

Face the Facts: Using Face Averaging to Visualize Gender-by-Race Bias in Facial Analysis Algorithms

Kentrell Owens^{*1}, Erin Freiburger^{*2}, Ryan Hutchings^{*2}, Mattea Sim², Kurt Hugenberg², Franziska Roesner¹, Tadayoshi Kohno¹

¹Paul G. Allen School of Computer Science & Engineering, University of Washington

²Psychological and Brain Sciences, Indiana University Bloomington

kentrell@cs.washington.edu, efreibur@iu.edu, ryjhutch@iu.edu, matsim@iu.edu, khugenb@iu.edu, franzi@cs.washington.edu, yoshi@cs.washington.edu

Abstract

We applied techniques from psychology — typically used to visualize human bias — to facial analysis systems, providing novel approaches for diagnosing and communicating algorithmic bias. First, we aggregated a diverse corpus of human facial images (N=1492) with self-identified gender and race. We tested four automated gender recognition (AGR) systems and found that some exhibited intersectional gender-by-race biases. Employing a technique developed by psychologists — face averaging — we created composite images to visualize these systems’ outputs. For example, we visualized what an “average woman” looks like, according to a system’s output. Second, we conducted two online experiments wherein participants judged the bias of hypothetical AGR systems. The first experiment involved participants (N=228) from a convenience sample. When depicting the same results in different formats, facial visualizations communicated bias to the same magnitude as statistics. In the second experiment with only Black participants (N=223), facial visualizations communicated bias significantly more than statistics, suggesting that face averages are meaningful for communicating algorithmic bias.

Introduction

Concerns about the social impact of artificial intelligence have risen in recent years (Tyson and Kikuchi 2023), motivating a greater focus on identifying biases within existing algorithms, particularly concerning facial analysis algorithms. For example, foundational 2018 “Gender Shades” work by Buolamwini and Gebru found that commercial automated gender recognition (AGR) systems showed higher misclassification rates for darker-skinned people, and particularly darker-skinned women (Buolamwini and Gebru 2018). To understand algorithmic biases in facial analysis systems, researchers have traditionally measured classification error rates across different demographic groups, with a focus on legally-protected categories such as gender or race. In this work, we applied social psychological research methods to build on prior work and **adopt a novel approach for visualizing algorithmic biases** in automated facial analysis systems.

^{*}These authors contributed equally.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.





	Labeled Correctly	Labeled Incorrectly
Black Women		
Black Men		

Figure 1: A table of composite images (i.e., face averages) that we created based on the output of the Kairos model. This figure was presented to participants in Studies 2 and 3. Qualitatively, one can observe that the Black women labeled correctly and the Black men labeled incorrectly are lighter-skinned than the other two groups—potentially indicating a bias towards classifying darker-skinned people as men.

Our approach was a result of a **collaboration between four social psychologists and three computer scientists**.

First, we aggregated a diverse corpus of high quality face images with self-identified gender and race information from prior psychology research. Second, mirroring the Gender Shades work, we ran these images through four public-source/access systems’ AGR algorithms (Deepface, Face++, Kairos, and Rekognition). We aimed to determine if facial analysis systems studied in previous work still exhibited classification biases. We then leveraged a visualization technique from computer science and psychology research known as *face averaging* (Sutherland, Rhodes, and Young 2017; Oldmeadow, Sutherland, and Young 2013), which social psychologists have used to understand humans’ biases when viewing faces. Here, we used this technique to understand algorithms’ biases in AGR. This technique, alongside our dataset of face images, allowed us to generate intersectional, *composite* images based on gender, race, and the models’ classification output.

On examination, one can qualitatively observe that the composite images generated from our approach (shown in Figure 1) reveal information about bias across gender and skin tone that might not otherwise be apparent to users and developers of facial analysis systems who focus solely on accuracy rates. For example, we perceived that most systems are biased to associate lighter skin tone with women and darker skin tone with men. To assess whether users could detect the systems’ biases from these visualizations, we ran a separate study in which we surveyed participants to determine how these composite images influenced their perceptions of bias and their acceptance of the use of AGR in different applications.

In this work, we asked the following research questions:

1. **Despite increasing social concern and research on intersectional biases in AGR, do previously-observed classification biases still persist? (Study 1)** We assessed present disparities in AGR accuracy across diverse face images for four popular facial analysis models. We found that intersectional disparities in AGR accuracy persist in popular facial analysis models, with a bias toward misclassifying women, and particularly Black women.
2. **What do we observe when we apply face averaging—a technique developed in social psychology—to AGR? (Study 1)** We employed face averaging to capture the features that relate to correct and incorrect gender classifications for many gender-by-race groups. Some of the visual biases observed, at least qualitatively, reify prior findings (Buolamwini and Gebru 2018) on intersectional disparities in AGR (e.g., darker skin tone more evident in Black women incorrectly labeled as men).
3. **Do the average facial depictions meaningfully communicate biases in automated facial analysis to lay audiences? (Studies 2 and 3)** We assessed whether viewing facial composite images, relative to statistics, better communicated bias in AGR classifications in two samples. We found that participants in a convenience sample (Study 2) rated similar levels of bias in both scenarios. However, we found in a sample of Black participants that facial composite images better communicated bias than traditional methods did (Study 3).

While our work uses the binary labels of “man” and “woman” because 1) the datasets with self-identified gender and race only included binary labels and 2) the algorithms we analyzed only output binary labels, we acknowledge that gender is not a binary and that seeking to remedy a system that only outputs binary gender is inherently flawed. As Keyes puts it, “a trans-inclusive system for non-consensually defining someone’s gender is a contradiction” (Keyes 2018). In this work we do not seek to improve AGR; we simply focus on a new technique to understand whether and how AGR systems are biased.

Study 1 addressed research questions 1 and 2. Specifically, we assessed current intersectional disparities in AGR across four popular facial analysis models. In contrast to prior work (Buolamwini and Gebru 2018; Ramachandran and Rattani 2022; Gustafson et al. 2023), we used faces of people who identified their own race and gender. Images were collected from prior psychological research, offering

greater control than naturalistic images (e.g., controlling for emotional expression). We then employed face averaging to visually represent the gender (mis)classifications. Overall, Study 1 allowed us to see what intersectional biases currently exist and what those biases look like in facial representations.

Studies 2 and 3 addressed research question 3. Historically, numerical data (i.e., the number of correct and incorrect gender classifications) have been used to communicate biases in facial analysis models. In Study 2 (convenience sample that was predominately white) and Study 3 (Black participant sample), we examined whether viewing AGR (mis)classifications via averaged facial representations, compared to numerical data, would lead participants to rate a facial analysis model as more biased and less appropriate for real-world applications (e.g., airport security screening). Study 3 specifically allowed us to observe whether a group historically discriminated by AGR systems (i.e., Black Americans) would be particularly sensitive to visual displays of bias as compared to numeric representations.

