biometric samples from different people are similar enough, biometric algorithms may label the two samples as matching, producing a false match. How often this error occurs for a given algorithm, on a given subset of people, is measured as the algorithm's false match rate (FMR).

The fairness[1] doctrine of disparate impact is relevant when considering demographic performance differentials in face recognition. Disparate impact occurs when the outcome of a process has different error rates for different groups, regardless of if those differences were unintentional or if the process was aware of the individual's group membership [40], [41]. There are at least two different criteria for what might be considered a fair face recognition algorithm in relation to FMR. The first is that FMR measured within specific groups should be equal for each group but that FMR measured between different groups can still take a different value. For example, a FMR of 1 in 1,000, when white males are compared to other white males and an equal FMR when black males are compared to other black males would satisfy this condition ($FMR_{(WM,WM)} == FMR_{(BM,BM)}$). However, under this criterion, FMR between black males and white males, for example, may be other than 1 in 1,000, presumably lower ($FMR_{(WM,BM)} < FMR_{(WM,WM)}$). In this scenario, a heatmap matrix of false match rates between various cohorts would appear as shown in Fig. 1A. For the purposes of this research, we call this criteria the "specific homogeneity fairness criterion."

A second possible face recognition fairness criterion, with respect to FMR, is that all within *and between* group FMRs should be equal. Under this condition the false match rate between white males when compared with other white males would equal both the false match rate when black males were compared with other black males and the false match rate when black males were compared to white males ($FMR_{(WM,WM)} == FMR_{(BM,BM)} == FMR_{(BM,WM)}$). In this scenario, a heatmap matrix of false match rates between various cohorts would appear as shown in Fig. 1B. For the purposes of this research we call this criteria the "broad homogeneity fairness criterion."

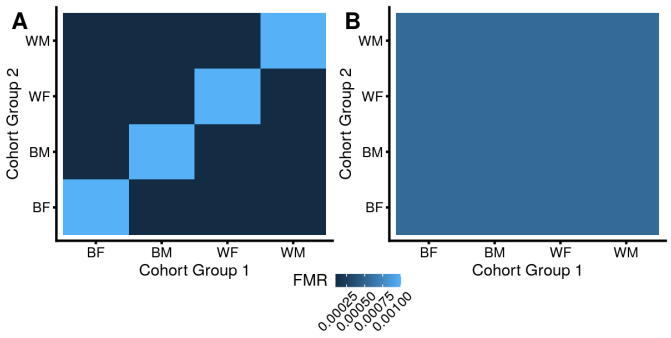


Fig. 1. Example heatmap matrices of false match rates within and across demographic cohorts using the **(A)** specific homogeneity fairness criterion and **(B)** broad homogeneity fairness criterion.

1. Fairness is a broad concept. The term itself does not currently have a concise technical definition [38], [39]. We use the term fairness criteria in the sense that these metrics relate to the *topic* of fairness, as they are used to reason about differential error rates. We believe these criteria are one element that can help inform perceptions of fairness more broadly. We do not attempt, nor do we believe it is possible, to measure fairness in the broader social and perceptual context with studies of algorithm outcomes alone.

2.3 The Consequences of Selecting Features Related to Demographic Groups, Imbalanced Galleries, and Identification Workflows in Face Recognition

Similarities in facial features can be related to demographics, including gender and race (Section 2.1). However, gender and race similarity alone are typically not enough to increase CFRA false match rates to unacceptable levels. Consequently, use of features related to gender and race has not been seen as a problem in the machine vision community. Nonetheless, small increases in one-to-one false match rate can lead to appreciable gains in one-to-many false positive identification rates, particularly when matching against large galleries [22]. This raises legitimate concerns about the fairness of CFRAs when matching against homogeneous galleries in law enforcement applications [16]. It is therefore important to understand the degree to which gender and race determine similarity scores produced by CFRAs.

A face recognition identification operation involves a one-to-many comparison of a unknown face probe (one) to a gallery composed of faces of known individuals (many). If the probe subject is not present in the gallery and any candidate is returned at a score above some threshold t, a false positive identification has occurred. How often this error occurs for a given algorithm, on a given subset of people, is known as the false positive identification rate (FPIR).

One argument for using the broad homogeneity fairness criterion (Fig. 1B) is that a biometric system that achieves equal FMR rates within, but not between groups, in one-to-one verification applications (i.e. satisfies the specific homogeneity fairness criterion; Fig. 1A) can still experience disparate impacts in one-to-many identification operations. Assuming each comparison in a one-to-many search is independent, the likelihood of a false positive identification occurring can be modeled as a function of the FMR of the underlying algorithm and the size of the search gallery (N), according to Equation 1.

$$\text{FPIR}(\text{FMR}_G, N) = 1 - (1 - \text{FMR}_G)^N \quad (1)$$

To illustrate the previous point, consider the case where a face recognition algorithm has achieved equal FMR rates within two demographic cohorts. For example, each intra-cohort FMR is 5.5e-5 and each cross cohort FMR is lower at 1e-7. For the purpose of this example, we will limit cohorts of interest to male (M) and female (F). This condition is shown in Fig. 2 (inset).

Now consider an unknown probe face image is searched against a gallery containing face images of 10,000 known individuals, but which is imbalanced such that females outnumber males with a ratio of 9:1. Expanding Equation 1 under the assumption that all comparisons are independent, but that within and between group FMR values differ, we can show that the likelihood of a female or male experiencing a false positive identification against this gallery can be calculated according to Equation 2. This is depicted in Fig. 2 as a function of the number of independent gallery comparisons performed.

$$\text{FPIR}_X = 1 - (1 - \text{FMR}_{(X,F)})^{N_F}(1 - \text{FMR}_{(X,M)})^{N_M} \quad (2)$$

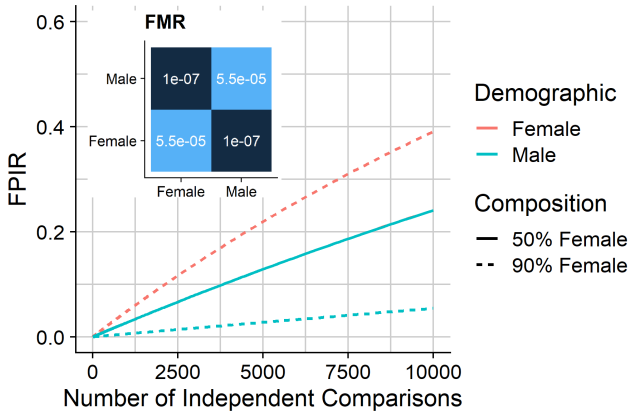


Fig. 2. FPIR for males and females as a function of the number of independent gallery comparisons given the specified gallery composition. The inset shows the FMR values used to generate the curves according to Equation 2. Note equal within-group FMR values lead to equal FPIR values for a balanced gallery (Composition: 50% female), but un-equal FPIR values for an imbalanced gallery (Composition: 90% female) [42].

