COMPARING HUMAN AND MACHINE BIAS IN FACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Much recent research has uncovered and discussed serious concerns of bias in facial analysis technologies, finding performance disparities between groups of people based on perceived gender, skin type, lighting condition, etc. These audits are immensely important and successful at measuring algorithmic bias but have two major challenges: the audits (1) use facial recognition datasets which lack quality metadata, like LFW and CelebA, and (2) do not compare their observed algorithmic bias to the biases of their human alternatives. In this paper, we release improvements to the LFW and CelebA datasets which will enable future researchers to obtain measurements of algorithmic bias that are not tainted by major flaws in the dataset (e.g. identical images appearing in both the gallery and test set). We also use these new data to develop a series of challenging facial identification and verification questions that we administered to various algorithms and a large, balanced sample of human reviewers. We find that both computer models and human survey participants perform significantly better at the verification task, generally obtain lower accuracy rates on dark-skinned or female subjects for both tasks, and obtain higher accuracy rates when their demographics match that of the question. Academic models exhibit comparable levels of gender bias to humans, but are significantly more biased against darker skin types than humans.

1 INTRODUCTION

Facial analysis systems have been the topic of intense research for decades, and instantiations of their deployment have been criticized in recent years for their intrusive privacy concerns and differential treatment of various demographic groups. Companies and governments have deployed facial recognition systems (Derringer, 2019; Hartzog, 2020; Weise & Singer, 2020) which have a wide variety of applications from relatively mundane, e.g., improved search through personal photos (Google, 2021), to rather controversial, e.g., target identification in warzones (Marson & Forrest, 2021). A flashpoint issue for facial analysis systems is their potential for biased results by demographics (Garvie, 2016; Lohr, 2018; Buolamwini & Gebru, 2018; Grother et al., 2019; Dooley et al., 2021), which make facial recognition systems controversial for socially important applications, such as use in law enforcement or the criminal justice system. To make things worse, many studies of machine bias in face recognition use datasets which themselves are imbalanced or riddled with errors, resulting in inaccurate measurements of machine bias.

It is now widely accepted that computers perform as well as or better than humans on a variety of facial recognition tasks (Lu & Tang, 2015; Grother et al., 2019) in terms of *accuracy*, but what about *bias*? The algorithm's superior overall performance, as well as speed to inference, makes the use of facial recognition technologies widely appealing in many domain areas and comes at enhanced costs to those surveilled, monitored, or targeted by their use (Lewis, 2019; Kostka et al., 2021). Many previous studies which examine and critique these technologies through algorithmic audits do so only up to the point of the algorithm's biases. They stop short of comparing these biases to that of their human alternatives. In this study, we question how the bias of the algorithm compares to human bias in order to fill in one of the largest omissions in the facial recognition bias literature.

We investigate these questions by creating a dataset through extensive hand curation which improves upon previous facial recognition bias auditing datasets, using images from two common facial recognition datasets (Huang et al., 2008; Liu et al., 2015) and fixing many of the imbalances and

erroneous labels. Common academic datasets contain many flaws that make them unacceptable for this purpose. For example, they contain many duplicate image pairs that differ only in their compression scheme or cropping. As a result, it is quite common for an image to appear in both the gallery and test set when evaluating image models, which distorts accuracy statistics when evaluating on either humans or machines. Standard datasets also contain many incorrect labels and low quality images, the prevalence of which may be unequal across different demographic groups.

We also create a survey instrument that we administer to a sample of non-expert human participants ($n = 545$) and ask machine models (both through academically trained models and commercial APIs) the same survey questions. In comparing the results of these two modalities, we conclude that, first, humans and academic models both perform better on questions with male subjects. Second, humans and academic models both perform better on questions with light-skinned subjects. Third, humans perform better on questions where the subject looks like they do. Fourth, commercial APIs are phenomenally accurate at facial recognition and we could not evaluate any major disparities in their performance across racial or gender lines. Finally, overall we found that academic models exhibit comparable levels of gender bias to humans, but are significantly more biased against darker skin types than humans.

2 Background and Prior Work

We provide a brief overview of facial recognition and additional related work. We further detail similar comparative studies which contrast the performance of humans and machines. Much of the discussion of bias overlaps with the sub-field of machine learning that focuses on social and societal harms. We refer the reader to Chouldechova & Roth (2018) and Barocas et al. (2019) for additional background of that broader ecosystem and discussion around bias in machine learning.

Facial Recognition In this overview, we focus on a review of the types of facial recognition technology rather than contrasting different implementations thereof. Within facial recognition, there are two large categories of tasks: verification and identification. Verification asks a 1-to-1 question: is the person in the source image the same person as in the target image? Identification asks a 1-to-many question: given the person in the source image, where does the person appear within a gallery composed of many target identities and their associated images, if at all? Modern facial recognition algorithms, such as He et al. (2016); Chen et al. (2018); Wang et al. (2018) and Deng et al. (2019), use deep neural networks to extract feature representations of faces and then compare those to match individuals. An overview of recent research on these topics can be found in Wang & Deng (2018). Other types of facial analysis technology include face detection, gender or age estimation, and facial expression recognition.

Bias in Facial Recognition Bias has been studied in facial recognition for the past decade. Early work, like that of Klare et al. (2012) and O'Toole et al. (2012), focused on single-demographic effects (specifically, race and gender), whereas the more recent work of Buolamwini & Gebru (2018) uncovers unequal performance from an intersectional perspective, specifically between gender and skin tone. The latter work has been and continues to be hugely impactful both within academia and at the industry level. For example, the 2019 update to NIST FRVT specifically focused on demographic mistreatment from commercial platforms (Grother et al., 2019).

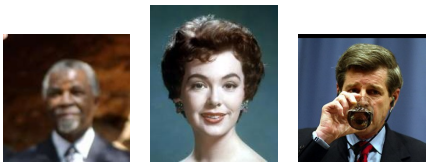
While our work focuses on the identification and comparison of bias, existing work on remedying the ills of socially impactful technology and unfair systems can be split into three (or, arguably, four (Savani et al., 2020)) focus areas: pre-, in-, and post-processing. Pre-processing work largely focuses on dataset curation and preprocessing (e.g., Feldman et al., 2015; Ryu et al., 2018; Quadrianto et al., 2019; Wang & Deng, 2020). In-processing often constrains the ML training method or optimization algorithm itself (e.g., Zafar et al., 2017a;b; Agarwal et al., 2018; Donini et al., 2018; Goel et al., 2018; Zafar et al., 2019; Diana et al., 2020; Lahoti et al., 2020; Martinez et al., 2020; Padala & Gujar, 2020; Wang & Deng, 2020), or focuses explicitly on so-called fair representation learning (e.g., Dwork et al., 2012; Zemel et al., 2013; Edwards & Storkey, 2016; Beutel et al., 2017; Madras et al., 2018; Wang et al., 2019; Adeli et al., 2021). Post-processing techniques adjust decisioning at inference time to align with fairness definitions (e.g., Hardt et al., 2016; Wang et al., 2020).

Human Performance Comparisons No work in the past to our knowledge has specifically focused on the question of comparing bias or disparity between humans and machines. Some prior work

(a) Incorrect Identities: these are labeled as the same; but the left is Zoë Lister and the right is Zoe Lister-Jones.

(b) Incorrect Labelling: this individual was labeled as not being pale skinned.

(c) Black and White Images: some identities only have black and white photos.



(d) Low-Quality Images

(e) Identical Attire/Background

(f) Multiple Distinct Faces

Figure 1: Shortcomings present in existing facial identification datasets

has looked at comparing overall performance or accuracy between the two groups. Tang & Wang (2004); O'Toole et al. (2007); Phillips & O'toole (2014) compare human and computer-based face verification performance. Lu & Tang (2015) was the first paper to show machine accuracy outpacing human accuracy. Hu et al. (2017); Phillips et al. (2018); Robertson et al. (2016) compared face recognition performance of human specific sub-populations whereas White et al. (2015) looked at comparing overall performance of humans who use the *outputs* of face recognition systems.