Background & Related Work

In this section we first present prior work on intersectional bias in facial analysis, visualizing algorithmic bias, and critiques of AGR systems. We then highlight relevant psychology literature on manipulating facial images to study humans.

Intersectional Bias in Facial Analysis Algorithms

Although there is a significant literature on gender bias in facial analysis systems (Raji and Buolamwini 2019; Feng and Shah 2022; Domnich and Anbarjafari 2021; Oh et al. 2020; Wu et al. 2020; Schwemmer et al. 2020; Manresa-Yee, Ramis Guarinos, and Buades Rubio 2022; Dominguez-Catena, Paternain, and Galar 2022; Ramachandran and Rattani 2022; Khan et al. 2011; Lin et al. 2016; Serna et al. 2021), the one most relevant to our approach is Buolamwini and Gebru’s “Gender Shades” paper (Buolamwini and Gebru 2018). This work leveraged the expertise of a professional dermatologist and had three major contributions. First, they analyzed two existing facial benchmark datasets and found that they were overwhelming comprised of lighter skinned faces. Secondly, they created a dataset (Pilot Parliament Benchmark) of face images that were balanced across skin tone. They avoided using race because it can be viewed as “an unstable social construct that changes based on geography and national norms for ethnic enumerations” (League 2020). Lastly, they used their dataset to evaluate the AGR algorithms of three commercial systems and found that darker-skinned women were disproportionately misclassified. We took a similar approach in Study 1 in that we aggregated our own dataset (see Table 1) and used it as input into AGR systems. Our approach differs from theirs in that we used *self-identified* race and gender, rather than skin tone and *perceived* gender. We acknowledge the multiple benefits of relying on skin tone (an objectively measurable feature); at the same time, there are also meaningful benefits of relying

on self-reported race (a culturally bound construct) to construct a dataset. First, racial phenotypicity is constituted by more than skin tone. Put simply, racial groups' faces vary in manifold ways, on average, not just skin tone. Our method allows us to capture those differences as well. Second, by relying on self-reported gender (rather than apparent gender), we can ensure we are capturing actual identity membership, rather than researchers' beliefs about others. This ensures any potential harms detected actually accrue to the groups measured.

Visualizing Algorithmic Bias/Fairness

We are, to our knowledge, the first work to use face averaging to study how visualizations impact perceptions of bias in AGR (not the impact of visualizations more generally, as done in (Yu et al. 2020)). However, there is relevant work in the HCI and Explainable AI (XAI) research communities studying people's understandings of algorithmic bias/fairness (Gaba et al. 2023). Although not focused on facial analysis technologies, Szymanski et al. found that while textual explanations for algorithmic decision-making are easiest to interpret for laypeople and experts, laypeople *prefer* visual aids despite the fact that they are more likely to misinterpret them (Szymanski, Millicamp, and Verbert 2021). Munechika et al. (Munechika et al. 2022) developed *Visual Auditor* to visualize under-performing subsets of data in a variety of machine learning domains; this tool represents these data as clusters plotted on a diagram displaying the intersections of features of interest. We similarly focus on under-performing subsets by averaging faces that were classified incorrectly along the feature of interest (i.e., gender). However, we anticipate that the effectiveness of face averaging may depend on people's ability and motivation to detect (often) subtle differences between faces (Hugenberg et al. 2010). Thus, the algorithmic bias depicted in facial averages may be most apparent to individuals who have meaningful previous experience viewing and remembering faces similar to the disproportionately-misclassified group.

Critical Examination of Automated Gender Recognition

Os Keyes conducted a content analysis of papers on AGR and HCI papers to understand how they used gender and what assumptions they made about gender (Keyes 2018). They found that AGR research fundamentally erases the existence of transgender people, and HCI researchers have used AGR to attempt to infer the gender of individuals for whom they did not have self-identified gender. Moreover, AGR research treated gender as a binary 95% of time. We took recommendations put forward by Keyes' work, including relying on *self-disclosed* gender information.

Facial Image Manipulation in Psychology

Sutherland et al. (Sutherland, Rhodes, and Young 2017) provides a thorough survey of facial image manipulation techniques and their applications in psychology literature. These techniques have led to key theoretical insights in social psychology in areas such as stereotyping, prejudice, and social

perception. These techniques include face averaging (averaging pixel values in facial regions defined by edges between facial landmarks), morphing, transforming, and caricaturing. We use face averaging to generate our composite images.

Methods

For research transparency, we added materials, data, and analysis scripts for all studies to the Open Science Framework.¹ In Study 1, we investigated current intersectional biases in AGR in four public-source/access facial analysis models. Prior work has shown that women with darker skin are more likely to be misclassified as men relative to women with lighter skin (Buolamwini and Gebru 2018). Given that skin-tone often covaries with race, we examined the accuracy of these systems' AGR algorithms as a function of self-reported race and ethnicity. We determine if this bias exists, first by inputting human faces (with self-identified gender and race), and then by employing face averaging to visualize those biases in faces. In Studies 2 and 3, we sought to understand if visualizing disparities in AGR accuracy in faces would meaningfully communicate bias.

Models

We selected four models to test in our experiments: Amazon Rekognition, Deepface, Face++, and Kairos. Prior work on skin tone bias in AGR included Microsoft and IBM's facial analysis models in addition to Rekognition, Face++, and Kairos (Raji and Buolamwini 2019; Buolamwini and Gebru 2018), but these models (or their AGR models) are no longer publicly available (Vincent 2022; Allyn 2020).

Ethical Considerations

We received approval from our University's IRB prior to data collection for Study 2 and 3. For all three studies, we complied with the usage policies of our image dataset sources, which were themselves designed to respect the rights and consent of the contributing participants. Our work was deemed exempt from human subjects review by our institution's Institutional Review Board, and we have complied with the usage policies of our image dataset sources.

Study 1

We procured 1492 face images of self-identified Asian, Black, Latinx, white, and Multiracial women and men displaying neutral expressions from publicly available databases (Ma, Correll, and Wittenbrink 2015; Ma, Kantner, and Wittenbrink 2021; DeBruine and Jones 2017; Chen, Norman, and Nam 2021; Minear and Park 2004; Conley et al. 2018; Righi, Peissig, and Tarr 2012); see Table 1.

