Who Wore It Best? And Who Paid Less? Effects of Privacy-Preserving Techniques Across Demographics

Xavier Merino, Michael King
Florida Institute of Technology
Melbourne, FL, USA
{xmerino2012,michaelking}@fit.edu

Abstract

Face recognition technologies, widely adopted across various domains, have raised concerns related to demographic differentials in performance and the erosion of personal privacy. This study explores the potential of "cloaking"—a privacy-preserving technique subtly altering facial images at the pixel level in order to reduce recognition accuracy—in addressing these concerns. Specifically, we assess the effectiveness of the state-of-the-art Fawkes algorithm across demographic groups categorized by race (i.e., African American and Caucasian) and gender. Our findings reveal African American males as the most significant beneficiaries of this protective measure. Moreover, in terms of cost-effectiveness, the African American demographic, as a collective, enjoys greater protection with fewer visual disruptions compared to Caucasians. Nevertheless, we caution that while cloaking techniques like Fawkes bolster individual privacy, their protection may not remain absolute as recognition algorithms advance. Thus, we underscore the persistent need for prudent online data-sharing practices.

1. Introduction

The growing prominence, versatility, and accuracy of face recognition technologies have become increasingly evident. At its core, face recognition relies on algorithms designed to identify individuals based on their unique facial features. The advancement of deep learning methods, a subset of machine learning in which neural networks process and analyze vast amounts of data, has significantly enhanced the capabilities of this technology. Deep learning enables the identification of faces even in challenging situations, such as low-quality images, varied poses, or different facial expressions [8].

Face recognition technology finds application in a wide spectrum of domains. It plays a crucial role in crime prevention, criminal investigations, and enhancing security by identifying individuals in surveillance footage [30]. Additionally, it simplifies everyday tasks like automatic photo tagging [32] and authentication to devices and services. Beyond these utilitarian uses, businesses harness face recognition to create personalized user experiences and launch targeted advertising campaigns [18, 52]. However, the integration of this technology into our daily lives raises concerns. Training machines with the task of identifying and categorizing human faces means that these algorithms not only replicate our discerning abilities but also our inherent shortcomings [35].

Coupled with privacy apprehensions, reports of performance differentials within these algorithms, influenced by factors such as race, gender, and ethnicity [49], also emerge as a significant concern. The human inclination to more accurately recognize faces of one's own race seems to also manifest within these systems [35], leading to discrepancies in accuracy for underrepresented demographics. Additionally, there are instances of false positive identifications, with a particularly high occurrence among people of color. The unsettling incident involving Porcha Woodruff [21], who, while eight months pregnant, was wrongly accused, detained, and arrested on carjacking charges, underscores the severe repercussions of such errors.

Amid these concerns and the ease of web scraping [3], particularly exemplified by companies like Clearview.ai amassing over 30 billion images without explicit consent [43], public apprehensions regarding the use and privacy implications of their personal photos have surged. This has driven a desire among individuals to regain control over their privacy and has led to the development of protective techniques such as de-identification or "cloaking." These methods aim to distort photographs at the pixel level, rendering them resistant to the recognition capabilities of machine learning models [10, 45].

In this work, we explore the privacy-preserving power of Fawkes [45], a state-of-the-art cloaking algorithm, and consider its efficacy relative to demographic groups for which differentials in performance have been characterized. More

specifically, we measure its effectiveness in reducing the accuracy of face recognition software across demographic categories based on race and gender. We say that the demographic group that experienced the most protection "wore it best." We also quantify the extent of the visual disturbances induced by the cloaking process, enabling us to ascertain which demographic experienced the least visual disruption. In addition, we introduce a protection-to-disturbance ratio that quantifies the added privacy in exchange for image quality in order to establish which demographic "paid less." To the best of our knowledge, this is the first study that explores the impact of privacy-preserving tools across demographics.

This work is structured in sections. Section 2 explores related works in face recognition systems pertaining to race, gender, and ethnicity, examining their societal impact and the efforts to mitigate them. Section 3 outlines the methodology used to assess the effectiveness of Fawkes in preserving privacy and details the "who wore it best" and "who paid less" objectives. Section 4 discusses the results and implications of this study. Section 5 provides concluding remarks and summarizes key insights. Section 6 outlines areas of future research.

2. Background and Related Works

Despite its technological progress, face recognition systems still grapple with demographic differentials shaped by factors such as race, gender, and age. Beyond affecting accuracy, performance differentials, in certain use cases, may introduce profound societal implications. In this chapter, we explore the intricacies of these challenges, their societal impact, and the efforts to counteract them.

2.1. Performance Differences in Face Recognition

The "other-race effect" (ORE) [48], wherein people recognize faces of their own race more accurately than other races, extends to computational algorithms. As highlighted by Cavazos et al. [9], the research over three decades unveils racial disparities in algorithm performance. Echoing this, Nagpal et al. [34] found that networks trained mostly on darker-skinned faces focus on the lips and eyebrows for recognition, while those trained on lighter-skinned faces center on the facial boundaries. Furthermore, research by Atzori et al. [1] found that in low-resolution face images, where features blur, algorithms exhibit diminished racial influences, offering consistent performance across racial groups. However, when these algorithms process high-resolution images, where facial details are more pronounced, the familiar racial disparities re-emerge. Such patterns resonate with human tendencies: darker-skinned individuals often emphasize features like lips, while lighter-skinned counterparts focus on facial shapes and irises [11, 17]. This parallel suggests that algorithms might inadvertently mirror human behavior [35], thereby encoding the ORE.