3 INTERRACE DATASET CURATION

We endeavor to answer two research questions: **(RQ1)** How and to what extent do humans exhibit bias in their accuracy in facial recognition tasks? **(RQ2)** How does this compare to machine learning-based models? In order to answer these questions, we created a set of challenging identification and verification questions which we posed to humans and machines from a novel dataset called InterRace for its application in intersectional facial recognition. The protocol around those experiments are described in Section 4.

To create our dataset, we first ensured that we had accurately labeled and balanced metadata. This required us to hand-check all the labels in the dataset. After removing poor quality and redundant images, we found that LFW lacked identities with dark skin tones, which is why further identities were drawn from CelebA. Though LFW does have an errata page, CelebA and other facial recognition datasets are known to have many missing or incomplete metadata, and so all CelebA images were examined by an author of this paper before adding them to the dataset. Finally, after randomly generating survey questions, we hand checked that there were no questions for which the answer is apparent or unclear for reasons other than properties of the faces (see Figure 1). In this section we detail our findings about the shortcomings in the metadata labels from LFW and CelebA and outline the steps we took to rectify and supplement these in the creation of the InterRace identities.

3.1 THE SHORTCOMINGS OF PREVIOUS DATASETS

In the process of trying to create a reasonable set of identification and verification questions, we identified that the LFW and CelebA datasets generally suffer from a range of problems that distort accuracy and bias metrics. We summarized these problems in Figure 1.

The first challenge we had to overcome is **incorrect identities**; this includes incorrect names, duplicated identities, as well as clearly incorrect matching between image and name. This problem is particularly harmful for facial recognition models which would be provided with galleries containing incorrect information about identities. In some cases, identities were split across multiple labels due to spellings. We found that this happened almost exclusively with non-canonically western names. E.g., Mesut Ozil (labelled as "Mesut Zil"), Jithan Ramesh (labelled as "Githan Ramesh"), Isha Koppikhar (labelled as "Eesha Koppikhar"), etc. Examples of incorrect identity labels include

Neela Rasgotra, a fictional character played by Parminder Singh and “All That Remains," a band name with the pictured individual being Philip Labonte. In other cases, multiple distinct identities were merged into the same label. In CelebA, Jennifer Lopez was grouped with Jennifer Driver, and Zoë Lister and Zoe Lister-Jones were both listed under “Zoe Lister" (pictured in Figure 1a).

Additionally, these datasets exhibit **metadata labelling problems** that manifest in two ways: (1) clearly defined labels being incorrectly or non-uniformly applied, and (2) vague and sometimes harmful metadata. In the first category, CelebA has features such as gender and age which often are incorrect or mislabeled (i.e. a pale-skinned person being labelled as not having pale skin, Figure 1b). Further, many categories in CelebA are subjective and/or harmful. For example, there is a label for “Attractive," “Big Nose/Lips," or “Chubby."

We found that some identities have **exclusively black and white images** (Figure 1c), making it trivial to identity two photos as being of the same label.

We filtered out **low-quality images** that could not be easily identified for reasons beyond properties of the face, such as poor light exposure, blurriness, facial obstruction, etc. We also removed “old-timey" photos that were easily associated with a specific time period, as this makes it easy to match them with other similar photos.

We found that many questions could be answered without considering face features at all, and these were removed. For example if the subject is **wearing identical attire and/or standing in front of an identical background in two images**. Many identities contained multiple images from the same red carpet event or award reception (Figure 1e). It *very* often happens that the same image appears multiple times in the dataset, but with slightly different crops, compression, or contrast adjustments.

Finally, some images **contained multiple faces**. Some of these pictures clearly have one person in the foreground and are therefore not problematic, but in others this is not the case, creating ambiguity as to which person is the target individual. See Figure 1f.

The image types above create inaccuracies when evaluating face recognition systems and distort measurements of bias when these problems occur at rates that differ across groups. For this reason, many datasets designed for training face analysis systems are not appropriate for evaluating bias.

3.2 THE INTERRACE IDENTITIES

After a thorough review of the LFW and CelebA datasets, random generation of survey questions, and rigorous hand-checking of questions to remove irregularities, we obtained a battery of survey questions for evaluating both humans and machines. We also selected survey questions that were balanced across gender, age, and skin type. Since LFW is highly skewed towards lighter identities, we included CelebA images and identities as well. We selected identities from LFW with at least two images of an individual, and then we hand labeled each identity for the following: their (1) birth date, (2) country of origin, (3) gender presentation, and (4) Fitzpatrick skin type. Labels 1-3 were assigned by an author of this paper, then that label was checked by at least two others, and modifications were made to achieve agreement among the labelers. Skin type labels (4) were assigned by 8 raters, and the mode was used as the final label.

We note that part of this work does reify categories of gender and skin type that have broader social and political implications. Further, we undertook a task of labeling and categorizing individuals who we do not know and have not received consent from for this task. Every identity for which we created these labels is indeed a celebrity in the public space with Wikipedia entries. Gender labels were rendered from the celebrity’s public comments on their own gender identity.

The **Fitzpatrick scale** (Fitzpatrick, 1988) was used to help balance the survey to include subjects with diverse skin types. This scale is widely used to classify skin complexions into 6 categories. While the Fitzpatrick scale is not perfect, it is the best systematic option currently for ensuring a broad representation.

We looked up each celebrity’s **birth date** online, mostly citing Wikipedia, and if we could not find it there, we continued to search on other websites. However, if we could still not find an individual’s date of birth, we did not list it. To find an individual’s **country of origin**, we again cited Wikipedia. If the individual came from a country that no longer existed (i.e. East and West Germany), we listed the current country. To label a person’s **gender presentation**, we took note of the person’s preferred

pronouns online and in interviews. In the event that their pronouns were not available online, we labeled their gender presentation. A major limitation of the CelebA and LFW datasets is that there were no individuals in our process who identified outside the gender binary or as gender queer.

At the end of our data collection, we collected metadata on 2545 identities which comprised a total of 7447 images. The identities themselves are rather imbalanced, though we selected a subgroup from these identities to create a balanced survey, discussed in Section 4. There are 1744 lighter-skinned individuals (as defined by Fitzpatrick skin types I-III) and 801 darker-skinned individuals (skin types IV-VI). There are 1660 males and 885 females. This sample is an improvement over previous datasets as it has been extensively evaluated to remove any errors in labeling and has a robust labeling for a wider array of skin types, unlike previous datasets which chose to label individuals as "pale." These data have a range of potential future use cases, such as being used for more evaluative facial recognition studies and commercial system audits.

4 EXPERIMENTS

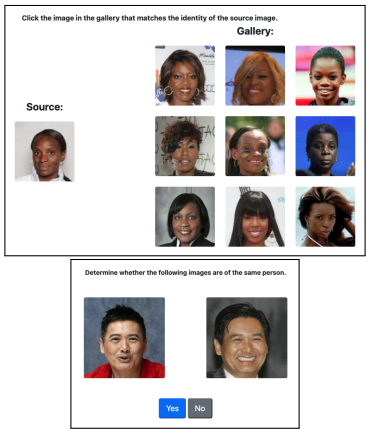


Figure 2: Example questions from the InterRace question bank. (Top) An example of an identification question. (Bottom) An example of a verification question. Notice that the demographics of all identities appearing in a question are matched to ensure the questions are not trivial.

With the high-quality metadata provided in the InterRace identities, we conduct two experiments that aim to answer our main research questions regarding the performance disparities of humans and machines. In this section, we outline how we selected the survey questions, administered the survey to human participants, and evaluated machine models. We describe the results in Section 5.

For both experiments, we create two types of questions: **identification** and **verification**. Both tasks contain a "source" image. In the identification task, 9 other images are presented in a grid, with one being of the same identity as the source and the others being of the same gender and skin type. For the verification task, a second image is selected with equal probability of being the same identity as the source image, or some other of the same gender and skin type as the source. Examples of these two types of questions can be seen in Figure 2.

We generated a static question bank with 78 identification questions and 78 verification questions for each of the 12 combinations of gender of skin type. Of those demographics with more than 78 identities, the source identity for the 78 questions were randomly chosen without replacement. This provided a total of 936 questions for each task. Finally, a pass was done over all questions to remove any for which context around the face (e.g., background or clothes) could be used to identify a person (e.g., a verification question where both images feature the same sports jersey). This final set has 901 identification and 905 verification questions.