Between December 2022 and May 2023, we fed these images to the four facial analysis models, which outputted an assessment of each model's AGR accuracy for each gender-by-race group. Specifically, we computed the number of correct (e.g., Black women labeled women) and incorrect (e.g., Black women labeled men) gender classifications for each intersectional group in each model.

¹ Available here: https://osf.io/jz2qb/?view_only=6bed3553f26041bfb2b3823915b7dc20

Dataset	Women/Men				
	Asian	Black	Latinx	Multiracial	White
Chicago Face Database (Ma, Correll, and Wittenbrink 2015)	57/52	104/93	56/52	0/0	90/93
Face Place (Righi, Peissig, and Tarr 2012)	28/15	12/11	15/2	17/5	45/39
Face Research Lab London Set (DeBruine and Jones 2017)	9/10	5/8	0/0	0/1	35/34
Lifespan Database (Minear and Park 2004)	10/35	38/22	1/3	0/0	103/87
RADIATE (Conley et al. 2018)	11/11	21/17	9/11	0/0	15/13
American Multiracial Face Database (Chen, Norman, and Nam 2021)	0/0	0/0	0/0	89/20	0/0
Chicago Face Database Multiracial (Ma, Kantner, and Wittenbrink 2021)	0/0	0/0	0/0	62/26	0/0

Table 1: A breakdown of self-identified gender & race for the images compiled from seven face datasets.

To visualize how specific models depict the faces of women and men, we relied on face averaging (Oldmeadow, Sutherland, and Young 2013; Sutherland, Rhodes, and Young 2017). For each model and gender-by-race groupings, we visualized the correctly and incorrectly-identified faces (e.g., Black women correctly identified as women by Deepface, and Black women incorrectly identified as men by Deepface). Specifically, for each model, self-reported racial/ethnic category, and gender category, we produced an aggregated visualization of the correctly and incorrectly-classified faces. Before averaging, we first identified 106 facial landmarks in each face through an auto-delineation process performed by one of the studied algorithms (i.e., Face++). Then, we aligned all the to-be-averaged faces using a Procrustes alignment algorithm (Salah, Alyüz, and Akarun 2008; Goodall 1991), which morphed each face in the group to the facial landmarks of one face. We averaged all the aligned faces to produce an average correctly and incorrectly-classified face for each model and gender-by-race grouping. We implemented these steps using the ‘webmorphR’ (DeBruine 2022) package in R (see Figure 2).²

Study 2

Study 2 (pre-registered on AsPredicted³) investigated whether the facial visualizations created in Study 1 would meaningfully communicate biases in AGR to laypeople. Historically, disparities in gender recognition have been communicated numerically (e.g., the number of correct vs. incorrect classifications). We examined whether depicting the same intersectional disparities in AGR would be perceived as more biased when presented as aggregate facial images vs. when presented numerically. Study 2 centered disparities in AGR for Black faces since 1) the gender misclassification of Black women has been focal in prior work (Buolamwini and Gebru 2018), and 2) Black women had the lowest classification accuracy rate for each model in Study 1 (see Figure 5). Moreover, we chose Kairos’ AGR model since the model’s disparities in classification accu-

racy between Black women and men was more subtle than in other models (e.g., Deepface, Face++), and we wanted to assess perceived bias in a context where significance of bias may be ambiguous. Indeed, if a model has extremely low classification accuracy (e.g., Deepface only correctly labeled 20.22% of Black women), then participants may judge the model as highly biased regardless of the presentation format.

We outlined three hypotheses in our pre-registration:

- H1: When a model’s bias is presented in the form of statistics and visualizations, people will find the model results with visualizations to be more biased.
- H2: When a model’s bias is presented in the form of statistics and visualizations, people will find usage of the model described with visualizations to be less acceptable.
- H3: Participants will select the model that displays their bias in the form of visualizations as more biased than the model that displays their bias in the form of statistics.

Participants. We determined the sample size using a power analysis for the smallest effect of interest ($d = 0.20$), which revealed that 200 participants would afford 80% power to detect an effect of $d = 0.20$ in a within-subjects t -test ($\alpha = .05$). In actuality, we collected a sample 239 participants from CloudResearch’s Connect platform. Participants completed the survey in a mean of 10.4 minutes and were paid \$1.25 USD, CloudResearch’s suggested hourly rate that is slightly higher than the federal minimum wage in the US. Following our pre-registration, we excluded 11 participants who did not respond “yes” to the question, “Did you take this study seriously?,” which was asked at the end of the survey. The final sample comprised 228 participants (126 men, 99 women, 3 did not report gender; 151 White, 23 Black or African American, 11 East Asian, 10 Hispanic/Latinx, 6 South East Asian, 4 South Asian, 1 American Indian or Alaskan, 1 Native Hawaiian or Other Pacific Islander, 17 reported multiple races/ethnicities or self-described, 4 did not report race/ethnicity; M age = 38.39, SD age = 11.34).

Materials and Procedure. Using the data from Study 1, we inputted four numbers into a table: the number of Black women correctly classified as women (164), the number of Black women misclassified as men (16), the number of Black men correctly classified as men (145), and the number of Black men misclassified as women (6). Then, we con-

²We could average up to 100 faces using Webmorph’s averaging feature. However, some samples included more than 100 faces, and thus, we separated the groups into equal sized subgroups, created subgroup averages, and averaged the subgroup averages to create one overall average.

³Pre-registration link here: <https://aspredicted.org/x4cv2.pdf>

structured facial averages (see Figure 1) using the same information as the tables; see procedure outlined in Study 1. To do so, we grouped faces of self-identified Black men and Black women based on whether they were classified correctly or incorrectly. Then, we averaged all the facial images in each group, irrespective of sample sizes, into one facial average.

Participants were presented with the numeric table (i.e., the numerical condition) and the face averages (i.e., the visualization condition) in separate randomly-ordered blocks. Both conditions were presented as the results of a new facial analysis algorithm for determining people's gender being developed by a different company (i.e., Company A and Company X, respectively). Importantly, while participants were told these results were from two different algorithms developed by different companies, the two presentation formats (numbers vs. visualizations) reflected the exact same data.

After presenting each condition, participants were asked to evaluate 1) how biased they believe the company's AGR algorithm is via a Likert scale (*Very biased* (1) to *Very unbiased* (7)), and 2) the acceptability of three real-world applications of the algorithm (targeted advertising, airport security, and identity verification) via a Likert scale (*Very unacceptable* (1) to *Very acceptable* (7)). After participants evaluated each company's algorithm separately, participants were asked to indicate which company's algorithm they believe was more biased, with three response options: Company A, Company X, or "*They are both similarly biased.*"

Study 3

Study 3 (pre-registered on AsPredicted⁴) examined the same hypotheses as Study 2 in a sample of Black participants.