If the probe image is of a female subject, they will experience a false match rate of 5.5e-3 against the 9,000 females in the gallery and a false match rate of 1e-7 against the 1,000 males. Using our numbers, we arrive at $\text{FPIR}_F = 39.0\%$. The situation is reversed for males with a resulting $\text{FPIR}_M = 5.4\%$. This disparate outcome in false positive identification rates exists despite completely equal within-cohort FMR (thus satisfying the specific homogeneity fairness criterion). Disparate outcomes in FPIR can be expected to occur in any face recognition algorithm that assigns greater similarity scores to individuals that share demographic characteristics, i.e. in any face recognition algorithm that does not satisfy the broad homogeneity fairness criterion.

2.4 Evaluation and Mitigation of Demographic Differentials in Commercial Algorithms versus Academic Algorithms

Face recognition algorithms and biometric recognition algorithms in general can be categorized as either commercial or academic. Much of the scientific literature, particularly around demographics in face recognition, focuses on academic algorithms [20], [43], [44], [45]. The implementation details of academic algorithms are usually published and the structure of their facial templates understood. However, leading commercial face recognition algorithms have superior performance relative to available academic algorithms [22] and come with the legal, financial, and operational support offered by commercial entities. Commercial face recognition algorithms are therefore often used by industry and government to make real-world decisions that may have a societal impact [7], [8]. Consequently, techniques to evaluate fairness and mitigate bias in commercial algorithms are required.

However, unlike academic algorithms, commercial algorithms are "black-boxes," meaning details of template structure [45] or face recognition algorithm architecture [20], [43] cannot be used in such evaluations and mitigations. The only available information for evaluating commercial face recognition algorithm performance are the similarity

scores they produce when comparing face images. Mitigation techniques similarly would only have this information available to them. This makes it necessary to develop methods of measuring and mitigating demographic differentials that rely only on these similarity scores, and not training, template data, or mechanistic algorithm insight.

3 METHODS

3.1 Dataset

Data used in this study were collected during the 2018 DHS S&T Biometric Technology Rally [46]. Biometric samples were collected from 333 diverse test subjects on 11 different face and five iris biometric acquisition systems. All acquisition system were commercially available systems from commercial biometric companies, available for purchase in 2018.

The test described in [46] produced 3,324 face and 1,414 left iris probe images (all devices failed to acquire images on some subjects). For this study, these probe images were compared to galleries of 1,205 face images and 1,083 left iris images previously gathered from the same subjects over a five-year period from 2012-2018. Five different CFRAs and one CIRA were used to independently generate biometric similarity scores. In total, this operation produced 21,558,281 similarity scores which form the basis of this study. All matching systems were commercially available systems from established biometric companies, available for purchase in 2019. To comply with information sharing agreements between the test organizers and technology providers, all algorithm names are aliased in this report as "face1", "face2", "face3", "face4", "face5", and "iris." Each algorithm produced an arbitrarily scaled similarity score for pairs of face or iris images. Larger scores corresponded to a greater likelihood that the two images belong to the same subject.

Demographic information, including race and gender, was self-reported by each of the 333 unique subjects (Fig. 3). Most subjects in our sample self-identified as Black or African American, or White. For this reason, comparisons of same gender and race versus different race and gender groups was restricted to these demographic groups.

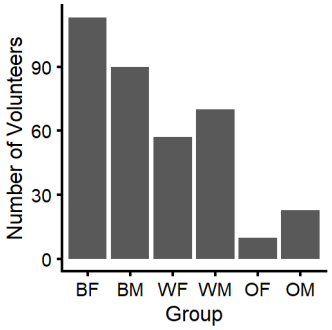


Fig. 3. Number of volunteers by self-reported demographic race (B, Black or African American; W, White; O, all others) and gender (F, Female; M, Male).

3.2 Analysis Techniques

3.2.1 99th Percentile Non-Mated Score

Biometric false match rates are driven by the behavior of the tail of the non-mated distribution. We quantified the characteristics of this tail using shifts in the 99th percentile score of the imposter distribution, similar to [24]. In Equation 1, $S_{(99,m)}$ is the subject-specific 99th percentile non-mated score, $I_{(n,m)}$ is the ordered set of non-mated similarity scores for subject m, and n is that subject's 99% highest non-mated similarity score.

3.2.2 Principal Component Analysis

The similarity scores described in Section 3.1, were arranged into a matrix, per recognition algorithm, where each entry at row i and column j represented the average similarity score between subjects i and j for each of our 333 subjects. All score matrices were symmetric, with the diagonal of each matrix corresponding to the average mated score for each subject. To understand the individual variations with the strongest association to subject similarity, we performed principal components analysis (PCA) on the matrix produced by each algorithm. PCA is a linear dimensionality reduction technique. It can be used to transform high dimensional data into a series of principal components (PCs) in such a way that the highest level of variance is found on the first component, PC_1. Each subsequent PC_k is orthogonal to the preceding and explains less variance ($\sigma_1^2 > \sigma_2^2 > ...\sigma_k^2 > ... > \sigma_n^2$). At some PC number k the full or a sufficient amount of the cumulative variance ($\sum_k \sigma_k^2$) has been explained by the PCs. The remaining $n-k$ PCs can be discarded, thus accomplishing dimensionality reduction. In this study, PC decomposition of each score matrix and subsequent operations were performed using built-in functions available in the R statistical programming language [47].