4.1 HUMAN EXPERIMENT

We conducted an institutional review board-approved survey through the crowdsource platform Cint. The survey had two parts, one for each type of question: identification and verification.

Each respondent was asked 36 identification questions and 72 verification questions, for a target survey length of around 10 minutes. The questions for each user were randomly sampled from the total question bank such that an even distribution of questions were asked for each demographic group. As such, each respondent was asked 3 identification questions and 6 verification questions for each intersectional demographic identity. When the user first entered the survey they were prompted with a consent form. After completing both tasks, respondents filled out a demographic self-identification form which asked the participants their age range, gender, and skin type. When asking respondents

to evaluate their own Fitzpatrick skin type scale, we provided a brief description of the scale and respondents were also shown three examples of each skin type from our dataset. The entire text of the survey, including the demographic questions, can be seen in Appendix D.

Within each task, an attention check question was presented after the first five questions and before the last five. For the identification task, the attention check questions used an identical image for the target and in the gallery. For verification, one question consisted of pairing a light skinned female with a dark skin male (obvious negative example), and the other contained two identical images (obvious positive). The images used in these questions do not appear elsewhere in the survey. If a user failed to answer an attention check question correctly, they were screened out and any of their responses were ignored in our analysis. Additionally, any user who passed the attention checks but took fewer than 4 minutes to complete the survey was dropped from the final analysis. The first 3 verification and identification questions seen by each user were removed, to account for the possibility that the user may have taken some time to adjust to the format of the questions.

Our survey sampled English-speaking participants who were 18 years or older and were US residents. Our final sample includes 545 participants. There are 146 self-identified as dark-skinned (Fitzpatrick IV-VI) females, 128 light-skinned (Fitzpatrick I-III) females, 140 dark-skinned males, and 131 light-skinned males. Most respondents (375) came from the 20-39 and 40-59 age demographics.

Table 1: Demographic breakdown of human survey respondents used in final analysis.

	Fitzpatrick	Age 0-19	Age 20-39	Age 40-59	Age 60-79	Age 80+	Total
Male	I-II	0	23	37	33	2	95
	III-IV	1	35	18	24	1	79
	V-VI	4	43	33	17	0	97
Female	I-II	0	31	26	36	0	93
	III-IV	4	33	26	27	0	90
	V-VI	1	43	27	20	0	91

4.2 Machine Experiments

Academic Models To measure disparities, we trained 6 face recognition models and evaluated them on InterRace questions. We trained ResNet-18, ResNet-50 (He et al., 2016) and MobileFaceNet (Chen et al., 2018) neural networks with CosFace (Wang et al., 2018) and ArcFace (Deng et al., 2019) heads, which are designed to improve angular separation of the learned features. The models are trained using the focal loss Lin et al. (2017). For the training data, we used $140\,000$ images of 7866 CelebA identities disjoint from identities selected for the InterRace dataset. The training data has equal number of female and male identities and images. At inference time, the models solve identification questions by finding the closest gallery image in the angular feature space. For verification questions, we threshold the cosine similarity between features extracted from images in the pair.

We trained neural networks for 100 epochs with a batch size of 512 using SGD optimizer with the initial learning rate of 0.1, momentum of 0.9 and weight decay of $5\text{e}{-4}$. The learning rate was decreased by a factor of 10 at epochs 35, 65 and 95. Prior to training, we aligned and re-scaled training images to 112×112, during the training we randomly apply horizontal flip to images. For data pre-processing and training routines we adapt the code from the publicly available GitHub repository `face.evoLVe.PyTorch`.

Commercial Models We evaluated three commercial APIs: AWS Rekognition, Microsoft Azure, and Megvii Face++. We were able to evaluate face verification and identification on AWS and Azure, and only face verification on Face++. The AWS CompareFace function, which compares a source and target image, was used for both identification and verification; the target image for identification was one image comprised of the nine gallery images stitched together. Azure has native identification and verification built into their Cognitive Services Face API. Face++ has a similar set up to AWS, however they only compare the largest detected faces in the source and target images; thus we were only able to perform face verification.

4.3 Analysis Strategy

We use a two-tailed t-test with matched pairs (with a given pair corresponding to a single respondent's or computer model's scores on the two sections) to compare the accuracy rates between tasks. We also use two-tailed, unpaired t-tests to compare the overall accuracy of humans on verification questions

Table 2: Overall gender and skin type disparities exhibited by the human survey respondents, academic models, and commercial APIs.

	Identification				Verification			
	Lighter		Darker		Lighter		Darker	
	Female	Male	Female	Male	Female	Male	Female	Male
Human	67.2%	78.3%	55.5%	73.1%	78.7%	83.1%	73.4%	80.1%
Academic Models	95.0%	97.5%	89.0%	93.6%	96.2%	96.7%	92.0%	94.2%
Commercial Models	96.7%	98.7%	96.7%	97.6%	97.6%	98.9%	97.8%	99.9%

with the overall accuracy of computer models on verification questions, and the overall accuracy of humans on identification questions with the overall accuracy of computer models on identification questions. The latter t-tests and all t-tests referred to in the rest of this section are conducted on the question-level: for instance, when comparing the verification accuracy of humans and machines, we use all verification responses from all human test-takers as one sample, and all verification responses from all machines as the other.

We then analyze the disparity along gender and skin-type categories within our computer algorithms and human survey results. Users and question subjects are binned by skin type. Since the Fitzpatrick is heavily skewed towards Western conceptions of skin tone, we use two categorizations: a binary categorization of "lighter" (I-III) and "darker" (IV-VI); and categorization by (I-II), (III-IV) and (V-VI). We use two-tailed unpaired t-tests to detect the presence of accuracy disparities based on the gender or Fitzpatrick type of the identities that formed the questions. We perform tests of this kind on data from the six individual computer models, and also on the aggregate data sets of all human question responses and all computer algorithm responses.

We use logistic regression in our analysis to allow us to control for confounding variables. Results are reported as odds ratios, which compare the ratio of odds for a baseline event with the odds for a different event. We consider a main model for human subjects which predicts whether an individual question taken by a respondent was answered correctly, with independent variables as the question target gender and skin-type, and test-taker age, gender, and skin-type. The logistic regressions we run on the computer model responses are similar, but do not include test-taker demographics. We do report separate results for different architectures.

5 RESULTS

Humans achieved higher accuracy on verification (78.9%) than identification (68.3%, significant with a two-tailed matched-pair t-test with $p < 0.001$). For computer models as a whole, this gap persists but is substantially narrowed – performance on verification is 94.6%, with 93.7% on identification – and is no longer statistically significant ($p = 0.129$).

The performance difference between machines and humans is highly significant ($p < 0.001$) on both tasks using unpaired t-tests which explore group-level changes between the two tasks. Furthermore, even when controlling for demographic effects in a logistic model, humans have a much lower odds compared to computers of getting a question right (OR $= 0.14$ for verification, $p < 0.001$, OR $= 0.21$ for identification, $p < 0.001$).

Humans and Computers Perform Better on Male Subjects For identification questions, we do not observe statistically significant performance for MobileFaceNetArcFace ($p = 0.09714$ for MobileFaceNetCosFace, and $p = 0.05629$ for ResNet50CosFace), but we do observe statistically significant disparities in favor of males for each of the other three ResNet models (all $p < 0.025$). In logistic regression, we observe an odds ratio for computer models on male identification subjects of 1.89 ($p < 0.001$). Similarly, humans have significantly ($p < 0.001$) better accuracy on identification questions with male subjects: 75.7% on male subjects versus 61.4% on female subjects. The same holds true for humans on verification questions: they attain an accuracy of 81.6% on male subjects, versus 76.1% on female subjects ($p < 0.001$). Interestingly, all demographics of survey respondents (when grouped by gender and skin-type) perform substantially better on males than on females for each task. The results of the human-only logistic models confirm human biases towards male subjects in both verification (OR $= 1.39$, $p < 0.001$) and identification (OR $= 1.97$, $p < 0.001$).

Academic models are found, through logistic regression, to exhibit a statistically significant difference in performance between verification questions with male or female subjects (OR $= 1.37$, $p = 0.01$).