Participants. In line with the power analysis from Study 2, we planned to collect data from 200 participants. In actuality, we collected a sample 237 participants from CloudResearch's Connect platform. Participants completed the survey in a mean of 10.7 minutes and were paid \$1.25 USD. Following our pre-registration, we removed data from 10 participants who did not respond "yes" to the question, "Did you take this study seriously?," which was asked at the end of the survey. In addition, we removed all data from 4 participants who did not select "Black or African American" as one of their racial/ethnic identities. The final sample comprised 223 participants (120 men, 99 women, 2 Genderqueer or Gender non-conforming, 2 self-described, and 1 did not report on gender; 208 Black or African American, 15 reported multiple races/ethnicities including Black or African American; M age = 33.31, SD age = 9.61).

Materials and Procedure. Study 3 had the same materials and procedure as Study 2, with the exception that the companies were renamed (i.e., "Company X" was renamed "Company B" due to a widely-known company being renamed "X").

⁴Pre-registration link here: <https://aspredicted.org/mj4rh.pdf>

Results

Study 1

Numeric Biases in Gender Classification Accuracy

Classification accuracy by model. The 1492 facial images were inputted into each of the four contemporary facial analysis models and we examined whether each model accurately labeled gender. Face++ did not recognize 15 facial images (13 Women, 2 Men), and Deepface did not recognize 3 facial images (2 Women, 1 Men); no other models had difficulty identifying faces in the images. To test for model performance differences, we computed a Pearson's Chi-squared test, in which we tested the observed number of correct and incorrect classifications against the expected number of correct and incorrect classifications given equivalent performance across the models. This test indicated that models significantly differed in their overall performance, $\chi^2(3) = 712.13, p < .001$. Deepface's gender classification accuracy rate was the lowest (72.53%), followed by Face++ (93.23%), Kairos (96.31%), and Rekognition (98.39%).

Classification accuracy by gender and model. We assessed whether classification accuracy differed as a function of self-reported gender for each model. To test this, we conducted another Pearson's Chi-squared test for each model indicating whether the observed number of correct and incorrect classifications depended on the self-reported gender of the images. Model performance significantly differed as a function of self-reported gender for DeepFace, $\chi^2(1) = 420.92, p < .001$, and Face++, $\chi^2(1) = 84.25, p < .001$, but did not for Kairos, $\chi^2(1) = 1.79, p = .182$, nor Rekognition, $\chi^2(1) = 2.91, p = .088$; see Figure 3.

Classification accuracy by race and model. Using a similar procedure as before, we conducted another Pearson's Chi-squared test for each model indicating whether the rate of observed correct and incorrect classifications depended on the self-reported racial identity of the facial images. Due to small cell sizes, we compute p -values via a bootstrapping procedure with 10,000 iterations. All models' performance differed as a function of self-reported racial identity: DeepFace: $\chi^2 = 68.96, p < .001$; Face++: $\chi^2 = 23.77, p < .001$; Kairos: $\chi^2 = 13.91, p = .007$; Rekognition: $\chi^2 = 9.91, p = .039$; see Figure 4.

Classification accuracy by gender, race, and model. Finally, we conducted one more Pearson's Chi-squared test separately for women and men in each algorithm, indicating whether the observed rate of correct and incorrect classifications depended on the self-reported racial identity of the facial images. Again, given small cell sizes, we computed p -values using the same bootstrapping procedure detailed above. For self-identified women, all models' performance differed as a function of self-reported racial identity: DeepFace: $\chi^2 = 93.19, p < .001$; Face++: $\chi^2 = 28.76, p < .001$; Kairos: $\chi^2 = 16.95, p = .003$; Rekognition: $\chi^2 = 11.97, p = .017$. In contrast, for self-identified men, no models' performance differed as a function of self-reported racial identity: DeepFace: $\chi^2 = 3.77, p = .430$; Kairos: $\chi^2 = 1.57, p = .814$; Rekognition: $\chi^2 = 1.59, p = .830$. We could not compute a