Moreover, differentials in the performance of face recognition span beyond just race. Robinson et al. [39] explored the intersection of race-gender, discovering that subjects most accurately identified individuals within their own race-gender cohort, followed by those of the same race but opposite gender. Contrastingly, performance declined when identifying those outside this subgroup. Buolamwini and Gebru's [6] work underscored differences in accuracy relative to gender and skin tone, noting that commercial classifiers misgendered one in three darker-skinned females compared to 0.8% of lighter-skinned males. Khalil et al. [26] then shed light on the role of non-demographic cues, like hairstyles, in gender identification, pointing to cultural influences on gender perceptions. Bhatta et al. [4] further explored the role of hairstyles, suggesting hair occlusions contribute significantly to the gender disparity.

Finally, Terhost et al.'s [47] expansive study on 47 non-demographic attributes, such as accessories, hairstyles, and facial expressions, revealed a decrease in accuracy for adorned faces, specific hairstyles, or certain facial features. Especially affected were darker-skinned female faces with glasses, emphasizing the compounded impact of multiple attributes. Given the complexities of face recognition, understanding how these performance differences manifest themselves in our daily experiences and societal structures becomes important.

2.2. Social Implications

Face recognition technology is deeply woven into modern living, spanning from smartphones and security systems to sectors like retail and finance. Yet, its roots in imagery expose long-standing implications of demographic under-representation in data used for technology development [27]. Historically, photography techniques, prevalent since the 1840s, tilted towards capturing white skin, a prejudice symbolized by the "Shirley card"—the industry's color balance standard until the 1990s [28]. Even with the advent of multi-racial cards [42], remnants of these deficiencies linger in modern cameras [7]. Examples include the HP face-tracking webcam's difficulty with darker complexions and Nikon's Coolpix S360 misinterpreting open Asian American eyes as blinking [41]. Beyond these inaccuracies, systemic flaws have broader implications [31], such as Amazon's AI recruitment tools displaying a male bias [12].

In the legal domain, face recognition errors have led to serious consequences, including the wrongful arrest of individuals like Porcha Woodruff [21] and Njeer Parks [20], predominantly affecting people of color [33]. Johnson et al. [25] highlight a marked racial disparity in such arrests across many U.S. cities in 2016, noting a significant 67% Black-White arrest gap in areas with advanced surveillance.

2.3. Addressing Performance Differentials

Face recognition differentials in accuracy stem primarily from data-driven and scenario-driven elements [9]. To produce equitable models, model trainers must grapple with both dimensions.

Data-driven elements hinge on the nature of training data, encompassing varied representation of demographics, data collection methods, and algorithmic designs catering to diverse facial features. Comprehensive training databases must span a spectrum of skin tones, backgrounds, ages, genders, facial expressions, and angles. Notably, Buolamwini and Gebru [6] stressed gender and skin type balance in datasets, while Robinson et al. [40] emphasized race and gender. Other research incorporated elements like hairstyles [4] or broader non-demographic features [47]. But quantity isn't quality. Khalil et al. [26] highlighted pitfalls in using open-source image databases, like skewed representation from celebrity dominance in certain ethnic groups. Equally important, human limitations can infiltrate training data labels [22], inadvertently perpetuating prejudices. Thus, a diverse developer team [15], bringing varied cultural and experiential perspectives, can help preempt and correct system inaccuracies.

Scenario-driven elements, meanwhile, deal with threshold adjustments aiming for consistent False Accept Rates (FAR) across subgroups. Proponents relate this to the "other-race effect"—the human inclination to better recognize faces from one's race—advocating adaptive thresholds based on attributes like race [9, 39]. Yet, concerns emerge in situations where racial and ethnic distinctions blur, complicating threshold settings for mixed-race or multicultural individuals [5, 19].

While developers possess some tools to address algorithm deficiencies, end-users are largely beholden to developers for impartial models, making them susceptible to residual system errors. Inaccuracies therein can lead to unintended consequences. This reality has propelled privacy-conscious users to seek alternative measures to shield themselves from unauthorized face recognition.

2.4. Fawkes Algorithm

Face recognition, while practical, raises issues of privacy and misidentification. To address this, privacy-centered solutions like Fawkes [45] have emerged, empowering users against unsanctioned face recognition.

Fawkes targets unauthorized web scraping of user images, subtly distorting them to hinder accurate recognition. Fawkes has been allegedly engineered to be effective against a spectrum of widely-used pre-trained face recognition models, specifically Microsoft's Azure Face, Amazon's Rekognition, and the Chinese Face++.

Initially, the algorithm selects random images from public datasets and computes face feature vectors for them. From this collection, it identifies a vector that bears the least resemblance to the user's image. Using this dissimilarity as a foundation, Fawkes crafts a "cloak" for the user's image. This cloak undergoes refinement, with the algorithm striving to minimize differences with the chosen target image's feature vector, all while operating within a defined perturbation budget.

A higher perturbation budget enhances privacy but risks visible distortions. Fawkes uses the "Structural Dissimilarity Index Measure" (DSSIM) to ensure cloaked images remain visually close to the originals but misleading in the feature space. By default, it aims for a DSSIM under 0.007, balancing privacy with visual fidelity. It has been noted, however, that using an aggressive cloaking preset might introduce visible artifacts—such as bluish spots or unusual indentations—that could deter users due to aesthetic concerns [24]. We deliberately choose Fawkes over other methods such as AnonymousNet [29], FSAP [51], DeepPrivacy [23] or IdentityDP [50] because Fawkes does not alter facial structure (e.g., jawlines, angles, ridges, etc.), features (i.e., masculinization or feminization), hair (i.e., head, facial hair), or pose. This ensures that individuals are not dissuaded from using the tool due to conspicuous, and potentially undesirable, changes in their appearance.

3. Methodology

This section details the methodology employed to assess Fawkes' effectiveness in preserving privacy across demographic cohorts categorized by race and gender. Drawing from the MORPHv3 dataset [38], which contains standardized mugshot-style photographs from the U.S., our goal is to identify which demographic benefits the most from a privacy standpoint—a concept we phrase as "who wore it best." By examining the cloaked images, we gauge the extent of perturbation induced by the cloaking process, enabling us to ascertain which demographic experienced the least visual disruption. Additionally, we establish a protection-to-disturbance ratio, termed "who paid less", which quantifies the compromise in image quality in exchange for enhanced privacy.