Humans and Computers Perform Worse on Darker-Skinned Subjects Humans collectively are proportionally 5.2% worse on dark-skinned subjects than light-skinned subjects for verification questions (80.9% versus 76.7%, $p < 0.001$) when we aggregate the Fitzpatrick scale as binary. On identification questions, this proportional difference grew to 11.7% in favor of light-skinned subjects (72.7% versus 64.2%, $p < 0.001$). This holds even when controlling for the demographics of the respondent: the odds ratio of dark-skinned compared to light-skinned question subjects for verification is 0.78 ($p < 0.001$) while for identification it is 0.67 ($p < 0.001$). When we aggregate the Fitzpatrick scale as three groups, I-II, III-VI, and V-VI, verification logistic regression finds statistically significant biases in favor of Fitzpatrick types I-II, over both III-VI and V-VI questions compared (OR $= 0.93$, $p = 0.023$ for III-VI; OR $= 0.85$, $p < 0.001$ for V-VI). For the identification task, even when controlling for respondent demographic, question subjects with Fitzpatrick values I-II have higher correct responses than that of values III-VI and V-VI (OR $= 0.92$, $p = 0.04$ for III-VI; OR $= 0.70$, $p < 0.001$ for V-VI).

On machines, we observe a similar higher performance on lighter-skinned subjects. When we aggregate the Fitzpatrick scale as just "light" and "dark", we observe a statistically significant performance disparity of 3.7% in favor of light-skinned question subjects on the verification task ($p < 0.001$), and for identification, we observe a 5.2% disparity in favor of light-skinned question subjects ($p < 0.001$). When we aggregate the Fitzpatrick scale into three categories, I-II, III-IV, and V-VI, we see a disparity for both tasks between the lightest (I-II) and darkest groups (V-VI) ($p < 0.0041$ and $p = 0.04$ for both verification and identification). Academic model performance is revealed to be significantly different, even when controlling for gender, between the types I-II and V-VI (OR $= 0.36$, $p < 0.001$ for identification; OR $= 0.47$, $p < 0.001$ for verification). However, I-II and III-VI do not show statistically significant differences for academically-trained models (OR $= 0.94$, $p = 0.714$ for identification; OR $= 0.92$, $p = 0.642$ for verification).

Human Test-Takers Perform Better on Subjects of Similar Demographic We hypothesized that humans would be more accurate on questions that contained subjects that looked like them. We find evidence to support this hypothesis in our data. On the verification task, humans perform significantly better on questions where the subjects match their gender identity (1.1%, $p = 0.02$), skin type (1.7%, $p = 0.002$), and gender identity and skin type (1.3%, $p = 0.009$). On the identification task, humans perform significantly better on questions where subjects match their skin type (2.3%, $p = 0.011$) and both their gender identity and skin type (3.3%, $p < 0.001$).

Humans and Machines Exhibit Comparable Gender Disparity, but Machines have Greater Skin Type Disparity than Humans To test for whether the levels of disparity described above are comparable between humans and machines, we look at the confidence intervals for the odds ratios of comparable models. For both tasks, recall that we observed a disparity on gender and skin type for humans and machines. For verification, we observe that the magnitude of the gender disparities are similar (OR 95% confidence intervals for humans are $[1.33, 1.46]$ and for academic models are $[1.07, 1.73]$). For identification, we observe that the magnitude of the gender disparities are also similar (OR 95% confidence intervals for humans are $[1.84, 2.10]$ and for academic models are $[1.50, 2.39]$). This allows us to conclude that when there is a gender disparity displayed by both humans and machines, the magnitudes and directions of that disparity are statistically similar.

On the other hand, the skin type disparity is more pronounced in academic models than in humans. Using the same analysis technique as above, we see that the OR 95% confidence intervals do not match for the darkest skin types in all cases (identification and verification as well as 2 and 3 skin type categories). Furthermore, the academically trained machines show a larger disparity (smaller odds ratio) than humans do. In identification, we have confidence intervals for binary skin types as $[0.62, 0.71]$ for humans and $[0.32, 0.52]$ for academic models. For tertiary skin types, we have confidence intervals of $[0.64, 0.75]$ for humans and $[0.27, 0.48]$ for the darkest subjects. In verification, we have confidence intervals for binary skin types as $[0.74, 0.82]$ for humans and $[0.38, 0.63]$ for academic models. For tertiary skin types, we have confidence intervals of $[0.80, 0.90]$ for humans and $[0.35, 0.62]$ for academic models. Regression tables can be found in Appendix C.

Commercial Facial Recognition Models Are Very Accurate The commercial models have very high accuracy, particularly AWS and Face++ which each scored above 97.3% accuracy on both verification and identification. As a result, these systems do not have enough incorrect responses to

have any statistically significant conclusions. On the other hand, Azure achieves verification accuracy of 93.3% and identification accuracy of 82.9%. In this case, we see a bias towards question gender in favor of males (OR $= 1.76$; $p = 0.041$) which is comparable to the bias observed with humans and academic models.

6 DISCUSSION

The study described in this work is the first to compare disparities and bias between humans and machines. We see that the gender and skin type biases of humans are also present in academic models. Interestingly the level of the disparities present in humans are comparable to that of the machines. These human disparities are present even when controlling for the demographics of the participant. We also find that humans perform better when the demographics of the question match their own. It might be easy to look back in hindsight on our results and say they are obvious. It may not be surprising that humans and machines have bias based on gender and skin type. However, we should not forget that this is the first study that directly compares the two with precisely the same questions which are pre-screened for difficulty. Our work is also the first to show a human preferential performance on subjects who look like them.

One key limitation of our human survey is that we analyze a crowdsourced sample. While it is demographically diverse, it does not represent a sample of expert facial recognizers. Our results should not be extrapolated too far outside the sample of non-expert crowd workers located in the US. Additionally, the results we have for the computer models are limited to those which we included and do not represent how all models work or behave.

Our findings contribute meaningfully to the ongoing work of understanding the benefits and harms presented by facial recognition technology. Specifically, we see that automated methods outperform non-expert humans across the board. When bias is detected in a machine, that bias is comparable to those exhibited by non-expert humans. In the future, further work should examine more targeted populations, such as the direct users of facial recognition technology, to understand how their native bias compares to the biases of machines or human-machine teams.

6.1 ACTIONABLE INSIGHTS

The field has focused on high-level, aggregate statistics at the dataset level for evaluating model performance. We see that in most areas, including facial recognition, where leader-boards drive innovation and improvements. However, our work puts that into perspective by examining the performance of systems when we look at accurately labeled subgroups of those overall datasets. We also see that commercial models have significantly less bias than academic models or people, and academic models are replicating and have only slightly less bias than people. The first result has never been documented before and leads one to conclude that the commercial companies have expended extra effort and resources to improve the accuracy and decrease the bias in facial identification and verification. This is likely related to incentives to minimize harm and maintain public images; but nevertheless, the different incentive has lead to improvements which academic models don't see because of the improper supremacy of the dataset-level metrics.This paper has actionable contributions to facial recognition for ML Practitioners and Dataset Curators:

ML Practitioners (1) Balanced training does not lead to unbiased performance, (2) the CosFace head and ResNet backbones yield lower bias than others, and (3) the ML solutions are always at least as biased as humans. This last point is very important. When we replace a human task with a computer task, obviously speed to do the task is important, but we also want to be better at humans in their biases. Our results, the first direct comparison of humans and machines in this domain, show that this has not been met yet. The models are always as biased, and in many cases, more biased than humans! Thus, we provide a stable benchmark ML practitioners can use to improve upon.

Dataset Curators The labels and the means of collecting those labels is very important, and should almost always involve human review. We found that existing datasets' labels (often which used computer-generated labeling methods) were widely insufficient and riddled with errors which have downstream implications on analysis.

REFERENCES

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2513–2523, 2021.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 60–69, 2018. URL http://proceedings.mlr.press/v80/agarwal18a.html.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

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*, volume 81, pp. 77–91, 2018. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pp. 428–438. Springer, 2018.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

William Derringer. A surveillance net blankets china's cities, giving police vast powers. *The New York Times*, Dec. 17 2019. URL https://www.nytimes.com/2019/12/17/technology/china-surveillance.html.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2796–2806, 2018.