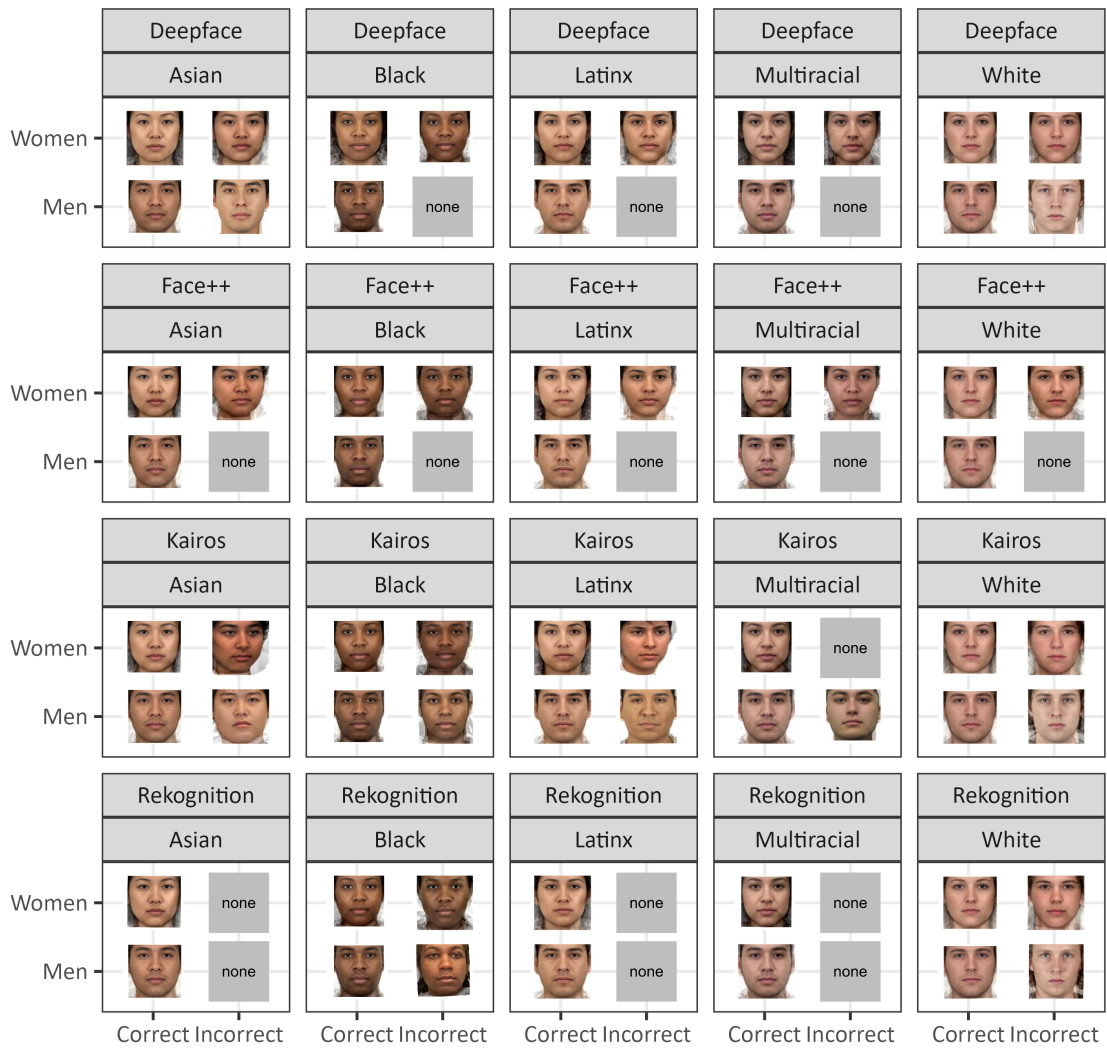


Figure 2: This figure displays the composite facial images for each self-identified gender-by-race group in each model, separated by classification accuracy

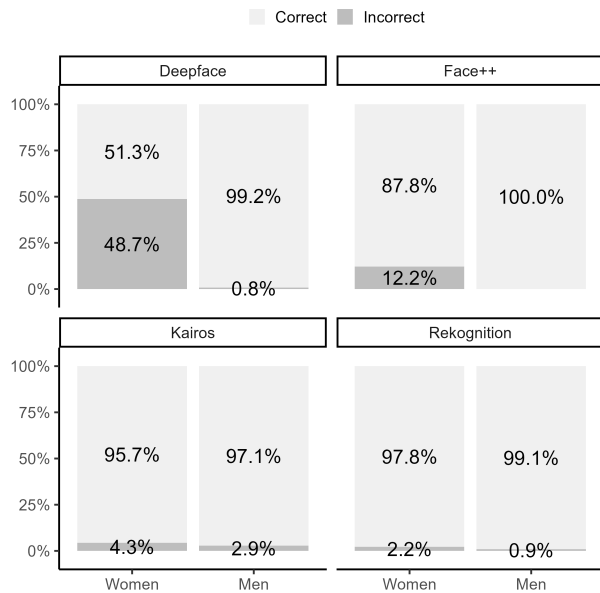


Figure 3: This figure displays the percentage of correct and incorrect gender classifications for each facial analysis model, separated by self-identified gender.

Pearson’s Chi-squared test for self-identified men in Face++ given perfect classification accuracy across all racial identities; see Figure 5.

Visualizing biases in gender classification accuracy.

Figure 2 displays the composite facial images for each self-identified gender-by-race group in each model, separated by classification accuracy. Importantly, each composite facial image depicts the facial features that underlie correct and incorrect gender classifications for a given gender-by-race group. For instance, qualitatively, it appears that darker skin tone is associated with incorrectly identifying Black women as men. Moreover, while qualitative, the facial averages also reveal features beyond skin tone that may underlie gender misclassifications, including face shape (e.g., roundness) and facial hair. We further discuss data-driven methods for diagnosing facial features underlying bias in AGR in the Discussion.

Overall, Study 1 revealed that intersectional disparities in gender classification accuracy persist in some popular facial analysis models (at the time of our latest analysis). For Deepface and Face++, gender misclassifications significantly differed by gender, with women being incorrectly classified more than men. Further, among women (but not men), classification accuracy differed by racial identity in every model. Gender misclassification rates were descriptively higher for Black women than any other gender-by-race group in all four models.

In short, we found that biases in AGR persist half a decade after the foundational findings of Buolamwini and Gebru (Buolamwini and Gebru 2018), and that these biases are visible in visual representations constructed by re-

searchers via face averaging. In Study 2, we investigate if these facial averages meaningfully communicate bias in face-based gender classification to laypeople.

Study 2

Hypothesis #1. We hypothesized that participants would find the model presented using composite images more biased relative to the model presented using numbers; we did not find support for this hypothesis. While in the predicted direction, there was no significant difference in the perceived bias of the gender classification models when results were presented as facial averages (Company X: $M = 3.87$, $SD = 1.72$) compared to when presented as numbers of (in)correct classifications (Company A: $M = 4.04$, $SD = 1.79$), $t(227) = 1.76$, $p = .080$, $d = 0.12$, 95% CI [-0.01, 0.25].

Hypothesis #2. We hypothesized that when a model’s bias was presented in the form of numbers and composite images, people would find usage of the model described with composite images to be less acceptable. This hypothesis was supported regarding targeted advertising, but not airport security or identity verification. We observed no significant difference in the perceived acceptability of using the gender classification models for airport security when participants were presented with facial averages (Company X: $M = 4.11$, $SD = 2.11$) compared to when presented with numbers of (in)correct classifications (Company A: $M = 4.19$, $SD = 2.12$), $t(227) = 0.82$, $p = .415$, $d = 0.05$, 95% CI [-0.08, 0.18]. Similarly, there was no significant difference in the perceived acceptability of model usage for identity verification when participants were presented with facial averages (Company X: $M = 4.08$, $SD = 2.13$) compared to when presented with numbers of (in)correct classifications (Company A: $M = 4.22$, $SD = 2.14$), $t(227) = 1.40$, $p = .163$, $d = 0.09$, 95% CI [-0.04, 0.22]. In contrast, participants found it more acceptable to use the gender classification model for targeted advertising when bias was presented as numbers of (in)correct classifications (Company A: $M = 4.16$, $SD = 2.00$) compared to when presented as facial averages (Company X: $M = 3.96$, $SD = 1.91$), $t(227) = 2.50$, $p = .013$, $d = 0.17$, 95% CI [0.03, 0.30].

Hypothesis #3. We hypothesized that participants would select the company that displays their bias in the form of composite images as more biased than the company that displays their bias in the form of numbers. We found that responses were not equally distributed across the options, $\chi^2(2) = 55.45$, $p < .001$. That is, 21.49% selected the model with results depicted as average facial visualizations, 21.93% selected the model with results depicted as numbers of (in)correct classifications, and 56.58% said that both models were similarly biased. However, contrary to predictions, no difference emerged between selecting the model with results as numbers relative to the model with results as visualizations, $\chi^2(1) < 0.001$, $p \approx 1$.