3.1. Dataset and Face Recognition Matchers

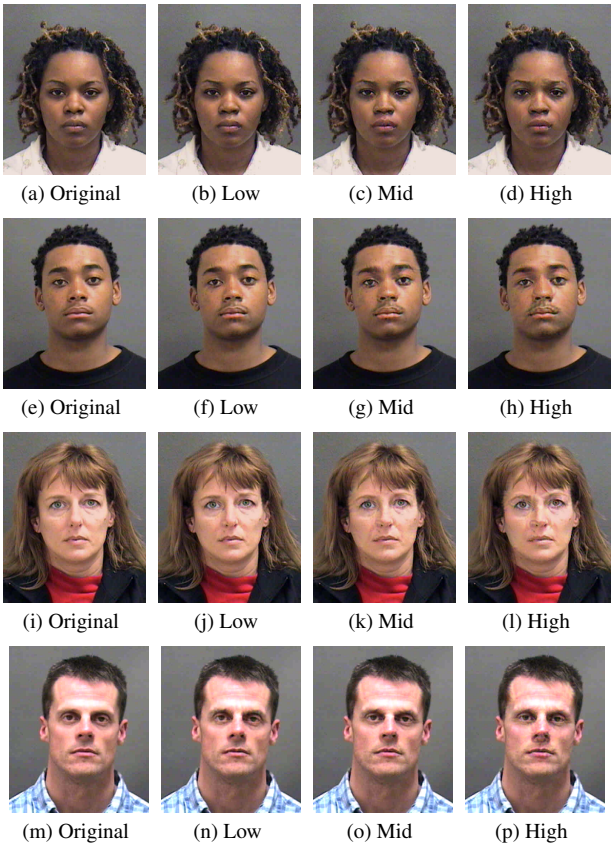
The MORPH dataset, sourced from public records, consists of images captured under conditions characteristic of frontal pose mugshots, ensuring consistent indoor lighting and a uniform 18% gray backdrop. Initially curated for facial aging research, this dataset's verified labels have since made it popular for characterizing the performance of face recognition technologies across demographics. For our analysis, we rely on MORPH v3, which includes:

- **Caucasian females (CF):** 10,941 images spanning 2,798 unique identities

- **Caucasian males (CM):** 35,276 images spanning 8,835 unique identities
- **African-American females (AAF):** 24,857 images spanning 5,929 unique identities
- **African-American males (AAM):** 56,246 images spanning 8,839 unique identities

We process this data using Fawkes, configuring it with three distinct preset perturbation levels: low, mid, and high. Within Fawkes, MTCNN handles face detection and alignment. Cloaking is then specifically applied to the facial region. This procedure yields three cloaked image variants for every original from the MORPH dataset, each corresponding to one of the perturbation levels. See Fig. 1 for an example of cloaked images.

For face recognition tasks, we employ ArcFace [13]. This integrates the SCRFD [16] model for detection and glintr100 model [46] for recognition. With this method, each image is converted into a 512-dimensional feature vector, which is subsequently matched using cosine similarity. Additionally, we leverage a COTS algorithm, which employs a proprietary matching technique.



(a) Original (b) Low (c) Mid (d) High

(e) Original (f) Low (g) Mid (h) High

(i) Original (j) Low (k) Mid (l) High

(m) Original (n) Low (o) Mid (p) High

Figure 1. Cloaked Images

3.2. Who wore it best?

This first objective focuses on quantifying the degree of protection against face recognition provided by varying cloaking intensities (low, mid, high) within African American and Caucasian demographics. The aim is to determine which demographic benefits the most.

To achieve this, we utilize the d-prime as a measure to quantify the separation between the original, non-cloaked, mated distribution and the mated cloaked distributions (low, mid, high) for each demographic subgroup (CF, CM, AAF, AAM). In addition to the d-prime, we also report the Earth Mover's Distance (EMD) as a supplementary metric. We say that the demographic subgroup with the most substantial decrease from the original authentic scores, based on the metrics discussed, benefits the most from cloaking and therefore "wore it best."

3.3. Who paid less?

This second objective seeks to determine the cost-effectiveness of cloaking by identifying which demographic subgroup incurred the least visual disturbance for the afforded degree of protection. In other words, the ratio of how much protection they acquired given the amount of disturbances introduced during cloaking. We say that the demographic subgroup with the most cost-effective protection, given a fixed cloaking level, "paid less."

To this end, we first utilize image quality assessment (IQA) metrics to quantify the visual disturbances added by Fawkes with respect to the original, non-cloaked, image. We then calculate the protection-to-disturbance ratio employing the previously obtained d-prime and EMD metrics as indicators of protection. We make use of three IQA metrics: VIF [44], PieApp [36], and DISTS [14].

- **Visual Information Fidelity (VIF):** Evaluates image quality by contrasting distorted images with a reference, targeting detailed textures and structures to mirror human perception. Scoring from 0 to 1, lower scores denote less fidelity. To exemplify its industry relevance, Netflix incorporates VIF into its video quality assessment processes [37].
- **PieApp:** Predicts perceptual image quality with a strong correlation to human opinion. The scores range between 0 and 1, indicating the likelihood that humans will prefer one image over another.
- **DISTS:** Measures variations in structure and texture, aligning with human ratings of image quality. The scores range from 0 to 1, with higher scores indicating higher fidelity.