Samuel Dooley, Tom Goldstein, and John P Dickerson. Robustness disparities in commercial face detection. *arXiv preprint arXiv:2108.12508*, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.05897.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

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

Clare Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11662.

Google. How google uses pattern recognition to make sense of images. https://policies.google.com/technologies/pattern-recognition?hl=en-US, 2021. Accessed: 2021-06-07.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Woodrow Hartzog. The secretive company that might end privacy as we know it. *The New York Times*, Jan. 18 2020. URL https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

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

Ying Hu, Kelsey Jackson, Amy Yates, David White, P Jonathon Phillips, and Alice J O'Toole. Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification. *Visual Cognition*, 25(4-6):492–506, 2017.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

Genia Kostka, Léa Steinacker, and Miriam Meckel. Between security and convenience: Facial recognition technology in the eyes of citizens in china, germany, the united kingdom, and the united states. *Public Understanding of Science*, pp. 09636625211001555, 2021.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Sarah Lewis. The racial bias built into photography. *The New York Times*, 25, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

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

Steve Lohr. Facial recognition is accurate, if you're a white guy. *New York Times*, 9, 2018.

Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on lfw with gaussianface. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3381–3390. PMLR, 2018. URL http://proceedings.mlr.press/v80/madras18a.html.

James Marson and Brett Forrest. Armed low-cost drones, made by turkey, reshape battlefields and geopolitics. https://www.wsj.com/articles/armed-low-cost-drones-made-by-turkey-reshape-battlefields-and-geopolitics-11622727370, Jun 2021. The Wall Street Journal.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6755–6764, 2020. URL http://proceedings.mlr.press/v119/martinez20a.html.

Alice J O'Toole, P Jonathon Phillips, Fang Jiang, Janet Ayyad, Nils Penard, and Herve Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1642–1646, 2007.

Alice J O'Toole, P Jonathon Phillips, Xiaobo An, and Joseph Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–176, 2012.

Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2277–2283. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/315. URL https://doi.org/10.24963/ijcai.2020/315.

Pew Research Center. In response to climate change, citizens in advanced economies are willing to alter how they live and work. Technical report, Pew Research Center, Washington, D.C., September 2021. URL https://www.pewresearch.org/global/wp-content/uploads/sites/2/2021/09/PG_2021.09.14_Climate_FINAL.pdf.

P Jonathon Phillips and Alice J O'toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74–85, 2014.

P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8227–8236. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00842. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Quadrianto_Discovering_Fair_Representations_in_the_Data_Domain_CVPR_2019_paper.html.

David J Robertson, Eilidh Noyes, Andrew J Dowsett, Rob Jenkins, and A Mike Burton. Face recognition by metropolitan police super-recognisers. *PloS one*, 11(2):e0150036, 2016.

Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2018.

Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Xiaoou Tang and Xiaogang Wang. Face sketch recognition. *IEEE Transactions on Circuits and Systems for video Technology*, 14(1):50–57, 2004.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9322–9331, 2020.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5310–5319, 2019.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation, 2020.

Karen Weise and Natasha Singer. Amazon pauses police use of its facial recognition software. *The New York Times*, Jul 2020. URL https://www.nytimes.com/2020/06/10/technology/amazon-facial-recognition-backlash.html.

David White, James D Dunn, Alexandra C Schmid, and Richard I Kemp. Error rates in users of automatic face recognition software. *PloS one*, 10(10):e0139827, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017a. doi: 10.1145/3038912.3052660. URL http://dx.doi.org/10.1145/3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 2017b. URL http://proceedings.mlr.press/v54/zafar17a.html.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

A How to extend InterRace

We see two main avenues for extension of the dataset; one could either further rectify existing face recognition datasets by correcting their errors, or build/augment new datasets with our labelling procedure. Both avenues for extending InterRace would rely, most likely and efficiently, on crowd workers. The protocol which we used for our labeling is fully described in our datasheet (Appendix E) to allow for easy extension and adaptation. As for extending the creation of the survey questions, this again would be very straightforward to do with a crowd, or could be done by the researcher if finer control were desired.

While our dataset was used here for one specific purpose, we know that our dataset and survey can be used for future evaluations of accuracy and bias. Furthermore, we hope our dataset curation process helps bring attention to the many pitfalls and weaknesses of large academic datasets.

B Ethics Statement

Our human subjects research was conducted in accordance with the rules, policies, and oversight of our institutional review board (IRB) which deemed our survey collection process to be Exempt. As is common practice with public figures, the data collected was done without the consent of those depicted in the images. This work contributes meaningfully by helping us better understand the

tendencies of both humans and machines in this socially important area of facial recognition. The work could potentially be used to improve facial recognition outcomes, concretize the inevitability of facial recognition technology even in morally questionable scenarios, or argue against the future development of facial recognition technologies on the basis of ongoing biases we describe.

C RESULTS TABLES

Table 3 reports the logistic regressions which depict the bias found between gender and skin type of the subject.

Table 4 reports the logistic regressions which depict the bias found between gender and skin type of the subject, even when controlling for respondent demographics.

The demographics of the subject in the question are represented with a q (`qgender` and `qskin_type`). The demographics of respondent are represented with an r (`rgender` and `rskin_type`).

Table 3: Logistic regressions for **machine** performance controlling for gender and skin types (when 2 Fitzpatrick categories are used and when 3 are used)

	Dependent variable:			
	answered_right			
	Identification	Verification	Identification	Verification
	(1)	(2)	(3)	(4)
qgenderMale	1.893	1.361	1.887	1.367
	t = 5.401***	t = 2.521**	t = 5.385***	t = 2.557**
qskin_type3III-IV	0.940	0.924		
	t = −0.367	t = −0.464		
qskin_type3V-VI	0.361	0.466		
	t = −7.114***	t = −5.109***		
qskin_type2dark			0.408	0.487
			t = −7.308***	t = −5.627***
Constant	18.060	21.435	19.513	23.231
	t = 22.649***	t = 23.013***	t = 26.913***	t = 26.993***
Observations	5,382	5,418	5,382	5,418
Log Likelihood	−1,209.843	−1,112.867	−1,218.424	−1,113.781
Akaike Inf. Crit.	2,427.686	2,233.735	2,442.848	2,233.562
Note:				*p<0.1; **p<0.05; ***p<0.01

D SURVEY TEXT

In this section, we include the text from the survey described in Section 4.1. Participants were compensated between $2.50 and $5.00.depending on whether the respondent belongs to a part of the population that is harder or easier to reach. Differential incentive amounts, standard in many survey panels (Pew Research Center, 2021), were designed to increase panel survey participation among groups that traditionally have low survey response propensities.

Landing page:

Table 4: Logistic regressions for **human** performance controlling for gender and skin types (when 2 Fitzpatrick categories are used and when 3 are used)

	Dependent variable:			
	answered_right			
	Identification	Verification	Identification	Verification
	(1)	(2)	(3)	(4)
qgenderMale	1.965 t = 20.488***	1.394 t = 13.050***	1.973 t = 20.556***	1.395 t = 13.069***
qskin_type3III-IV	0.919 t = −2.065**	0.931 t = −2.273**		
qskin_type3V-VI	0.697 t = −9.055***	0.846 t = −5.378***		
qskin_type2dark			0.667 t = −12.341***	0.779 t = −9.846***
rgenderMale	0.949 t = −1.591	0.895 t = −4.386***	0.955 t = −1.403	0.896 t = −4.325***
rskin_type3III-IV	1.078 t = 1.869*	1.104 t = 3.182***		
rskin_type3V-VI	1.215 t = 4.944***	1.128 t = 3.950***		
rskin_type2dark			1.239 t = 6.525***	1.139 t = 5.119***
Constant	1.734 t = 12.960***	3.394 t = 36.895***	1.789 t = 15.985***	3.576 t = 44.590***
Observations	17,877	37,605	17,877	37,605
Log Likelihood	−10,865.250	−19,292.960	−10,825.090	−19,254.730
Akaike Inf. Crit.	21,744.500	38,599.910	21,660.190	38,519.460

Note: *p<0.1; **p<0.05; ***p<0.01

> Welcome to this survey! It was created at the [author's identity] during the summer of 2021, made possible by the [author's support].
>
> The survey will take approximately approximately 10 minutes to finish. You will be performing two tasks, with each task taking approximately 5 minutes. After finishing the first task, you will be routed to the other task. You may take a short break in between the two tasks, but the survey is intended to be taken in one sitting. If at any time in the middle of a task you need to take a break, be sure to refresh the page.
>
> Once you feel ready, press 'Next' to get routed to your first task.