Overall, the results of Study 2 demonstrate that facial visualizations communicate bias to the same magnitude as numbers, suggesting that these visuals are meaningful communication tools even in contexts where bias is subtle (i.e.,

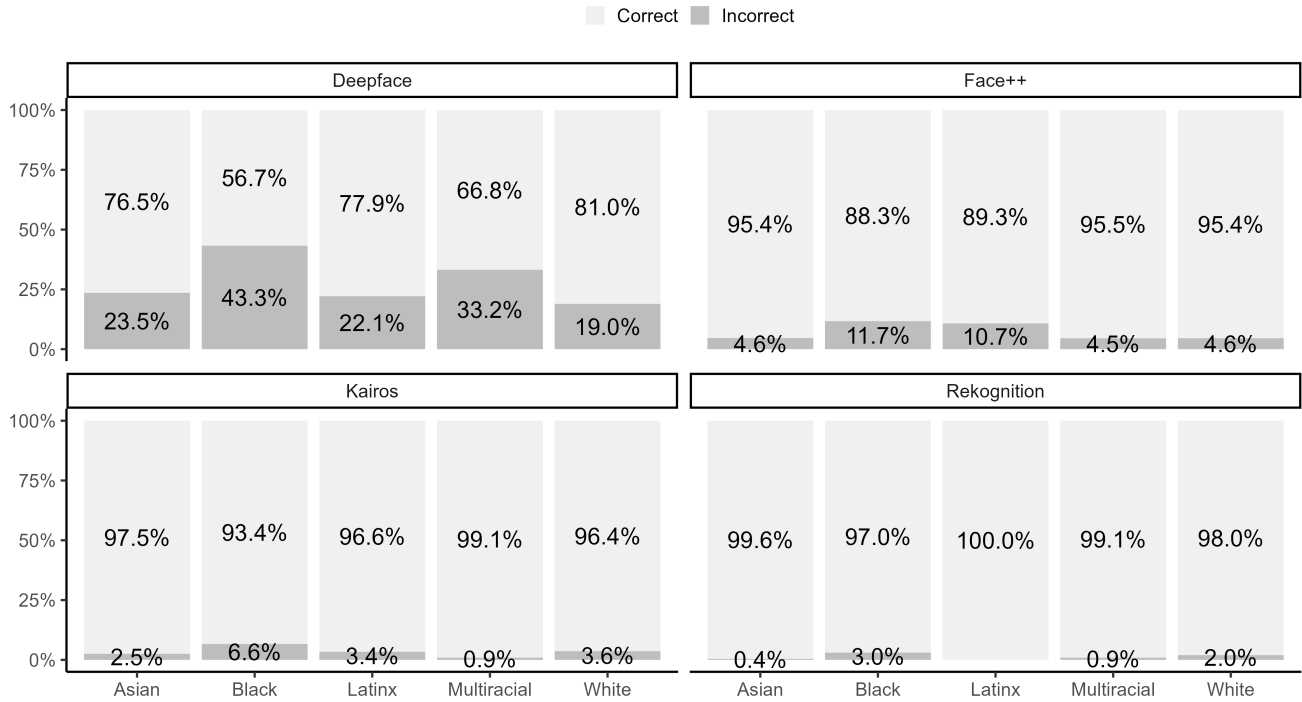


Figure 4: The percentage of correct and incorrect gender classifications for each facial analysis model, separated by self-identified race.

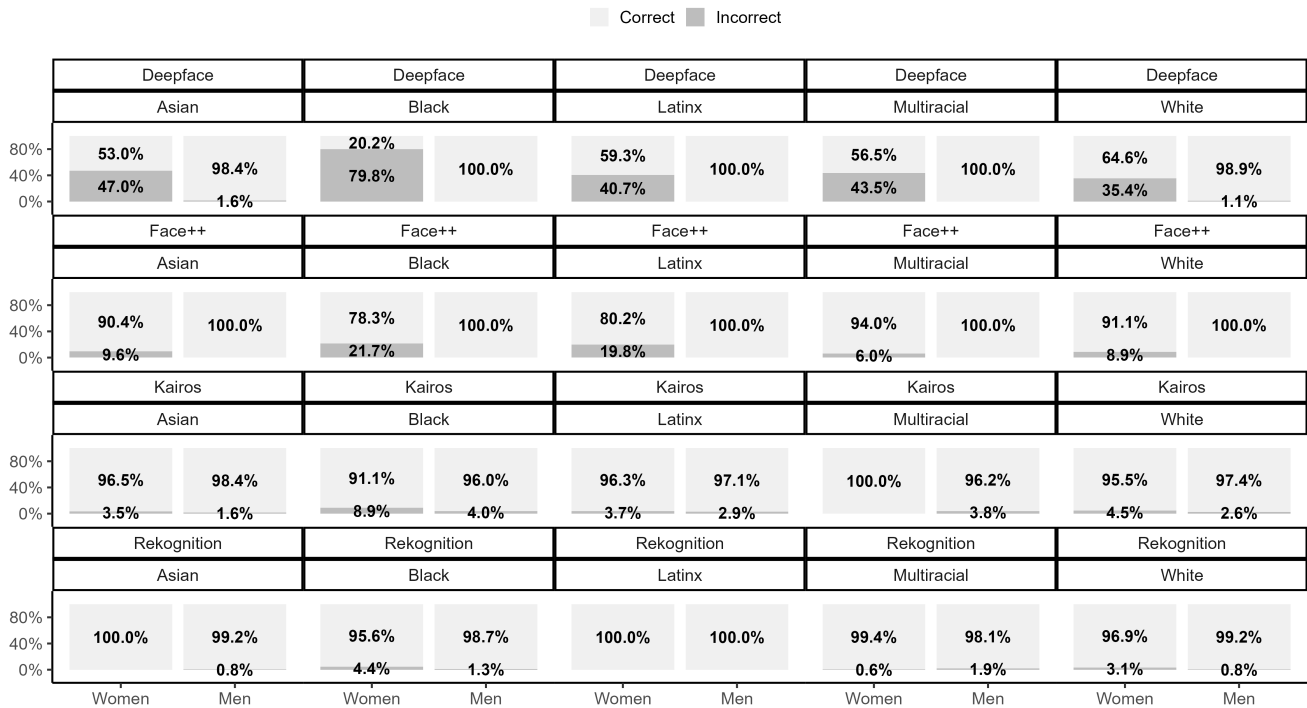


Figure 5: This figure displays the percentage of correct (e.g., Black women labeled women) and incorrect (e.g., Black women labeled men) gender classifications for each facial analysis model, separated by self-identified gender and race.