4. Discussion and Results

4.1. Who wore it best?

This section focuses on quantifying the degree of protection against face recognition provided by varying cloaking intensities (low, mid, high) for each demographic subgroup (CF, CM, AAF, AAM) to determine which cohort benefits the most. The comparisons are done between each cloaking level's authentic distribution and the original, non-cloaked, authentic distribution, using the original pictures as the gallery and the cloaked counterparts as probes. We present the resulting matching distributions in Fig. 3. For each plot, yellow represents a baseline comparison, green represents a comparison between low-level cloaked images and the original, red represents a mid-level cloak, and blue represents a high-level cloak. Mated distributions are colored, while non-mated distribution are colored and hatched. Our focus is directed to Fawkes' ability to reduce the accuracy of mated comparisons, with the impostor comparisons included for completeness only.

The authentic distributions in Fig. 3 show a pronounced downward shift in the recognition scores as the cloaking intensifies. To measure the impact of the cloak, we employ d-prime to quantify the separation between the cloaked distributions and the original. In all cases, the separations are the result of the cloaked distribution shifting towards lower scores. We use this as a proxy to measure protection, with higher separations toward lower scores corresponding to higher privacy-preserving power. To address potential limitations of d-prime, especially as it relates to unequal variances and distribution shapes, we report the Earth Mover's Distance (EMD) as a complement to this analysis. The EMD captures differences due to shifts, shape, and spread, providing a more comprehensive measure of difference. As with d-prime, we use EMD as a proxy to measure protection, with higher distances toward lower scores representing higher privacy-preserving power. See Tab. 1 for a compilation of the measures across demographic subgroups and cloak level.

In our analysis, both matchers concur via the d-prime and the EMD in designating African-American males (AAM) as the cohort benefitting the most from cloaking at all intensities, followed by African-American females (AAF). Among the Caucasian cohorts, the results diverge based on the matcher, with females (CF) and males (CM) seemingly tied when considering their standing (i.e., each achieving 6 wins) across matchers.

See Fig. 2 for a visual representation of the protection afforded as a cohort progresses to each cloaking level. Across all demographics, both d-prime and EMD suggest that the transition from the baseline to a mid cloaking level results in the most pronounced enhancement in protection. This can also be seen in the histograms in Fig. 3. Concerning ArcFace, there's a proportional relationship between increased cloaking and enhanced protection. However, the shift from a medium to high cloaking level displays a more muted increase, suggesting diminishing returns at greater cloaking intensities. For the COTS matcher, while the low-level cloaking provides very little protection, both medium and high levels yield an increase in privacy. It is worth noting that the COTS matcher exhibits a degree of robustness against the Fawkes algorithm.

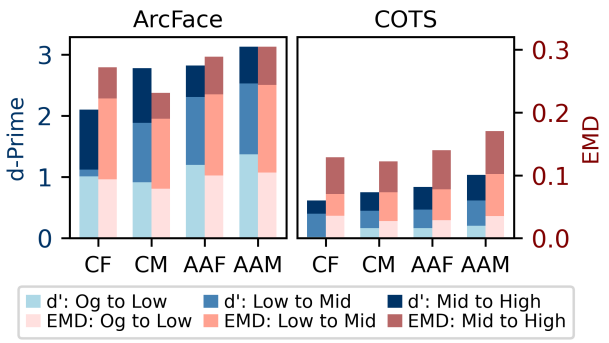


Figure 2. Protection as Cloaking Intensifies

We say that the demographic subgroup with the most substantial decrease from the original authentic scores benefits the most from cloaking and therefore "wore it best." This distinction belongs to AAM, followed by AAF, CM, and CF.

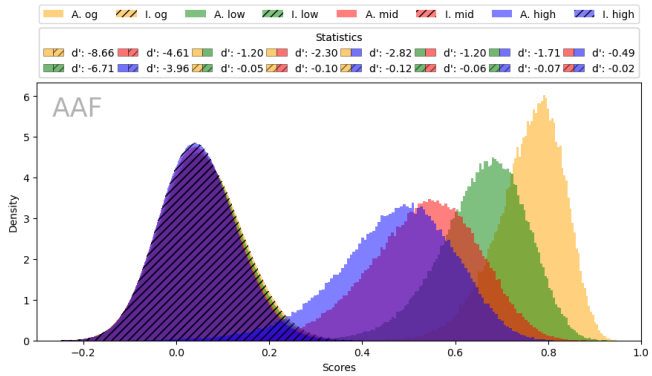
4.2. Who paid less?

This section delves into evaluating the cost-effectiveness of cloaking by identifying the demographic subgroup that experienced the least visual disturbances for each protection level. We begin by quantifying these disturbances using IQA tools and subsequently assess the economics of cloaking. This assessment involves calculating the ratio of protection obtained by each cohort relative to the cost incurred through the disturbances, allowing us to determine which cohort "paid less".

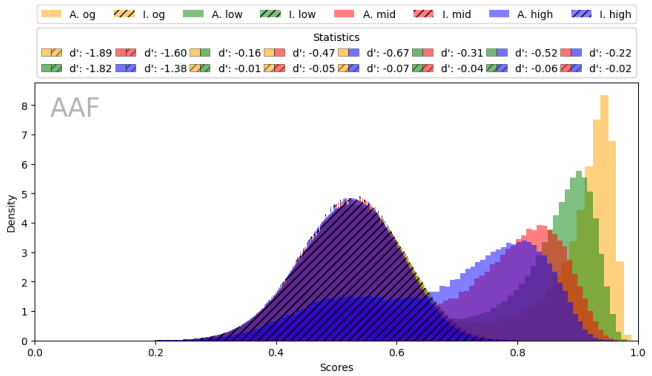
4.2.1 Assessing Visual Disturbances via IQA

We measured the visual disturbances arising from cloaking by contrasting the cloaked image to its original counterpart, utilizing three IQA methods: VIF, PieApp, and DISTS. These algorithms grade fidelity on a scale of 0 to 1; thus, we determined the disturbance measure by taking the complement (i.e., subtracting the fidelity score from 1).