Verification instructions:

> Welcome to Task A! This task will take approximately 5 minutes. It has 74 questions. Each question will have two images, each of a single face. Your job is to identify whether the faces in these two images are of the same person or not.

> You can either click on the buttons ‘Yes’ / ‘No’ or press ‘y’ for ‘Yes’ and ‘n’ for ‘No’. After you click a button or press one of the ‘y’ or ‘n’ keys, you will not be allowed to change your answer, so keep that in mind. Please try to verify whether the two images are of the same person to the best of your ability. If at any time in the middle of a task you need to take a break, be sure to refresh the page. After finishing the last question, you will be directed to Task B. Once you feel ready, click the ‘Next’ button to start this task.

Verification task heading:

> ***Task A:*** **Determine whether the following images are of the same person.**

The “***Task A:***" is a link to a popup that displays the verification instructions again.

Identification instructions:

> Welcome to Task B! This task will take approximately 5 minutes. It has 38 questions. Each question will have ten images, each of a single face. One image will appear on the left of your screen — this is the target image. The other nine images will appear on the right of your screen in a 3-by-3 grid. Exactly one of these nine images will match the identity of the target image. Your job is to click the image in the grid that matches the target. After you click a picture in the gallery, you will not be allowed to change your answer, so keep that in mind. Please try to identify the matching image to the best of your ability. If at any time in the middle of a task you need to take a break, be sure to refresh the page. After finishing the last question, there will be a brief questionnaire asking about your individual information. Once you feel ready, click the ‘Next’ button to start this task.

Identification task heading:

> ***Task B:*** **Click the image in the gallery that matches the identity of the target image.**

Similarly, “***Task B:***" is a link to a popup that displays the verification instructions again.

Note that there is a 50-50 chance for starting on verification or identification. In this case, verification was presented first (which is referred to by “task A") and identification was presented second (which is referred to by “task B").

User information page:

> Please enter your information.
>
> All information will be kept strictly private on secure university servers and will be erased after the completion of this study.
>
> Select your age: [0-19, 20-39, 40-59, 60-79, 80+, Prefer not to say]
>
> Select your gender: [Male, Female, Other]
>
> Feel free to elaborate on your gender presentation: [text-box]
>
> Please select the category that best represents your skin tone.
>
> What are the Fitzpatrick Skin Types?
>
> [Pale White Skin, White Skin, Light Brown Skin, Moderate Brown Skin, Dark Brown Skin, Deeply Pigmented Dark Brown Skin]

What are the Fitzpatrick Skin Types? popup:

> The **Fitzpatrick Skin Phototypes** were developed by dermatologist Thomas B. Fitzpatrick. It is a system commonly used to classify skin complexions and their various reactions to exposure to ultraviolet radiation, or sun exposure. There are 6 catogeries, ranging from extremely sensitive skin which always burns instead of tannning, to very resistant skin which is deeply pigmented and almost never burns.

Thanks page:

> **Thanks for taking the [author's identity] 2021 Survey!**
> Thanks again for finishing the **[author's identity] 2021** survey! Have a great rest of your day!

E DATASHEETS FOR DATASETS

E.1 MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The main purpose for creating this dataset was to create a set of challenging face verification and face identification questions which were used in a series of experiments which compared the bias of humans in machines in these tasks.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by [author's identity].

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

N/A.

Any other comments?

No.

E.2 COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is an identity. Each identity can have one or more images picturing them.

How many instances are there in total (of each type, if appropriate)?

There are 2545 unique identities with 7447 images in total.

Gender	Skin Tone	Identities	Images
Female	1	111	236
Female	2	269	760
Female	3	150	443
Female	4	139	303
Female	5	138	301
Female	6	78	156
Male	1	126	250
Male	2	647	2284
Male	3	441	1668
Male	4	189	488
Male	5	171	382
Male	6	86	176

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset is a sample from both Labelled Faces in the Wild and CelebA. This sample is not representative of the larger set in order to cover a more diverse range of perceived gender and Fitzpatrick rating pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of one of more images of an identity.

Is there a label or target associated with each instance? If so, please provide a description.

Identity name, approximate age upon release of containing dataset, perceived gender, country of origin, and Fitzpatrick skin rating is labelled for each instance.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Not that we are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset contains references to images and identities in the Labelled Faces in the Wild and CelebA datasets, whose websites have persistent data catalogs. LFW has an errata page which indicates any errors or updates, but there have been none recently.

`http://vis-www.cs.umass.edu/lfw/`

`http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does identify subpopulations, specifically by age, perceived gender, country of origin, and Fitzpatrick skin tones. Age was calculated as the difference between the min of year of death and the release of the containing dataset with the date of birth. Perceived gender was gathered from preferred pronouns. Country of origin was obtained from Wikipedia or another celebrity information page if not available. Fitzpatrick skin tone ratings were determined by at least a 5/8 majority among the [author's identity].

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes since the names are used as the identifiers.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

This dataset does not contain sensitive information to our knowledge.

Any other comments?

No.

E.3 COLLECTION PROCESS

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The date of birth, perceived gender, and country of origin for a given instance was acquired by manually searching up the given identity's Wikipedia page or any celebrity information page. The Fitzpatrick skin ratings was obtained in a majority vote among the [author's identity].

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Manual human curation was used to collect the data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Identities were sampled from the larger LFW and CelebA dataset. Identities from LFW with more than one image were primarily taken. Identities from CelebA were sampled to improve the

underrepresented intersectional demographic identities (such as those with Fitzpatrick ratings of I or IV-VI) with a goal of bringing each intersection to above 75 identities.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

[author's identity] were involved in the data collection sample and they were compensated as a part of [author's identity].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected over a timeframe of two months.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

There was no IRB review conducted for this dataset collection.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We obtained the data from the individuals through third parties.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Individuals were not notified about the data collection.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

No.

E.4 PREPROCESSING/CLEANING/LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

N/A

Any other comments?

No.

E.5 USES

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, the data were used for face recognition bias tests in a survey project run by the dataset creators.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?

This dataset could also be used in mitigating bias in facial recognition models as a training dataset. The facial categorization task could also utilize this dataset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Facial recognition audits.

Any other comments?

No.

E.6 DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the data will be shared publicly.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

GitHub.

When will the dataset be distributed?

2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

E.7 MAINTENANCE

Who will be supporting/hosting/maintaining the dataset?

This dataset will be hosted on GitHub and the authors of this paper will continue to support the dataset, performing any necessary maintenance.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

[author's identity]

Is there an erratum? If so, please provide a link or other access point.

A list of erratum is displayed and updated in the README of the project's GitHub.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Problematic images will be addressed when brought to attention, and an amended dataset will be released through GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Images and identities were curated from other well established datasets CelebA and Labeled Faces in the Wild. Please defer to the practices followed in said parent datasets.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

All versions of the dataset will continue to be hosted on GitHub as different releases of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

GitHub pull requests.

Any other comments?

No.

F RECOGNIZABILITY

The present results do use CelebA, but that does not mean that every subject is widely recognizable to every participant. While we did not ask our participants if they specifically knew the subject in a question, we can approximate the level of recognisability of a subject. We have performed an update to account for the recognisability of the subject. To do this, we split each subject into one of three groups: "very famous", "somewhat famous", and "less famous". These three groups are equally sized. Membership in each group was determined by ordering the number of Google search results that were observed when a search for their name was performed. The cutoff for the "somewhat famous" group was 1,640,000 search results, and the cutoff for the "very famous" group was 10,700,000.

With these groups, we can see strong evidence that our main conclusions hold for the most part least and most famous groups, with only moderate differences in the conclusions drawn for those somewhat famous. This additional analysis and result should appease concerns that the conclusions from our study only consist of widely recognizable subjects.