Prior work in demographics has measured performance variation in face recognition algorithms by comparing score distributions, error rates at fixed thresholds, thresholds required to obtain equal error rates, and area under ROC curves (AUC) [18], [19], [21], [22], [24]. PCA has several advantages over those approaches. First, PCA allows score matrices of different biometric algorithms to be compared using common units of explained variance. Scores of different algorithms are scaled arbitrarily and error rates depend critically on thresholds, which must be determined separately for each algorithm. While AUC offers the ability to compare overall algorithm accuracy as a function of demographics, modern face recognition algorithms may make no errors on some datasets, producing uniform AUC (AUC = 1). Our measure allows algorithm comparisons in the absence of errors. Finally, PCA allows us to reconstruct the score matrix after the removal of select PCs, which we will leverage to analyze the impact on score distributions when certain principal components are removed.

3.2.3 Demographic Clustering

Each PC computed as described in Section 3.2.2 corresponds to a pattern of score variation across 333 subjects. The similarity of face features between subjects in our dataset determines CFRA similarity scores. The PCs that explain

the most variance for each algorithm correspond to the shared feature patterns that are most heavily weighted by each algorithm in determining similarity. We assessed the degree of association of these features with gender and race by measuring the distribution of these groups across each PC. Specifically, we measured the degree of demographic clustering by calculating a clustering index C_k for each PC_k by taking the ratio of within group deviation from the mean to overall deviation from the mean across all subjects i in our sample according to Equation 2, where D is the set of subjects belonging to a specific demographic group and x_i is the value for subject i on the PC.

$$C_k = 1 - \frac{\sum_D \sum_{i \in D} (x_i - \bar{x_D})^2}{\sum_i (x_i - \bar{x})^2} \tag{3}$$

We assessed whether the clustering index value for each PC_k was statistically significant by comparing the calculated C_k values, which rely on the variance between subjects in real demographic groups ($\bar{\sigma}^2_{D,k} = \frac{1}{N} \sum_D \sum_{i \in D} (x_i - \bar{x_D})^2$), to the 99^{th} percentile of the distribution of C_{null}, where C_{null} is calculated by 500 shuffles assigning subjects to randomized demographic groups D. Given 333 PCs with no significant clustering, this criteria would, by chance label 3 as clustered.

Finally, to assess the overall demographic clustering for an algorithm, we measured the proportion of total variance in scores explained by demographic clustering according to Equation 4 where σ^2_k is the variance of PC_k, σ^2_{tot} is the total variance across the entire dataset, and C_k is as described in Equation 3.

$$C_{tot} = \frac{1}{\sigma^2_{tot}} \sum_k \sigma^2_k C_k \tag{4}$$

3.2.4 D-Prime Analysis

Since the PCs of algorithm similarity score matrices are orthogonal, it's possible to discard certain PCs and reconstruct score matrices as if these components did not exist. The reconstructed score matrices will have different distributions of mated (diagonal) and non-mated similarity (non-diagonal) scores. To quantify the separation between these two distributions, and the impact of this reconstruction step, we use the d-prime metric [48]. Previous studies of demographic effects in face recognition have also measured broad relative shifts in mated and non-mated distributions using d-prime [44]. We calculated the d-prime according to Equation 5 where μ and σ^2 are the mean and variance, and M and NM refer to the mated and the non-mated distributions of average similarity scores, respectively.

$$d' = \frac{\mu_M - \mu_{NM}}{0.5\sqrt{\sigma^2_M + \sigma^2_{NM}}} \tag{5}$$

4 Results

4.1 Consistent Effects of Broad Demographic Homogeneity across Commercial Face Recognition Algorithms

Prior work has shown, using a single CFRA, that the tail of the non-mated similarity score distribution between subjects of the same gender and race is higher than the tail of the distributions between subjects of different genders and race [24]. All five CFRAs in our sample reliably followed this broad homogeneity pattern (Fig. 4) Conversely, no effect of gallery homogeneity was observed for the CIRA.

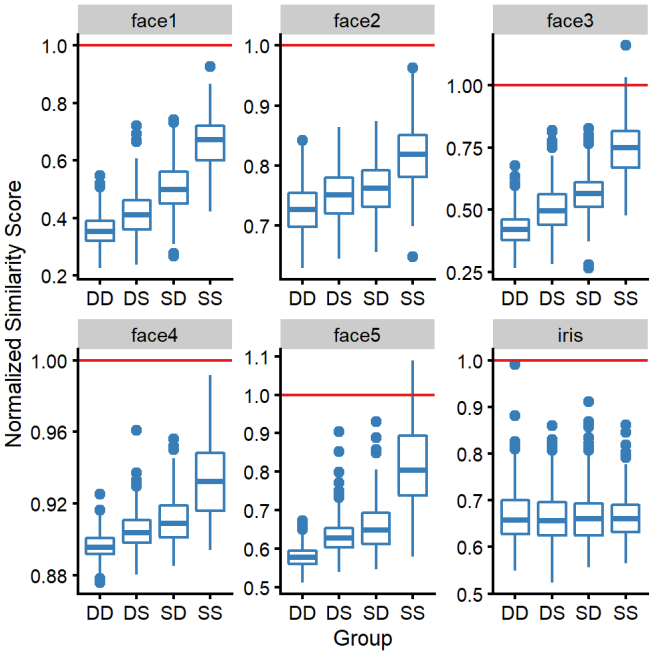


Fig. 4. Group homogeneity strongly modulates the tail of the non-mated distribution. Each facet corresponds to a different biometric algorithm and plots the 99th percentile of the non-mated score distribution (Section 3.2.1) across individuals. Scores on the y-axis for each algorithm are divisively normalized such that scores at the 1:10,000 threshold (red line) get a value of 1. Groups along the x-axis are as follows: DD, different gender and race; DS, different gender and same race; SD, same gender and different race; SS, same gender and race.