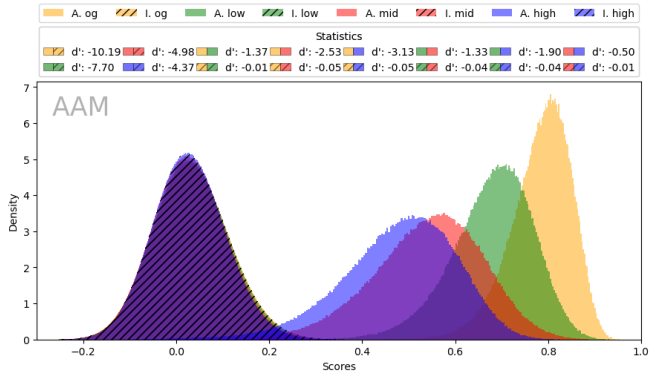
While there are dedicated algorithms for Face Image Quality Assessment (FIQA), general-purpose IQA methods have demonstrated greater consistency with respect to race and gender [2]. Notably, FIQA methods tend to fa-



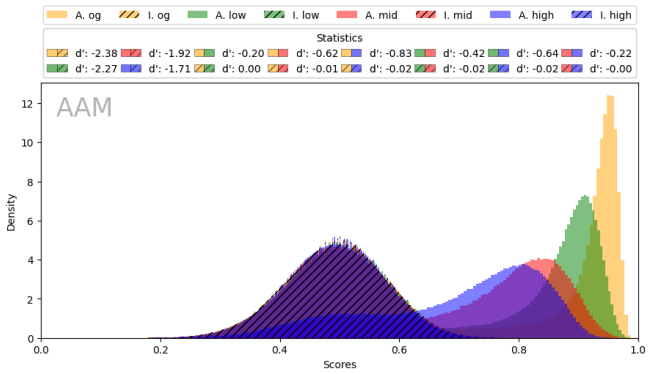
(a) African-American Females — ArcFace



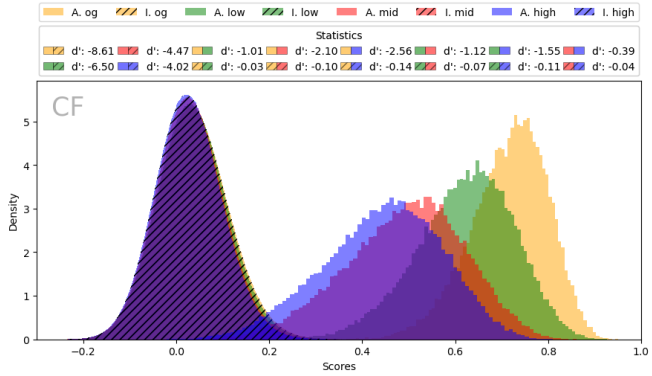
(b) African-American Females — COTS



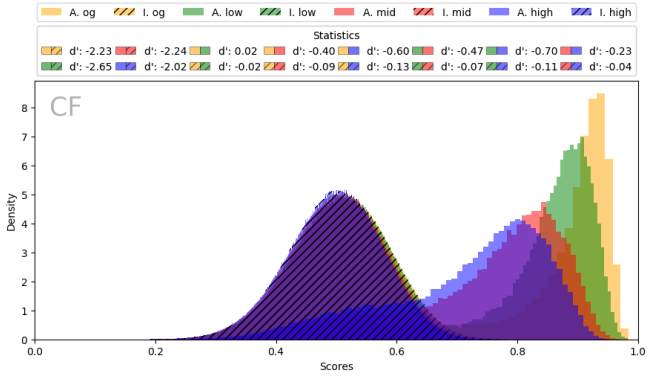
(c) African-American Males — ArcFace



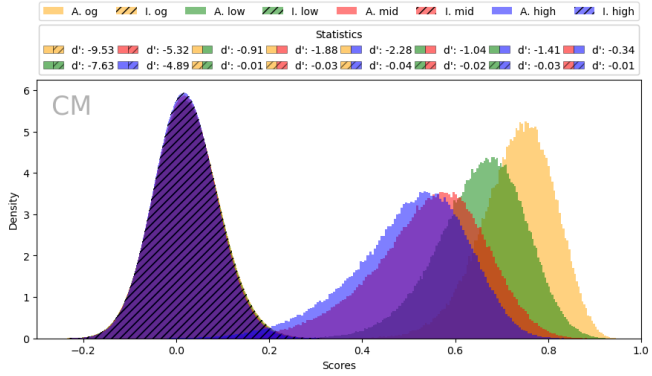
(d) African-American Males — COTS



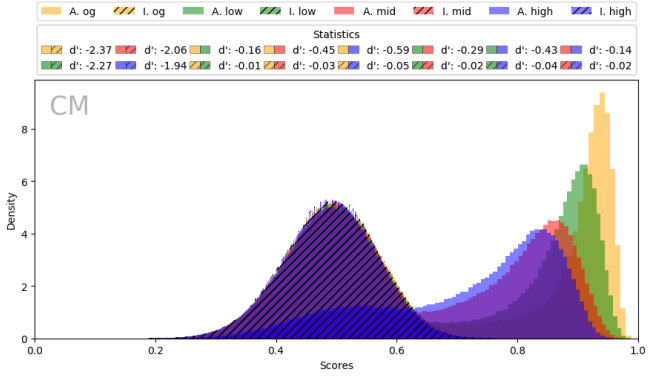
(e) Caucasian Females — ArcFace



(f) Caucasian Females — COTS



(g) Caucasian Males — ArcFace



(h) Caucasian Males — COTS

Figure 3. Authentic and Impostor Score Distributions

Table 1. Privacy Protection Measures Across Demographic Subgroups

			d-prime				Earth Mover's Distance			
			African-Americans		Caucasians		African-Americans		Caucasians	
Gallery	Probe	Matcher	Females	Males	Females	Males	Females	Males	Females	Males
Original	Low Cloaking Level	ArcFace	1.1988	**1.3703**	1.0114	0.9145	0.1007	**0.1050**	0.0945	0.0792
Original	Mid Cloaking Level	ArcFace	2.3048	**2.5258**	1.1203	1.8847	0.2305	**0.2456**	0.2242	0.1914
Original	High Cloaking Level	ArcFace	2.8202	**3.1284**	2.1010	2.7778	0.2908	**0.3071**	0.2741	0.2329
Original	Low Cloaking Level	COTS	0.1629	**0.2042**	0.0185	0.1624	0.0286	0.0353	**0.0359**	0.0275
Original	Mid Cloaking Level	COTS	0.4659	**0.6153**	0.3976	0.4460	0.0780	**0.1022**	0.0703	0.0729
Original	High Cloaking Level	COTS	0.6739	**0.8282**	0.5998	0.5862	0.1113	**0.1349**	0.0928	0.0946