Kairos). Moreover, we find that visualizing gender classification biases may be particularly meaningful for certain applications (i.e., targeted advertising), communicating to a lay audience that AGR is not acceptable in particular domains. In Study 3, we examined Black participants' judgments, investigating whether visualizations may be a more potent bias communication tool than traditional numbers for a population disproportionately impacted by AGR bias.

Study 3

Hypothesis #1. As in Study 2, we hypothesized that participants would find the model presented using composite images more biased relative to the model presented using numbers; in contrast to Study 2, we found support for this hypothesis. Participants perceived the gender classification model as more biased when bias was presented as facial averages (Company B: $M = 3.61$, $SD = 1.60$) compared to when bias was presented as numbers of (in)correct classifications (Company A: $M = 4.09$, $SD = 1.61$), $t(222) = 4.05$, $p < .001$, $d = 0.27$, 95% CI [0.14, 0.40].

Hypothesis #2. As in Study 2, we hypothesized that when a model's bias was presented in the form of numbers and composite images, people would find usage of the model described with composite images to be less acceptable. This hypothesis was supported across use cases (i.e., targeted advertising, airport security, and identity verification). Participants found it more acceptable to use the gender classification model for airport security when bias was presented as numbers of (in)correct classifications (Company A: $M = 4.05$, $SD = 2.01$) compared to when presented as facial averages (Company B: $M = 3.57$, $SD = 1.93$), $t(222) = 4.35$, $p < .001$, $d = 0.29$, 95% CI [0.16, 0.43]. The same was true for identity verification: participants found it more acceptable to use the gender classification model when bias was presented as numbers of (in)correct classifications (Company A: $M = 4.11$, $SD = 2.04$) compared to when presented as facial averages (Company B: $M = 3.63$, $SD = 1.99$), $t(222) = 4.12$, $p < .001$, $d = 0.28$, 95% CI [0.14, 0.41]. Finally, participants found it more acceptable to use the gender classification model for targeted advertising when bias was presented as numbers of (in)correct classifications (Company A: $M = 4.29$, $SD = 1.77$) compared to when presented as facial averages (Company B: $M = 3.86$, $SD = 1.82$), $t(222) = 4.48$, $p < .001$, $d = 0.30$, 95% CI [0.17, 0.43].

Hypothesis #3. We hypothesized that participants would select the company that displays their bias in the form of composite images as more biased than the company that displays their bias in the form of numbers. We found that responses were not equally distributed across the options, $\chi^2(2) = 36.66$, $p < .001$. That is, 30.49% selected the model with results depicted as average facial visualizations, 18.39% selected the model with results depicted as numbers of (in)correct classifications, and 51.12% said that both models were similarly biased. A significant difference emerged between selecting the model with results as numbers relative to the model with results as visualizations, $\chi^2(1) = 8.21$, $p = .004$.

Discussion

In Study 1, we found that biases in automated gender classification persist in facial analysis models, and that researchers can construct visual representations of those biases through face aggregation. For instance, it is evident that darker skin tone underlies a bias to incorrectly label Black women as men even in contemporary public-source/access AGR algorithms. In Studies 2 and 3, we found that facial visualizations communicate bias just as strongly (Study 2) or more strongly (Study 3) than statistics, suggesting that these visuals meaningfully communicate algorithmic bias to laypeople. We recommend that future researchers attempting to diagnose biases use composite images alongside statistics, affording insights into *how* algorithms yield intersectional gender-by-race biases.

Limitations & Future Work

The current work sparks many questions that may motivate future research. First, Studies 2 and 3 compared people's evaluations of *either* statistical presentation of bias *or* face averaging presentation of bias. An open question for future work, therefore, is how both bias presentations jointly shape evaluations. Prior work has demonstrated that in some cases, a combination of both visuals and statistics might be useful for communicating information (Cheng et al. 2019). Conversely, other work finds that overwhelming people with information can have adverse effects (Jacoby 1984). Although not the goal of the current work, an interesting direction for future research will be to investigate how to integrate composite images with other forms of information to provide a more comprehensive understanding of an algorithm's biases. Relatedly, another question for future research is whether perceived bias from facial averages and statistics vary depending on the magnitude of the gender-by-race biases. Future work would benefit from exploring whether perception of bias vary for outputs of more inaccurate models (e.g., Deepface, Face++).

Second, we found somewhat inconsistent results in Studies 2 and 3. Although trends were often in the same direction, we identified a stronger and more reliable impact of the facial averages in Study 3, which recruited exclusively Black participants, than in Study 2, which did not restrict recruitment based on participants' racial/ethnic identity. Each experiment had unique contributions. Study 2 afforded a robust test of our hypothesis among a convenience sample of predominantly white raters that may overlook algorithmic bias (Smith et al. 2019) or lack sensitivity to subtle differences in the facial features of Black faces (Meissner and Brigham 2001) — factors that could have reduced the impact of our manipulation. Study 3 centered a sample of participants who often face algorithmic biases, and thus may have greater sensitivity to its impact, as well as a group of participants who may have more expertise in perceiving and remembering (same-race) Black faces. The study results in sum suggest that facial averages may better communicate bias to individuals who are personally targeted by the technology; however, future research should more systematically examine when and for whom facial averages relative to traditional statistics most effectively communicate algorithmic bias.

Third, the current work raises questions about how researchers can continue to leverage social psychological methods in the context of computing and bias. In particular, can psychological methods help us more deeply analyze and understand algorithmic biases? We find that face averaging techniques may be a compelling tool to communicate bias. However, face averaging and related data-driven methods (Todorov et al. 2013) may also provide tools to more deeply understand *how* an algorithm is biased. For instance, while qualitative, the face averages generated in the current work reveal facial features beyond skin tone that may underlie gender misclassifications, including face shape (e.g., roundness) and facial hair. Future work would benefit from directly investigating the features associated with correct and incorrect gender classifications. Social psychologists, for example, have leveraged synthetic faces to analyze the underlying facial components that contribute to human evaluations and biases. Applying such methods typically reserved for understanding the human mind to now understanding an algorithm’s “mind” may provide valuable insights about what contributes to race and gender biases in face classification, as well as how designers could intentionally target and alleviate such biases.

Conclusion

As long as AGR and other facial analysis systems continue to be used, researchers need to employ techniques to 1) diagnose bias in those systems and 2) communicate their findings effectively to relevant parties. In this work we find that *face averaging* — from psychology research — may be an effective means of supporting both of these tasks. More interdisciplinary work is necessary to uncover how other techniques from psychology may help researchers diagnose biased facial analysis systems, understand the underlying features contributing to decision-making, and develop novel explainability toolkits for computer vision researchers and laypeople.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation under awards #2205171 and #2207019. We thank Kaiming Cheng and Alexandra Michael for their feedback on the final manuscript.