4.2 Face Recognition Score Matrices have Block-Diagonal Demographic Structure

Faces of different pairs of subjects have different features in common, only some of which are relevant to face recognition algorithms (Section 2.1). The patterns of similarity scores for individuals known to share various features can reveal how these features are weighted by the algorithm in calculating face similarity. Variation in CFRA similarity scores is driven both by face features as well as by the properties of the images used in the comparison [23]. To isolate the effect of face features for each algorithm, we computed 110,889 average similarity scores between each unique pair of the 333 subjects in our sample. Fig. 5 plots these average subject-to-subject similarity scores as a "score matrix" with rows and columns sorted based on the gender and race of each subject in our dataset (Section 3.2.2). Each score in this matrix is an average of 72 similarity scores between probe and gallery face images of the subjects and 28 similarity scores between probe and gallery left iris images of the subjects. As expected from Fig. 5, CFRA score matrices showed a clear block-diagonal structure with higher similarity scores for subject pairs within the same demographic group than between subjects in different demographic groups. This structure indicates the presence

of correlations in the data that could be leveraged by a dimensionality reduction technique, such as PCA.

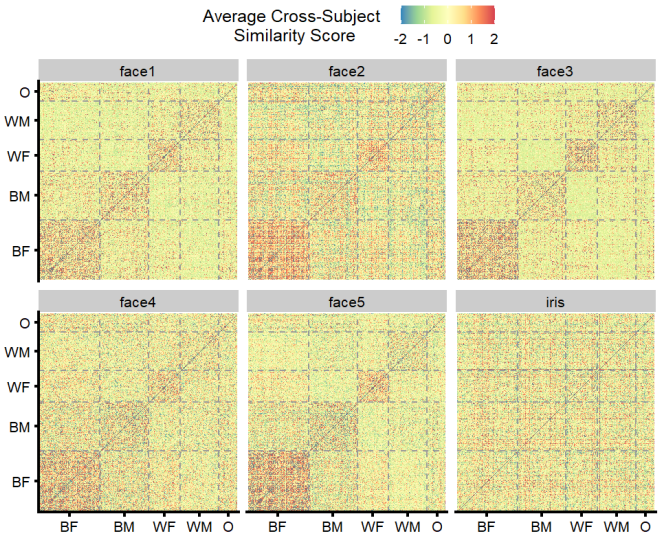


Fig. 5. CFRAs produce higher similarity scores within demographic groups. Each facet shows a raster plot of the average similarity scores produced for each pair of individuals in our sample. To aid visualization, scores have been normalized such that the non-mated scores have $\mu = 0$ and $\sigma = 1$. Dashed lines separate demographic groups. Note block diagonal structure present for all CFRAs, but not for the CIRA

4.3 Face Recognition Algorithms Cluster Individuals by Race and Gender

Fig. 5 and Fig. 4 suggest that all CFRAs show homogeneity effects. However, it is difficult to compare the magnitude of the effects across algorithms because similarity scores returned by each black-box CFRA are scaled arbitrarily. We used PCA (Section 3.2.2) to reduce the dimensionality of the similarity matric (Fig. 6) and isolate the contribution of particular components to the overall variation in the data.

After applying PCA, each PC corresponds to a score pattern across individuals in our sample. Assuming that score patterns are related to the face features of individuals, those patterns that explain the largest proportion of similarity score variance should therefore separate subjects based on the relative contribution of this feature to score variance. For instance, if similarity scores were determined solely by the relative width of the nose, then our subjects would be ordered based on nose width along the first PC of the score matrix. If, on the other hand, scores were not related to nose length, but rather related to the intercanthal width, then subjects would be ordered by distance between the eyes and not by nose length. Though we cannot know the important features used by the black-box CFRAs, we can examine the extent to which the order of subjects along each PC corresponds to demographic groups. Further, since each PC has a known contribution to overall score variance, we can quantify the extent to which known demographic categories determine the similarity scores.

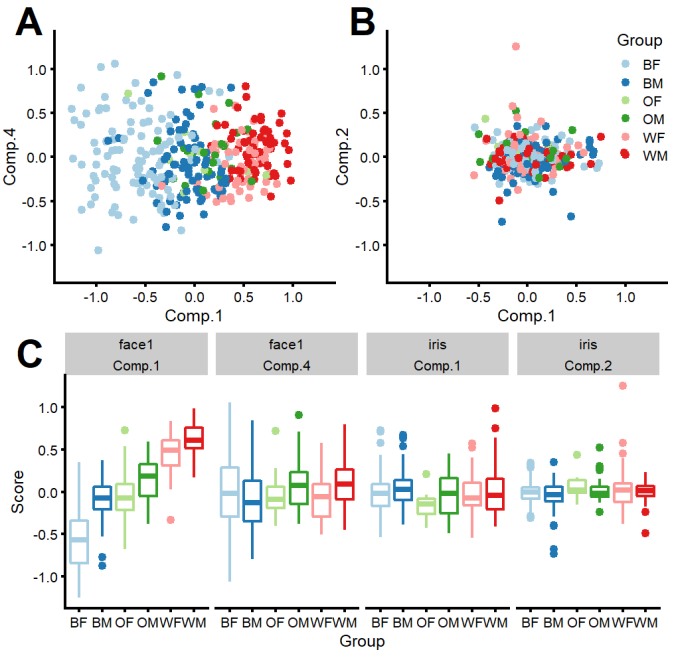


Fig. 6. Visualization of select principal components. **A.** Component 1 for algorithm face1 shows distinct clustering by demographic group, but component 4 does not. **B.** Components 1 and 2 for algorithm iris do not show demographic clustering. **C.** Distributions of component values visualized in association with different demographic groups.

4.4 Comparing Demographic Clustering Across Commercial Face Recognition Algorithms

We quantified the clustering illustrated in Fig. 6 by computing a clustering index for each PC (Section 3.2.3, Equation 3). The clustering index is bounded between 0 and 1, with zero signaling that the variance within each gender and race group is the same as overall variance. A clustering index of 1 indicates that there is no variance across individuals within each gender and race group.

Fig. 7 shows the clustering index for the first ten PCs of each algorithm. All five CFRAs showed statistically significant clustering for the first two PCs according to the test described in Section 3.2.3. Additionally, the first two PCs explained between 12% and 27% of the variance in similarity scores, depending on CFRA. None of the first ten PCs had significant demographic clustering for CIRA similarity scores.