Table 2. Measuring Perceived Disturbances via IQA Methods

		VIF				PieApp				DISTS			
Reference	Degraded	AAF	AAM	CF	CM	AAF	AAM	CF	CM	AAF	AAM	CF	CM
Original	Low Cloak	0.0996	0.1145	**0.0847**	0.1051	**0.1583**	0.1813	0.1777	0.1900	0.0145	0.0216	**0.0105**	0.0152
Original	Mid Cloak	0.1658	0.1925	**0.1501**	0.1872	**0.2433**	0.2784	0.2718	0.2842	0.0290	0.0419	**0.0225**	0.0315
Original	High Cloak	0.2046	0.2371	**0.1819**	0.2265	**0.2966**	0.3363	0.3311	0.3440	0.0391	0.0554	**0.0298**	0.0415

Table 3. Protection per Disturbance Point Spent

				VIF				PieApp				DISTS			
Reference	Degraded	Matcher	Protection	AAF	AAM	CF	CM	AAF	AAM	CF	CM	AAF	AAM	CF	CM
Original	Low Cloak	ArcFace	d-prime	**0.1205**	0.1197	0.1194	0.0870	**0.0757**	0.0756	0.0569	0.0481	0.8254	0.6353	**0.9625**	0.6000
Original	Mid Cloak	ArcFace	d-prime	**0.1390**	0.1312	0.0746	0.1007	**0.0947**	0.0907	0.0412	0.0663	**0.7940**	0.6035	0.4988	0.5982
Original	High Cloak	ArcFace	d-prime	**0.1378**	0.1320	0.1155	0.1226	**0.0951**	0.0930	0.0634	0.0807	**0.7207**	0.5642	0.7055	0.6688
Original	Low Cloak	ArcFace	EMD	0.0101	0.0092	**0.0111**	0.0075	**0.0064**	0.0058	0.0053	0.0042	0.0694	0.0487	**0.0899**	0.0520
Original	Mid Cloak	ArcFace	EMD	0.0139	0.0128	**0.0149**	0.0102	**0.0095**	0.0088	0.0082	0.0067	0.0794	0.0587	**0.0998**	0.0608
Original	High Cloak	ArcFace	EMD	0.0142	0.0130	**0.0151**	0.0102	**0.0098**	0.0091	0.0083	0.0068	0.0743	0.0554	**0.0920**	0.0561
Original	Low Cloak	COTS	d-prime	0.0164	**0.0178**	0.0022	0.0155	0.0103	**0.0113**	0.0010	0.0085	**0.1122**	0.0947	0.0176	0.1066
Original	Mid Cloak	COTS	d-prime	0.0281	**0.0320**	0.0265	0.0238	0.0191	**0.0221**	0.0146	0.0157	0.1605	0.1470	**0.1770**	0.1416
Original	High Cloak	COTS	d-prime	0.0329	**0.0349**	0.0330	0.0259	0.0227	**0.0246**	0.0181	0.0170	0.1722	0.1494	**0.2014**	0.1411
Original	Low Cloak	COTS	EMD	0.0029	0.0031	**0.0042**	0.0026	0.0018	0.0019	**0.0020**	0.0014	0.0197	0.0164	**0.0341**	0.0180
Original	Mid Cloak	COTS	EMD	0.0047	**0.0053**	0.0047	0.0039	0.0032	**0.0037**	0.0026	0.0026	0.0269	0.0244	**0.0313**	0.0231
Original	High Cloak	COTS	EMD	0.0054	**0.0057**	0.0051	0.0042	0.0038	**0.0040**	0.0028	0.0028	0.0284	0.0243	**0.0312**	0.0228

vor white individuals, primarily due to the face recognition frameworks they utilize.

Table 2 presents the outcomes from the IQA techniques, quantifying the perceived disturbances caused by varying cloaking levels across different cohorts. Both VIF and DISTS are congruent in indicating that the CF group experiences the least disturbance per cloaking level, while PieApp attributes this to the AAF cohort. It's noteworthy that VIF and DISTS primarily focus on analyzing structural and textural nuances. In contrast, PieApp gauges the probability of a human observer preferring one image over another—here, the cloaked versus the original.

4.2.2 The Economy of Cloaking

To evaluate the cost-effectiveness of cloaking, we establish a ratio that represents the balance between protection achieved and the visual disturbances "paid" for that protection. We use the d-prime and EMD metrics from previous sections as indicators of protection, while disturbances are gauged based on the IQA metrics. This ratio, framed as protection divided by disturbance, illustrates the level of protection obtained per unit of disturbance.

Table 3 presents these calculations, where higher values denote superior value—meaning more protection for each unit of disturbance. When examining results based on the d-prime metric, African Americans stand out as consistently achieving the best value in protection cost-effectiveness. ArcFace favors the AAF category, whereas the COTS matcher points to AAMs. In contrast, EMD-based findings are more varied, contingent upon the IQA method used. For instance, DISTS attributes that highest value to CFs, while VIF's oscillates between CFs or AAMs. PieApp leans towards AA, favoring AAF with ArcFace and

AAM with the COTS matcher. To establish a clear ranking, we count how frequently each demographic subgroup attains a specific placement (first, second, third, and fourth) per matcher. The summary of this can be found in Tab. 4.