Below, we present a re-analysis of the main body of the paper but for each of the three groups of recognizability. The summary of the results are found in Table 5.

F.1 LESS FAMOUS GROUP ANALYSIS

F.1.1 VERIFICATION IS EASIER THAN IDENTIFICATION; COMPUTERS ARE MORE ACCURATE THAN HUMANS

Humans achieved higher accuracy on verification (78.9%) than identification (68.3%, significant with a two-tailed matched-pair t-test with $p < 0.001$). For computer models as a whole, this gap persists but is substantially narrowed – performance on verification is 94.5%, with 93.7% on identification – and is no longer statistically significant ($p = 0.351$). The performance difference between machines and humans is highly significant ($p < 0.001$) on both tasks using unpaired t-tests which explore group-level changes between the two tasks.

Furthermore, even when controlling for demographic effects in a logistic model, humans have a much lower odds compared to computers of getting a question right (OR = 0.11 for verification, $p < 0.001$, OR = 0.19 for identification, $p < 0.001$).

	Identification				Verification			
	Overall	Less Famous	Somewhat Famous	Very Famous	Overall	Less Famous	Somewhat Famous	Very Famous
Computer are Better than Humans	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Humans are Better on Males	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Machines are Better on Males	Only ResNets	Only ResNets	No	Yes	Yes	Yes	No	Yes
Machines More Biased than Humans on Gender	Yes	Yes	No	No	Yes	No	No	No
Machines at least as Biased as Humans on Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Humans are Better on Lighter Skinned	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Machines are Better on Lighter Skinned	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Machines More Biased than Humans on Skin Type	Yes	Yes	No	Yes	Yes	No	No	Yes
Machines at least as Biased as Humans on Skin Type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Humans Better on Matched Gender Identity	Yes	Yes	Yes	No	Yes	No	Yes	No
Humans Better on Matched Skin Type	Yes	No	No	No	Yes	No	No	No

Table 5: A summary of the conclusions from the analysis for each Recognizability strata compared to the main conclusions from all questions ("Overall").

F.1.2 Humans and Computers Perform Better on Male Subjects

For identification questions, we do not observe statistically significant performance MobileFaceNet, ($p > 0.0788$), but we do observe statistically significant disparities in favor of males for each of the other three ResNet models (all $p < 0.003$). Similarly, in logistic regression, we observe an odds ratio for computer models on male identification subjects of 8.34 ($p < 0.001$).

Humans also have significantly ($p < 0.001$) better accuracy on identification questions with male subjects: 77.5% on male subjects versus 55.3% on female subjects. The same holds true for humans on verification questions: they attain an accuracy of 81.2% on male subjects, versus 72.8% on female subjects ($p < 0.001$). The results of the human-only logistic models confirm human biases towards male subjects in both verification (OR = 1.57, $p < 0.001$) and identification (OR = 2.64, $p < 0.001$).

Academic models are found, through logistic regression, to exhibit a statistically significant difference in performance between verification questions with male or female subjects (OR = 1.98, $p = 0.007$).

F.1.3 Humans and Computers Perform Worse on Darker-Skinned Subjects

Humans collectively perform 6.8% worse on dark-skinned subjects than dark-skinned subjects for verification questions (79.0% versus 73.6%, $p < 0.001$). On identification questions, this gap is widened to 20.0% in favor of light-skinned subjects (72.5% versus 20.4%, $p < 0.001$). Even when controlling for the demographics of the respondent, logistic regressions yield the same conclusions: odds ratio of dark-skinned compared to light-skinned question subjects for verification is 0.78 ($p < 0.001$) while for identification is 0.56 ($p < 0.001$).

When we aggregate the Fitzpatrick scale as three groups, I-II, III-VI, and V-VI, verification logistic regression finds statistically significant biases in favor of Fitzpatrick types I-II, over V-VI questions compared (OR = 0.98, $p = 0.745$ for III-VI; OR=0.76, $p < 0.001$ for V-VI). For the identification task, even when controlling for respondent demographic, question subjects with Fitzpatrick values I-II have higher correct responses than that of values V-VI (OR = 0.98, $p = 0.833$ for III-VI; OR=0.62, $p < 0.001$ for V-VI).

The results are more nuanced for machines.

When we aggregate the Fitzpatrick scale as just "light" and "dark", we observe a statistically significant performance disparity of 3.6% in favor of light-skinned question subjects on the verification

task ($p < 0.003$), and for identification, we observe a 7.4% disparity in favor of light-skinned question subjects ($p < 0.001$).

When we aggregate the Fitzpatrick scale into three categories, (I-II),(III-IV), and (V-VI), we see a disparity for both tasks between the lightest (I-II) and darkest groups (V-VI) ($p < 0.023$ and $p < 0.001$ for both verification and identification). Subject Fitzpatrick type is revealed to be significantly different, even when controlling for gender, between the types I-II and V-VI (OR = 0.33, $p < 0.001$ for identification; OR=0.56, $p = 0.040$ for verification). This means that I-II and III-VI do not show statistically significant differences for academically-trained models (OR = 1.49, $p = 0.227$ for identification; OR=0.74, $p = 0.322$ for verification).

F.1.4 Human Test-Takers Do Not Perform Better on Subjects of Similar Demographic

Humans do not perform significantly better on questions where the subjects match their gender identity , skin type , or gender identity and skin type. On the identification task, humans do not perform significantly better on questions where subjects match their skin type ($p > 0.142$) but do peform better on questions where subjects math both their gender identity and Fitzpatrick type (4.5%, $p < 0.01$).

F.1.5 Machines are at least as biased as Humans

To test for whether the levels or disparity described above are comparable between humans and machines, we look at the confidence intervals for the odds ratios of comparable models. For both tasks, recall that we observed a disparity on gender and skin type for humans and machines. For verification, we observe that the magnitude of the gender disparities are similar (OR 95% confidence intervals for humans are [1.42, 1.72] and for academic models are [1.22, 3.32]). For identification, we observe that the magnitude of the gender disparities are dissimilar (OR 95% confidence intervals for humans are [2.37, 3.03] and for academic models are [4.22, 17.03]). This allows us to conclude that when there is a gender disparity displayed by both humans and machines, the machines are at least as biased as the humans.

On the other hand, the skin type disparity is more pronounced in academic models than in humans. Using the same analysis technique as above, we see that the OR 95% confidence intervals do not match for the darkest skin types in all cases (identification and verification as well as 2 and 3 skin type categories). Furthermore, the academically trained machines show a larger disparity (smaller odds ratio) than humans do. In identification, we have confidence intervals for binary skin types as [0.50, 0.63] for humans and [0.19, 0.48] for academic models. For tertiary skin types, we have confidence intervals of [0.53, 0.71] for humans and [0.20, 0.56] for the darkest subjects. In verification, we have confidence intervals for binary skin types as [0.71, 0.85] for humans and [0.34, 0.85] for academic models. For tertiary skin types, we have confidence intervals of [0.81, 0.98] for humans and [0.32, 0.97] for academic models.

F.2 Somewhat Famous Group Analysis

F.2.1 Verification is Easier Than Identification; Computers are More Accurate Than Humans

Humans achieved higher accuracy on verification (78.9%) than identification (68.3%, significant with a two-tailed matched-pair t-test with $p < 0.001$). For computer models as a whole, this gap persists but is substantially narrowed – performance on verification is 94.6%, with $93.\%$ on identification – and is no longer statistically significant ($p = 0.836$) The performance difference between machines and humans is highly significant ($p < 0.001$) on both tasks using unpaired t-tests which explore group-level changes between the two tasks.

Furthermore, even when controlling for demographic effects in a logistic model, humans have a much lower odds compared to computers of getting a question right (OR = 0.12 for verification, $p < 0.001$, OR = 0.20 for identification, $p < 0.001$).

F.2.2 HUMANS PERFORM BETTER ON MALE SUBJECTS; BUT COMPUTERS DO NOT

For identification questions, we do not observe statistically significant performance for any of the academic models ($p > 0.145$). Similarly, in logistic regression, we observe an odds ratio for computer models on male identification subjects of 1.05 ($p = 0.808$).