To compare the extent to which different algorithms exhibited demographic clustering, we next measured the clustering index across all PCs (Section 3.2.3, Equation 4). On average, we found demographic clustering accounted for 10% of total CFRA score variance, ranging from 6% for "face4" to 14% for "face3" (Fig. 7B). Clustering accounted for less than 2% of the variance in similarity scores produced by the CIRA. Of the 333 PCs calculated for the CFRAs, on average 14 showed significant clustering, compared with one for the CIRA (Fig. 7C). Components with no significant clustering accounted, on average for 62% of total score variance for CFRAs. These components reflect face feature variances that are not associated with gender or race in our sample.

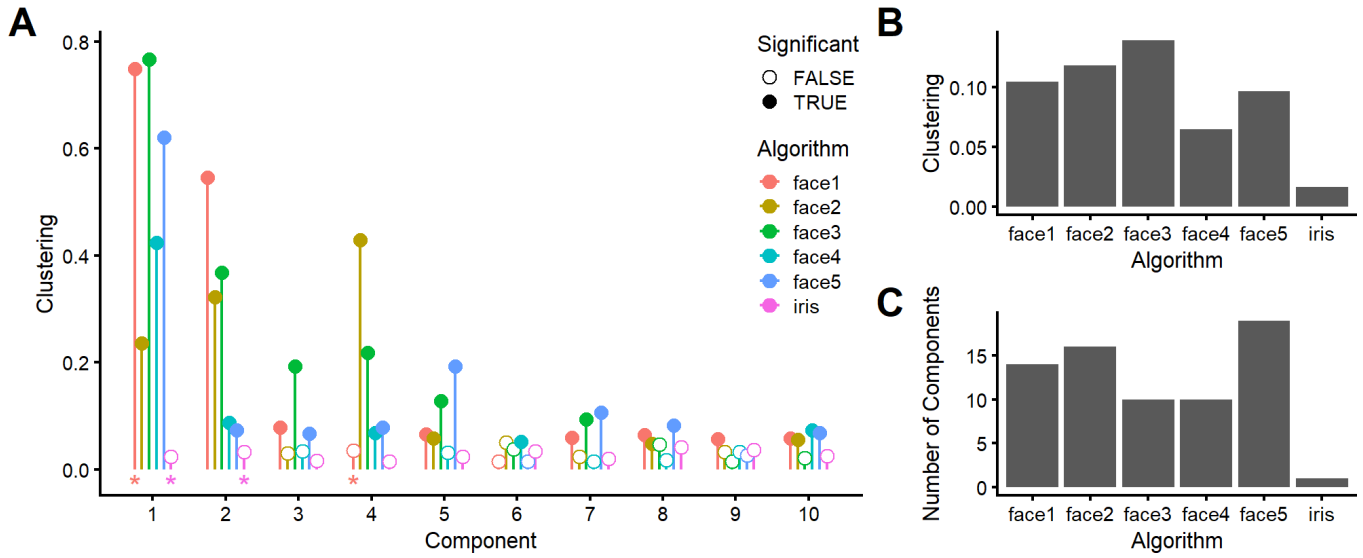


Fig. 7. Quantification of demographic clustering across algorithms. **A.** Stem plot of the demographic clustering index computed for each component. Filled circles correspond to components with statistically significant clustering (Section 3.2.3). Asterisks mark components visualized in Fig. 6. **B.** The total proportion of similarity score variance explained by demographic clustering for each algorithm. **C.** Number of principal components with statistically significant clustering for each algorithm $\alpha = 0.01$ (3 are expected by chance at this α value)

4.5 Estimating the Effects of Ignoring Demographically Clustered Features

We estimated the potential performance impact of having CFRAs ignore face features associated with gender and race. To do this, we reconstructed average similarity score matrices after removing all components with significant clustering and then compared the effects on the mated and non-mated distributions using the d-prime statistic (Section 3.2.4). Fig. 8 shows the distributions of average similarity scores in the original and reconstructed score matrices. As expected, removing PCs with significant clustering brought the non-mated distributions of average scores between subjects of the same gender and race (SS) closer to the non-mated average scores between subjects of different genders and races (DD). However, the operation also brought the overall mated and non-mated distributions closer together, decreasing d-prime. Nonetheless, even after reconstruction, d-prime values remained high for most algorithms. For example, if the range of original d-prime values (6.90 to 18.17) is considered to be the "useful range" (i.e. the range useful for commercial applications of face recognition technology), then 5 of 6 algorithms would still be in the useful range after reconstruction (5.22 to 12.70). Indeed, the leading CFRA after reconstruction, maintained a d-prime value larger than all other CFRAs before reconstruction. This suggests that ignoring face features associated with gender and race may maintain CFRA performance within a commercially useful range.

5 Discussion

5.1 Sample Size and Statistical Considerations

In this study, we tested whether algorithms cluster individuals based on race and gender in principal component space. Our null hypothesis was that there is no clustering. The statistical error of primary concern in this setting is a false positive, i.e. detecting significant clustering by chance when the effect does not actually exist. Consequently, we use a stringent significance level (α) of 0.01, meaning we have a 1% chance of detecting clustering by chance. Our sample size of 333 individuals was sufficient to detect significant clustering above chance in all CFRAs under test and to fail to detect significant clustering above chance level in the CIRA. While our sample was sufficient for detecting these effects in CFRAs, future studies employing larger samples may detect smaller effects or reveal additional interactions between demographic factors and clustering.