Noticeably, both matchers consistently chose the same racial demographic for each placement, but differed in their gender selection. Given the marginal differences in disturbance scores across race-gender categories, we explored the statistical significance of these variations. We found no meaningful disparity between race-gender groups (e.g., AAF, AAM) in terms of disturbances. However, there were statistical significant differences when comparing broader racial groups (e.g., African Americans vs. Caucasians). Consequently, for our cost-effectiveness reporting, we focus on racial groups rather than the combined race-gender categories. Overall, African Americans "paid less" in terms of visual disturbances for their degree of protection than Caucasians, suggesting that African Americans tend to get "the most bang for their buck."

Table 4. Tally of Cost-Effectiveness Placement Records

		African-Americans		Caucasians	
Ranking	Matcher	Females	Males	Females	Males
1st	ArcFace	**11**	0	7	0
2nd		7	**10**	1	0
3rd		0	4	5	**9**
4th		0	4	**5**	9
1st	COTS	1	**10**	7	0
2nd		**14**	2	1	1
3rd		3	5	**6**	4
4th		0	1	4	**13**

5. Conclusions

Face recognition technologies offer undeniable benefits but come with significant challenges in differential outcomes and potential breaches of individual privacy. As we entrust machines with the task of recognizing and categorizing human faces, these algorithms inherit not only our ability to discern but also our shortcomings and predispositions. The technology is capable of both streamlining daily operations and perpetuating systemic disparities. For those deeply concerned about their privacy, these technologies pose a twofold dilemma: the risk of misidentification, and the unsettling prospect of being easily identified and tracked. The lack of clarity surrounding the use of one's facial data further intensifies these concerns, prompting many to seek proactive defensive measures.

In this context, cloaking has emerged as a privacy-preserving technique that introduces pixel-level perturbations to facial images. These alterations reduce the accuracy of recognition software while maintaining visual fidelity for human viewers. This work emphasizes the relevance of cloaking, particularly in light of its ability to empower demographic groups more vulnerable to negative impacts from recognition algorithms. Our results, when using the Fawkes algorithm, reveal that, in terms of protection effectiveness, African American males "wore it best" and benefited the most from the cloak, followed by African American females. Additionally, in terms of cost-effectiveness—balancing protection against the introduction of visual disturbances—African Americans, as a broader racial group, appeared to acquire more protection for fewer visual disruptions when compared to their Caucasian counterparts, and hence "paid less."

The use of cloaking techniques like Fawkes reinforce individual privacy and offer a degree of protection against the unauthorized use of facial data. For those deeply conscious of their digital privacy, such techniques not only provide a shield against unsanctioned face recognition but also serve as a powerful statement, highlighting the need for fair and equitable outcomes in face recognition. Nevertheless, it is essential to acknowledge that cloaking's protection is not perpetual. As face recognition algorithms evolve, they may grow more resilient against techniques. As a result, privacy-minded individuals should still exercise discretion when sharing their data online.

6. Future Work

This work presented a comparative analysis of the privacy-preserving protection provided by Fawkes across various demographics. Other existing algorithms were not considered due to their potential to introduce undesirable changes in facial structure. However, it is essential to undertake efforts to evaluate the privacy protection offered by these alternative methods across diverse demographics, as the results of those works predominantly feature individuals with lighter skin tones. Furthermore, considering the potential for significant intraclass variations within self-reported race categories [5, 19], and the ambiguity surrounding race classifications, particularly for mixed-race individuals, it is advisable to explore other factors, such as skin tone. A more granular study on the effectiveness of cloaking algorithms based on assessed skin tones can provide insights into the nuances of privacy preservation in a more inclusive manner.

References

[1] Andrea Atzori, Gianni Fenu, and Mirko Marras. Demographic Bias in Low-Resolution Deep Face Recognition in the Wild. *IEEE Journal of Selected Topics in Signal Processing*, 17(3):599–611, May 2023. Conference Name: IEEE Journal of Selected Topics in Signal Processing. 2

[2] Ziga Babnik and Vitomir Štruc. Assessing Bias in Face Image Quality Assessment. In *2022 30th European Signal*

Processing Conference (EUSIPCO), pages 1037–1041, Aug. 2022. ISSN: 2076-1465. 5

[3] Giordano Benitez Torres and Michael C. King. Harvesting faces from social media photos for biometric analysis. In *2020 IEEE International Symposium on Technology and Society (ISTAS)*, pages 230–234, 2020. 1

[4] Aman Bhatta, Vitor Albiero, Kevin W. Bowyer, and Michael C. King. The Gender Gap in Face Recognition Accuracy Is a Hairy Problem. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–10, Waikoloa, HI, USA, Jan. 2023. IEEE. 2, 3

[5] Joan Ferrante Brown, Prince. Classifying People by Race. In *Race And Ethnic Conflict*. Routledge, 1994. Num Pages: 10. 3, 8

[6] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, Jan. 2018. ISSN: 2640-3498. 2, 3

[7] S. Butt, H. Butt, and D. Gnanappiragasam. Unintentional consequences of artificial intelligence in dermatology for patients with skin of colour. *Clinical and Experimental Dermatology*, 46(7):1333–1334, 10 2021. 2

[8] Laura Rodríguez Carlos-Roca, Isabelle Hupont Torres, and Carles Fernández Tena. Facial recognition application for border control. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, July 2018. 1

[9] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):101–111, Jan. 2021. Conference Name: IEEE Transactions on Biometrics, Behavior, and Identity Science. 2, 3

[10] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition, Jan. 2021. 1

[11] Patrick Chiroro and Tim Valentine. An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48(4):879–894, 1995. 2

[12] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, Oct. 2018. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. 4