References

Allyn, B. 2020. IBM Abandons Facial Recognition Products, Condemns Racially Biased Surveillance. *NPR*.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Chen, J. M.; Norman, J. B.; and Nam, Y. 2021. Broadening the stimulus set: Introducing the American multiracial faces database. *Behavior Research Methods*, 53: 371–389.

Cheng, H.-F.; Wang, R.; Zhang, Z.; O’Connell, F.; Gray, T.; Harper, F. M.; and Zhu, H. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Conley, M. I.; Dellarco, D. V.; Rubien-Thomas, E.; Cohen, A. O.; Cervera, A.; Tottenham, N.; and Casey, B. 2018. The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, 270: 1059–1067.

DeBruine, L. 2022. *webmorphR: Reproducible Stimuli*. R package version 0.1.1.

DeBruine, L.; and Jones, B. 2017. Face Research Lab London Set.

Dominguez-Catena, I.; Paternain, D.; and Galar, M. 2022. Gender Stereotyping Impact in Facial Expression Recognition. arXiv:2210.05332:2210.05332.

Domnich, A.; and Anbarjafari, G. 2021. Responsible AI: Gender Bias Assessment in Emotion Recognition. arXiv:2103.11436:2103.11436.

Feng, Y.; and Shah, C. 2022. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 11882–11890.

Gaba, A.; Kaufman, Z.; Cheung, J.; Shvaker, M.; Hall, K. W.; Brun, Y.; and Bearfield, C. X. 2023. My model is unfair, do people even care? visual design affects trust and perceived bias in machine learning. *IEEE Transactions on Visualization and Computer Graphics*.

Goodall, C. 1991. Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2): 285–321.

Gustafson, L.; Rolland, C.; Ravi, N.; Duval, Q.; Fu, C.-Y.; Hall, M.; Ross, C.; and Adcock, A. 2023. FACET: Fairness in Computer Vision Evaluation Benchmark. *Meta*.

Hugenberg, K.; Young, S. G.; Bernstein, M. J.; and Sacco, D. F. 2010. The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological review*, 117(4): 1168.

Jacoby, J. 1984. Perspectives on information overload. *Journal of consumer research*, 10(4): 432–435.

Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.*, 2.

Khan, S. A.; Nazir, M.; Akram, S.; and Riaz, N. 2011. Gender Classification Using Image Processing Techniques: A Survey. In *2011 IEEE 14th International Multitopic Conference*, 25–30.

League, T. A. J. 2020. Request Data Sets.

Lin, F.; Wu, Y.; Zhuang, Y.; Long, X.; and Xu, W. 2016. Human Gender Classification: A Review. *International Journal of Biometrics*, 8(3/4): 275.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47: 1122–1135.

- Ma, D. S.; Kantner, J.; and Wittenbrink, B. 2021. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53: 1289–1300.
- Manresa-Yee, C.; Ramis Guarinos, S.; and Buades Rubio, J. M. 2022. Facial Expression Recognition: Impact of Gender on Fairness and Expressions. In *XXII International Conference on Human Computer Interaction*, 1–8. ACM. ISBN 978-1-4503-9702-5.
- Meissner, C. A.; and Brigham, J. C. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1): 3.
- Minear, M.; and Park, D. C. 2004. A lifespan database of adult facial stimuli. *Behavior research methods, instruments, & computers*, 36: 630–633.
- Munehika, D.; Wang, Z. J.; Reidy, J.; Rubin, J.; Gade, K.; Kenthapadi, K.; and Chau, D. H. 2022. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*, 45–49.
- Oh, D.; Dotsch, R.; Porter, J.; and Todorov, A. 2020. Gender Biases in Impressions from Faces: Empirical Studies and Computational Models. *Journal of Experimental Psychology: General*, 149: 323–342.
- Oldmeadow, J. A.; Sutherland, C. A. M.; and Young, A. W. 2013. Facial Stereotype Visualization Through Image Averaging. *Social Psychological and Personality Science*, 4(5): 615–623.
- Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 429–435. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Ramachandran, S.; and Rattani, A. 2022. Deep Generative Views to Mitigate Gender Classification Bias Across Gender-Race Groups. arXiv:2208.08382:2208.08382.
- Righi, G.; Peissig, J. J.; and Tarr, M. J. 2012. Recognizing disguised faces. *Visual Cognition*, 20(2): 143–169.
- Salah, A. A.; Alyüz, N.; and Akarun, L. 2008. Registration of three-dimensional face scans with average face models. *Journal of Electronic Imaging*, 17(1): 011006.
- Schwemmer, C.; Knight, C.; Bello-Pardo, E. D.; Oklobdzija, S.; Schoonvelde, M.; and Lockhart, J. W. 2020. Diagnosing Gender Bias in Image Recognition Systems. *Socius*, 6: 2378023120967171.
- Serna, I.; Peña, A.; Morales, A.; and Fierrez, J. 2021. Inside-Bias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 3720–3727.
- Smith, A.; et al. 2019. More than half of US adults trust law enforcement to use facial recognition responsibly. *Pew Research Center*, 5.
- Sutherland, C. A. M.; Rhodes, G.; and Young, A. W. 2017. Facial Image Manipulation: A Tool for Investigating Social Perception. *Social Psychological and Personality Science*, 8(5): 538–551.
- Szymanski, M.; Millecamp, M.; and Verbert, K. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces, IUI '21*, 109–119. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8017-1.
- Todorov, A.; Dotsch, R.; Porter, J.; Oosterhof, N.; and Falvello, V. 2013. Validation of Data-Driven Computational Models of Social Perception of Faces. *Emotion (Washington, D.C.)*, 13.
- Tyson, A.; and Kikuchi, E. 2023. Growing public concern about the role of artificial intelligence in daily life. *Pew Research Center*.
- Vincent, J. 2022. Microsoft to retire controversial facial recognition tool that claims to identify emotion.
- Wu, W.; Protopapas, P.; Yang, Z.; and Michalatos, P. 2020. Gender Classification and Bias Mitigation in Facial Images. In *12th ACM Conference on Web Science, WebSci '20*, 106–114. Association for Computing Machinery. ISBN 978-1-4503-7989-2.
- Yu, B.; Yuan, Y.; Terveen, L.; Wu, Z. S.; Forlizzi, J.; and Zhu, H. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference, DIS '20*, 1245–1257. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369749.