The aims of our study are distinct from biometric evaluations that include images of millions of individuals [22]. Tests such as [22] aim to find performance, particularly false match and non-match error rates, differences between algorithms. The null hypothesis in these settings is that no performance differences exist. The statistical error of primary concern is a false negative, i.e. failing to detect a true difference between algorithm A and algorithm B. Given the fact that biometric error rates can be extremely small, large sample sizes are needed to increase the power of the test ($1 - \beta$), allowing for low false negative rates. Indeed, future studies of this sort may be needed to compare the magnitude of demographic effects between algorithms but the effect sizes we document in this study are of sufficient magnitude that our sample size is adequate

5.2 Summary of Findings

In this paper, we discuss the extent to which five commercial face recognition algorithms (CFRAs) and one commercial iris recognition algorithm (CIRA) utilize face features associated with gender and race in determining individual identity. We first show that, non-mated similarity scores of all five CFRAs were higher between subjects of the same gender and race. We go on to quantify the proportion of score variance explained by gender and race information

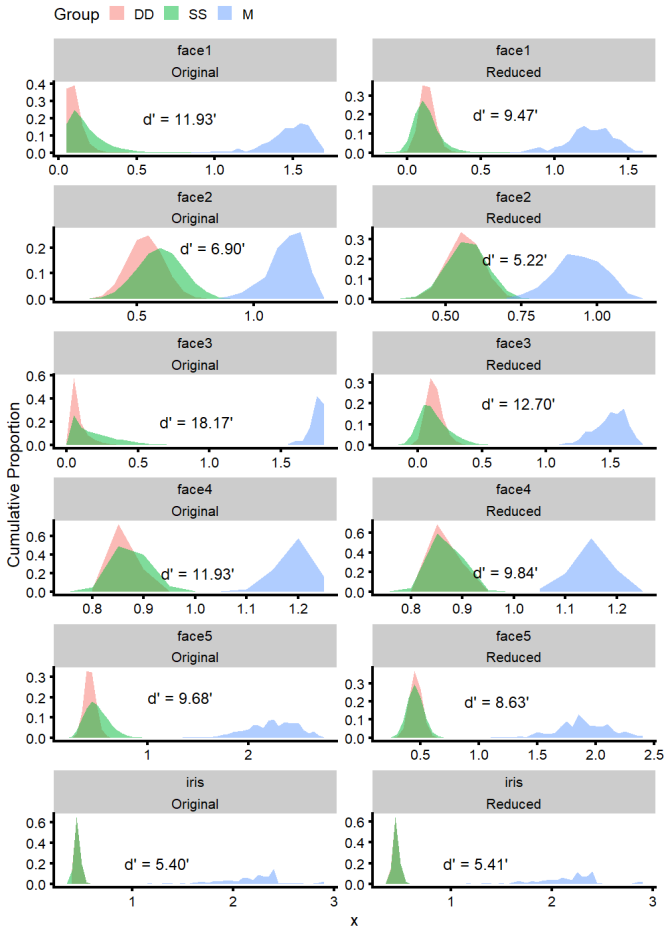


Fig. 8. Performance impact of removing components with significant demographic clustering. Each facet shows the empirical distributions of mated and non-mated scores in the original score matrix and after dimensionality reduction. Distributions are color coded as: SS, same gender and race; DD, different gender and race; M, mated. Black line shows the overall non-mated distribution. The SS and DD non-mated distributions include only individuals identifying as "Black" or as "White". Values of d' are based on the comparison of the mated distribution (M) to the full non-mated distribution (black line). Note that SS and DD distributions for reduced scores are closer together. Note also that M and the overall non-mated distributions (black line) move closer in the reduced face recognition matrices as quantified by d'.

across CFRAs using principal component analysis and show that some principal components cluster individuals by race and gender whereas most do not.

Use of face features associated with gender and race by CFRAs creates concerns regarding the fairness of these algorithms in some applications. Recent work by privacy groups [16] highlighted the fact that law enforcement face image galleries can be demographically homogeneous, with African American males comprising a majority of the faces. The demographic clustering documented in this research means that performing identifications against such galleries using images of out-of-gallery Black males would yield higher rank-1 similarity scores relative to White females. This practice could be seen as running contrary to a central fairness doctrine known as disparate impact [40], [41] (see Section 2.2). Additionally, to the extent that our sample of CFRAs and the sample tested in Annex 5 of [22] are representative, it would appear this condition exists when any current CFRA is used to search against a large homogenous gallery.

However, our research also shows that this disparate impact based on race and gender is likely avoidable. We found that most variation in CFRA similarity scores is not associated with race and gender. Further, separation between mated and non-mated score distributions reconstructed exclusively using PCs that do not cluster individuals by race and gender was only modestly reduced, suggesting CFRAs can maintain acceptable performance even when ignoring face features associated with race and gender. Indeed, recent work suggests that demographic features can be removed from face images while maintaining subsequent face recognition [36], [37]. This is what has long been observed in iris recognition. The periocular images used in iris recognition contain both the iris texture used for identification and the surrounding facial imagery. As such, they bear features related to demographics and both humans and algorithms can readily identify race and gender from periocular images [49], [50], [51]. Nonetheless, iris recognition algorithms based on iris-codes do not utilize these features in making identity determinations [35], a property likely closely linked with their ability to distinguish the irises of monozygotic twins [52].

5.3 Implications and Future Work

Our comparison between face and iris algorithms demonstrates some important points. First, it shows that some features of the face used by CFRAs do not carry obvious race and gender information. Second, it demonstrates an existence proof that it may be possible to select only these features and still perform accurate facial identifications. However, this does not appear to be the current commercial practice. One possible reason for this is that the DCNN technology that underpins most facial recognition algorithms, post-2014, maximizes recognition performance based on all available information rather than select features that have desirable behavior.

Since 2014, DCNNs have revolutionized face recognition technology, driving down error rates and supporting increasing technology deployment. Our work suggests that the gains made by the top performing algorithms have been so substantial that a modest reduction in performance from ignoring race and gender features may now be a worthwhile tradeoff to obtain, not only commercially useful, but also fair identity systems.

DCNN approaches to object classification have been shown to take "short-cuts" when accomplishing a task. For example, using correlated but ultimately spurious features in image data to arrive at a classification determination [53], [54]. Regardless of root cause, these effects are large enough in current CFRA's to be observed with relatively small populations. This is a positive in the sense that large image datasets are not required to study this issue, but it is also discouraging that these effects exist at this magnitude. Iris recognition developers considering a migration to DCNN technology should carefully evaluate their results for similar demographic effects.

Our research suggests prudence when using current CFRAs when performing identifications against large, homogeneous galleries and points to a need for audits of

operational systems to measure the extent to which the differential performance demonstrated here leads to differential outcome in operational use. Human review with orthogonal information may mitigate such occurrences.