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2020. 4

[15] Steven Fraser and Dennis Mancl. Dimensions of Diversity, Equity, and Inclusion. *SIGSOFT Softw. Eng. Notes*, 48(2):18–21, Apr. 2023. 3

[16] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and Computation Redistribution for Efficient Face Detection, May 2021. arXiv:2105.04714 [cs]. 4

[17] Jan B. Deregowski Hadyn D. Ellis and John W. Shepherd. Description of white and black faces by white and black subjects. *International Journal of Psychology*, 10(2):119–123, 1975. 2

[18] Joseph F. Hair and Marko Sarstedt. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *Journal of Marketing Theory and Practice*, 29(1):65–77, Jan. 2021. 1

[19] Janet E. Helms, Maryam Jernigan, and Jackquelyn Mascher. The meaning of race in psychology and how to change it: A methodological perspective. *American Psychologist*, 60(1):27–36, 2005. Place: US Publisher: American Psychological Association. 3, 8

[20] Kashmir Hill. Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. *The New York Times*, Dec. 2020. 2

[21] Kashmir Hill. Eight Months Pregnant and Arrested After False Facial Recognition Match. *The New York Times*, Aug. 2023. 1, 2

[22] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, May 2019. Association for Computing Machinery. 3

[23] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, Lecture Notes in Computer Science, pages 565–578, Cham, 2019. Springer International Publishing. 3

[24] Danny Janssen and Simon Carton. An exploration of the suitability of Fawkes for practical applications. page 12. 3

[25] Thaddeus L. Johnson, Natasha N. Johnson, Denise McCurdy, and Michael S. Olajide. Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly*, 39(4):101753, Oct. 2022. 2

[26] Ashraf Khalil, Soha Glal Ahmed, Asad Masood Khattak, and Nabeel Al-Qirim. Investigating Bias in Facial Analysis Systems: A Systematic Review. *IEEE Access*, 8:130751–130761, 2020. Conference Name: IEEE Access. 2, 3

[27] David Leslie. Understanding Bias in Facial Recognition Technologies. *SSRN Journal*, 2020. 2

[28] Sarah Lewis. The Racial Bias Built Into Photography. *The New York Times*, Apr. 2019. 2

[29] Tao Li and Lei Lin. AnonymousNet: Natural Face De-Identification With Measurable Privacy. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 56–65, June 2019. ISSN: 2160-7516. 3

[30] Saibal Manna, Sushil Ghildiyal, and Kishankumar Bhimani. Face Recognition from Video using Deep Learning. In *2020*

5th International Conference on Communication and Electronics Systems (ICCES), pages 1101–1106, June 2020. 1

[31] Kirsten Martin. Ethics of Data and Anaytics: Concepts and Cases. In *Ethics of Data and Analytics: Concepts and Cases*. CRC Press, May 2022. Google-Books-ID: E51kEAAAQBAJ. 2

[32] Jinesh Mehta, Eshaan Ramnani, and Sanjay Singh. Face detection and tagging using deep learning. In *2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP)*, pages 1–6. IEEE, 2018. 1

[33] Muncipality of Anchorage, Alaska. Work Session FRT Fact Sheet v3. 2

[34] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini K. Ratha. In-Group Bias in Deep Learning-Based Face Recognition Models Due to Ethnicity and Age. *IEEE Transactions on Technology and Society*, 4(1):54–67, Mar. 2023. Conference Name: IEEE Transactions on Technology and Society. 2

[35] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1–14:11, Feb. 2011. 1, 2

[36] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual Image-Error Assessment through Pairwise Preference, June 2018. arXiv:1806.02067 [cs] version: 1. 4

[37] Reza Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–2, 2017. 4

[38] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006. 3

[39] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face Recognition: Too Bias, or Not Too Bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, June 2020. ISSN: 2160-7516. 2, 3

[40] Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing Biases and Preserving Privacy on Balanced Faces in the Wild. *IEEE Transactions on Image Processing*, 32:4365–4377, 2023. Conference Name: IEEE Transactions on Image Processing. 3

[41] Adam Rose. Are Face-Detection Cameras Racist? *Time*, Jan. 2010. 2

[42] Lorna Roth. Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication*, 34(1):111–136, Mar. 2009. 2

[43] Jaron Schneider. Clearview ai has scraped more than 30 billion photos from social media, Mar 2023. 1

[44] Kalpana Seshadrinathan, Thrasyvoulos N. Pappas, Robert J. Safranek, Junqing Chen, Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik. Chapter 21 - Image Quality Assessment. In Al Bovik, editor, *The Essential Guide to Image Processing*, pages 553–595. Academic Press, Boston, Jan. 2009. 4

[45] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. page 17. 1, 3

[46] SthPhoenix. InsightFace-REST, Nov. 2023. original-date: 2019-08-15T14:55:43Z. 4

[47] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, Mar. 2022. Conference Name: IEEE Transactions on Technology and Society. 2, 3

[48] Ian M. Thornton, Duangkamol Srismith, Matt Oxner, and William G. Hayward. Other-race faces are given more weight than own-race faces when assessing the composition of crowds. *Vision Research*, 157:159–168, Apr. 2019. 2

[49] Nitasha Tiku, Kevin Schaul, and Szu Yu Chen. How AI is crafting a world where our worst stereotypes are realized. *Washington Post*, Nov. 2023. 1

[50] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. IdentityDP: Differential private identification protection for face images. *Neurocomputing*, 501:197–211, Aug. 2022. 3

[51] Hanyu Xue, Bo Liu, Xin Yuan, Ming Ding, and Tianqing Zhu. Face image de-identification by feature space adversarial perturbation. *Concurrency and Computation*, 35(5):e7554, Feb. 2023. 3

[52] Gozde Yolcu, Ismail Oztel, Serap Kazan, Cemil Oz, and Filiz Bunyak. Deep learning-based face analysis system for monitoring customer interest. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):237–248, Jan. 2020. 1