Developing demographically-blind CFRAs that explicitly ignore face features associated with race and gender will help maintain fairness as use of this technology grows. This is fundamentally different than striving for equal false match error rates within demographic groups and lower false match rates between groups. Achieving the specific homogeneity fairness criteria, i.e. FMR within males equals FMR within females and is greater than FMR between males and females (see Figure 1A), only ensures that identification error rates will be equal if two, improbable conditions are met. First, the number of individuals in the identification gallery that belong to each demographic cohort must be the same. This is unlikely to ever be the case across all possible demographic groups (male, female, white, black, young, old, etc.). Second, even if cohort parity is achieved, un-equal identification error rates for members of different demographic groups would still occur in systems where individuals are matched against a gallery independent of their demographic group membership because of differences between homogeneous and heterogeneous FMRs, i.e. the FMR within males is greater than the FMR between male and female individuals. Thus, implementing accurate group detection and within-group-only matching would also be required, a practice not currently in wide use and prone to classification errors.

The most viable solution to this disparate impact in identification scenarios is to develop CFRA's that satisfy the broad homogeneity fairness criteria we have described here (see Figure 1B). This would effectively ensure that individuals are not spuriously matched to a gallery on the basis of their race or gender. We believe that reducing broad homogeneity effects, with the ultimate goal of achieving the broad homogeneity fairness criteria, should be a major focus for entities using or creating face recognition technology in the future.

6 Acknowledgments

This research was funded by the Department of Homeland Security, Science and Technology Directorate on contract number W911NF-13-D-0006-0003. The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers. The data used in these studies were collected from adults over the age of 18 who consented for their data to be used in biometrics related research. Data may be made available to select researchers. Please contact the authors. The data were acquired using the IRB protocol "Development and Evaluation of Enhanced Screening" number 120180237, approved by New England IRB.

The authors thank the following staff at the Maryland Test Facility for their contributions to this report: Andrew Blanchard, Kirsten Huttar, Michael Odio, and Roland Bell for providing software engineering support; Laura Rabbitt for human factors support; Kevin Slocum, Frederick Clauss, Gary Lake, and Brian Glatfeltner for providing network and integration engineering support; Jacob Hasselgren for directing Rally execution; Rebecca Rubin for technical document support and editing; Cynthia Cook for statistics support; as well as Colette Bryant and Rebecca Duncan for support in Rally organization and execution. The authors thank Jerry Tipton and Patty Hsieh for their broad support as well as review and comment on this manuscript.

The paper authors acknowledge the following author contributions: John J. Howard and Yevgeniy B. Sirotin conceived the work, generated biometric templates and comparisons, developed and performed analyses, and wrote the paper; Jerry Tipton and Arun Vemury conceived the work and edited the paper.

References

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

[2] "Microsoft Cognitive Services – General availability for Face API, Computer Vision API and Content Moderator." [Online]. Available: https://azure.microsoft.com/en-us/blog/

[3] "Amazon Rekognition - Automate your image and video analysis with machine learning." [Online]. Available: https://aws.amazon.com/rekognition/

[4] "The Future is Here: iPhone X." [Online]. Available: https://www.apple.com/newsroom/2017/09/the-future-is-here-iphone-x/

[5] "Say "Hello" to Windows Hello on Windows 10." [Online]. Available: https://blogs.windows.com/windowsexperience/2015/07/25/say-hello-to-windows-hello-on-windows-10/.

[6] U.S. Customs and Border Patrol, "CBP Expands Simplified Arrival in Arizona." [Online]. Available: https://www.cbp.gov/newsroom/national-media-release/cbp-expands-simplified-arrival-arizona#

[7] C. Manaher, "Privacy impact assessment for the traveler verification service," 2018. [Online]. Available: https://www.dhs.gov/publication/dhscbppia-056-traveler-verification-service

[8] J. P. O'Neill, "How facial recognition makes you safer," *New York Times*, 2019.

[9] "NYPD Questions and Answers Facial Recognition." [Online]. Available: https://www1.nyc.gov/site/nypd/about/about-nypd/equipment-tech/facial-recognition.page

[10] J. Jouvenal, "Police used facial-recognition software to identify suspect in newspaper shooting," *The Washington Post*, 2019, Accessed: 2021-03-15.

[11] C. McCarthy, "How NYPD's facial recognition software ID'ed subway rice cooker kook," *The New York Post*, 2019, Accessed: 2021-03-15.

[12] U.S. Custom and Border Protection, "Dulles CBP's New Biometric Verification Technology Catches Third Impostor in 40 Days," Accessed: 2021-03-15. [Online]. Available: https://www.cbp.gov/newsroom/national-media-release/dulles-cbp-s-new-biometric-verification-technology-catches-third

[13] K. Hill, "Wrongfully Accused by an Algorithm," *The New York Times*, 2020, Accessed: 2021-03-15. [Online]. Available: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[14] K. Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," *The New York Times*, 2020, Accessed: 2021-03-15. [Online]. Available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

[15] E. Anderson, "Controversial Detroit facial recognition got him arrested for a crime he didn't commit," *The Detriot Free Press*, 2020, Accessed: 2021-03-15. [Online]. Available: https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/

[16] C. Garvie, A. Bedoya, and J. Frankle, "The perpetual line-up. unregulated police face recognition in america. georgetown law center on privacy & technology, october 18, 2016," 2016.

[17] J. Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots," *The ACLU of Northern California*, 2018. [Online]. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

[18] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, pp. 1–11, 2011.

[19] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic effects on estimates of automatic face recognition performance," *Image and Vision Computing*, vol. 30, no. 3, pp. 169–176, 2012.

[20] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[21] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *IEEE transactions on biometrics, behavior, and identity science*, vol. 3, no. 1, pp. 101–111, 2020.

[22] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects," United States National Institute of Standards and Technology, Tech. Rep., 2019.

[23] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, 2019.

[24] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

[25] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[26] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 330–347.

[27] L. G. Farkas, M. J. Katic, and C. R. Forrest, "International anthropometric study of facial morphology in various ethnic groups/races," *Journal of Craniofacial Surgery*, vol. 16, no. 4, pp. 615–646, 2005.

[28] M. J. Kesterke, Z. D. Raffensperger, C. L. Heike, M. L. Cunningham, J. T. Hecht, C. H. Kau, N. L. Nidey, L. M. Moreno, G. L. Wehby, M. L. Marazita *et al.*, "Using the 3d facial norms database to investigate craniofacial sexual dimorphism in healthy children, adolescents, and adults," *Biology of sex differences*, vol. 7, no. 1, p. 23, 2016.

[29] A. Samal, V. Subramani, and D. Marx, "Analysis of sexual dimorphism in human face," *Journal of Visual Communication and Image Representation*, vol. 18, no. 6, pp. 453–463, 2007.

[30] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao *et al.*, "Modeling 3d facial shape from dna," *PLoS genetics*, vol. 10, no. 3, 2014.

[31] D. K. Liberton, *An investigation into genes underlying normal variation in facial morphology in admixed populations*. The Pennsylvania State University, 2012.

[32] I. Heulens, M. Suttie, A. Postnov, N. De Clerck, C. S. Perrotta, T. Mattina, F. Faravelli, F. Forzano, R. F. Kooy, and P. Hammond, "Craniofacial characteristics of fragile x syndrome in mouse and man," *European Journal of Human Genetics*, vol. 21, no. 8, pp. 816–823, 2013.

[33] V. Macho, A. Coelho, C. Areias, P. Macedo, and D. Andrade, "Craniofacial features and specific oral characteristics of down syndrome children," *Oral Health Dent Manag*, vol. 13, no. 2, pp. 408–11, 2014.

[34] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[35] K. Hollingsworth, K. W. Bowyer, S. Lagree, S. P. Fenker, and P. J. Flynn, "Genetically identical irises have texture similarity that is not detected by iris biometrics," *Computer Vision and Image Understanding*, vol. 115, no. 11, pp. 1493–1502, 2011.

[36] V. Mirjalili, S. Raschka, and A. Ross, "Privacynet: Semi-adversarial networks for multi-attribute face privacy," *arXiv preprint arXiv:2001.00561*, 2020.

[37] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in *European Conference on Computer Vision*. Springer, 2014, pp. 682–696.

[38] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[39] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[40] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[41] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[42] Y. Sirotin and A. Vemury, "Demographic variation in the performance of biometric systems: insights gained from large-scale scenario testing," *European Association of Biometrics, Virtual Event Series - Demographic Fairness in Biometric Systems (presentation)*, 2021, Accessed: 2021-04-27. [Online]. Available: https://mdtf.org/publications/EAB2021-Demographics.pdf

[43] V. Albiero, K. Bowyer, K. Vangara, and M. King, "Does face recognition accuracy get better with age? deep face matchers say no," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 261–269.

[44] K. Vangara, M. C. King, V. Albiero, K. Bowyer *et al.*, "Characterizing the variability in face recognition accuracy relative to race," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[45] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper, "Comparison-level mitigation of ethnic bias in face recognition," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020, pp. 1–6.

[46] J. J. Howard, A. J. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. R. Vemury, "An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–7.

[47] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[48] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behavior research methods, instruments, & computers*, vol. 31, no. 1, pp. 137–149, 1999.

[49] S. Lagree and K. W. Bowyer, "Predicting ethnicity and gender from iris texture," in *2011 IEEE international conference on technologies for homeland security (Hst)*. IEEE, 2011, pp. 440–445.

[50] A. Kuehlkamp and K. Bowyer, "Predicting gender from iris texture may be harder than it seems," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 904–912.

[51] D. Bobeldyk and A. Ross, "Predicting gender and race from near infrared iris and periocular images," *arXiv preprint arXiv:1805.01912*, 2018.

[52] J. Daugman and C. Downing, "Epigenetic randomness, complexity and singularity of human iris patterns," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1477, pp. 1737–1740, 2001.

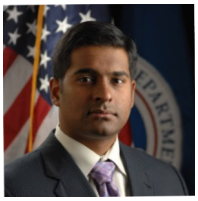
[53] S. Jabbour, D. Fouhey, E. Kazerooni, M. W. Sjoding, and J. Wiens, "Deep learning applied to chest x-rays: Exploiting and preventing shortcuts," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 750–782.

[54] A. Narla, B. Kuprel, K. Sarin, R. Novoa, and J. Ko, "Automated classification of skin lesions: from pixels to practice," *Journal of Investigative Dermatology*, vol. 138, no. 10, pp. 2108–2110, 2018.

John Howard. Dr. Howard received his Ph.D. in Computer Science from Southern Methodist University. His thesis was on pattern recognition models for identifying subject specific match probability. His current research interests include biometrics, computer vision, machine learning, testing human machine interfaces, pattern recognition, and statistics. He has served as the principal investigator on numerous R&D efforts across the intelligence community, Department of Defense, and other United States Government agencies. He is a member of the SAIC Identity and Data Sciences Lab and currently the Principal Data Scientist at the Maryland Test Facility.

Yevgeniy Sirotin. Dr. Sirotin holds a Ph.D. in Neurobiology and Behavior from Columbia University and has diverse research interests in behavior and human computer interaction. His past research spans mathematical psychology (cognitive modeling), neurophysiology (multi-spectral imaging of the brain), psychometrics (mechanisms of visual and olfactory perception), biometrics (design and testing of identity systems), and human factors (usability). He currently works as Principal Investigator and Manager of the Identity and Data Sciences Laboratory at SAIC which supports applied research in biometric identity technologies at the Maryland Test Facility.

Jerry Tipton. Jerry Tipton is the Program Manager and Director of SAIC's Identity and Data Sciences Lab. He has over 20 years experience in the biometric industry with over 15 years managing research portfolios in support of various United States Government agencies. He currently supports the S&T at the Maryland Test Facility.



Arun Vemury Arun Vemury received his Master of Science in Computer Engineering from George Washington University. His current research interests include biometrics, pattern recognition, machine learning, and operations research. He serves as the Director of the Biometrics and Identity Technology Center for S&T.