Humans have significantly ($p < 0.001$) better accuracy on identification questions with male subjects: 73.7% on male subjects versus 62.7% on female subjects. The same holds true for humans on verification questions: they attain an accuracy of 79.7% on male subjects, versus 75.4% on female subjects ($p < 0.001$). The results of the human-only logistic models confirm human biases towards male subjects in both verification (OR = 1.29, $p < 0.001$) and identification (OR = 1.67, $p < 0.001$).

Academic models are found, through logistic regression, to exhibit a statistically significant difference in performance between verification questions with male or female subjects (OR = 1.93, $p = 0.747$).

F.2.3 HUMANS AND COMPUTERS PERFORM WORSE ON DARKER-SKINNED SUBJECTS

Humans collectively perform 3.2% worse on dark-skinned subjects than dark-skinned subjects for verification questions (78.9% versus 76.3%, $p < 0.001$). On identification questions, this gap is widened to 4.2% in favor of light-skinned subjects (69.5% versus 66.6%, $p < 0.023$). Even when controlling for the demographics of the respondent, logistic regressions yield the same conclusions: odds ratio of dark-skinned compared to light-skinned question subjects for verification is 0.84 ($p < 0.001$) while for identification is 0.87 ($p = 0.014$).

When we aggregate the Fitzpatrick scale as three groups, I-II, III-VI, and V-VI, verification logistic regression finds no statistically significant biases in favor of Fitzpatrick types I-II, over both III-VI and V-VI questions compared (OR = 0.94, $p = 0.297$ for III-VI; OR=1.00, $p = 0.986$ for V-VI). For the identification task, even when controlling for respondent demographic, question subjects with Fitzpatrick values I-II do not have higher correct responses than that of values III-VI and V-VI (OR = 1.03, $p = 0.686$ for III-VI; OR=.89, $p = 0.106$ for V-VI).

The results are more nuanced for machines.

When we aggregate the Fitzpatrick scale as just "light" and "dark", we observe no statistically significant performance disparity between skin types for either identification or verification. When we aggregate the Fitzpatrick scale into three categories, (I-II),(III-IV), and (V-VI), we also see no disparity for both tasks between the lightest (I-II) and darkest groups (V-VI) ($p > 0.1499$ for both verification and identification). Subject Fitzpatrick type is revealed to not be significantly different, even when controlling for gender, between the types I-II and III-IV or V-VI in either verification or identification, except III-IV is easier for the academic models compared to I-II (OR=3.26 with $p < 0.001$).

F.2.4 HUMAN TEST-TAKERS PERFORM BETTER ON SUBJECTS OF SIMILAR DEMOGRAPHIC

Humans perform significantly better on questions where the subjects match their gender identity ($p < 0.006$) and gender identity and skin type ($p < 0.001$), but not for skin type ($p > 0.138$), On the identification task, humans do not perform significantly better on questions where subjects match their skin type ($p > 0.410$) but do peform better on questions where subjects math both their gender identity and Fitzpatrick type (4.5%, $p < 0.02$).

F.2.5 HUMANS AND MACHINES EXHIBIT COMPARABLE SKIN TYPE AND GENDER PERFORMANCE

To test for whether the levels or disparity described above are comparable between humans and machines, we look at the confidence intervals for the odds ratios of comparable models. For all variables and tasks, the Odds Ratios confidence intervals overlap meaning the biases (to the extent there are any) are comparable between humans and machines.

F.3 VERY FAMOUS GROUP ANALYSIS

F.3.1 VERIFICATION IS EASIER THAN IDENTIFICATION; COMPUTERS ARE MORE ACCURATE THAN HUMANS

Humans achieved higher accuracy on verification (78.9%) than identification (68.3%, significant with a two-tailed matched-pair t-test with $p < 0.001$). For computer models as a whole, this gap persists but is substantially narrowed – performance on verification is 94.6%, with 92.7% on identification – though still statistically significant ($p = 0.009$) The performance difference between machines and humans is highly significant ($p < 0.001$) on both tasks using unpaired t-tests which explore group-level changes between the two tasks.

Furthermore, even when controlling for demographic effects in a logistic model, humans have a much lower odds compared to computers of getting a question right (OR = 0.19 for verification, $p < 0.001$, OR = 0.25 for identification, $p < 0.001$).

F.3.2 HUMANS PERFORM BETTER ON MALE SUBJECTS; BUT COMPUTERS DO NOT

For identification questions, we do not observe statistically significant performance for any of the academic models ($p > 0.379$). Similarly, in logistic regression, we observe an odds ratio for computer models on male identification subjects of 1.42 ($p > 0.079$).

Humans though have significantly ($p < 0.001$) better accuracy on identification questions with male subjects: 75.6% on male subjects versus 67.1% on female subjects. The same holds true for humans on verification questions: they attain an accuracy of 83.3% on male subjects, versus 81.5% on female subjects ($p < 0.002$). The results of the human-only logistic models confirm human biases towards male subjects in both verification (OR = 1.18, $p < 0.001$) and identification (OR = 1.59, $p < 0.001$).

Academic models are found, through logistic regression, to exhibit no statistically significant difference in performance between verification questions with male or female subjects (OR = 1.42, $p = 0.125$).

F.3.3 HUMANS AND COMPUTERS PERFORM WORSE ON DARKER-SKINNED SUBJECTS

Humans collectively perform 6.9% worse on dark-skinned subjects than dark-skinned subjects for verification questions (85.3% versus 79.4%, $p < 0.001$). On identification questions, this gap is widened to 10.4% in favor of light-skinned subjects (75.3% versus 67.4%, $p < 0.001$). Even when controlling for the demographics of the respondent, logistic regressions yield the same conclusions: odds ratio of dark-skinned compared to light-skinned question subjects for verification is 0.66 ($p < 0.001$) while for identification is 0.65 ($p < 0.001$).

When we aggregate the Fitzpatrick scale as three groups, I-II, III-VI, and V-VI, verification logistic regression finds statistically significant biases in favor of Fitzpatrick types I-II, over both III-VI and V-VI questions compared (OR = 0.74, $p < 0.001$ for III-VI; OR=0.71, $p < 0.001$ for V-VI). For the identification task, even when controlling for respondent demographic, question subjects with Fitzpatrick values I-II have higher correct responses than that of values V-VI (OR=0.64, $p < 0.001$).

The results are more nuanced for machines.

When we aggregate the Fitzpatrick scale as just "light" and "dark", we observe a statistically significant performance disparity of 6.6% in favor of light-skinned question subjects on the verification task ($p < 0.001$), and for identification, we observe a 7.2% disparity in favor of light-skinned question subjects ($p < 0.001$). When we aggregate the Fitzpatrick scale into three categories, (I-II),(III-IV), and (V-VI), we see a disparity for both tasks between the lightest (I-II) and darkest groups (V-VI) ($p < 0.001$ for both verification and identification).

Subject Fitzpatrick type is revealed to be significantly different, even when controlling for gender, between the types I-II and V-VI (OR = 0.29, $p < 0.001$ for identification; OR=0.12, $p < 0.001$ for verification). This means that I-II and III-VI do not show statistically significant differences for academically-trained models for identification (OR = 0.77, $p = 0.402$) but do for verification(OR=0.15, $p =< 0.001$).

		Very Famous – Verification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.811	0.807	0.808	0.849	0.797	0.8
	2	0.854	0.893	0.888	0.895	0.884	0.89
	3	0.86	0.838	0.837	0.868	0.855	0.861
	4	0.745	0.777	0.772	0.766	0.794	0.737
	5	0.773	0.785	0.794	0.82	0.837	0.783
	6	0.774	0.796	0.835	0.792	0.811	0.806

Table 6: For the very famous, verification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

F.3.4 Human Test-Takers Do Not Perform Better on Subjects of Similar Demographic

Humans do not perform significantly better on questions where the subjects match their gender identity , skin type , or gender identity and skin type. On the identification task, humans do not perform significantly better on questions where subjects match their skin type or both their gender identity and Fitzpatrick type.

F.3.5 Humans Have More Gender Disparity and Machines, but Machines have More Skin Type Disparity than Humans

To test for whether the levels or disparity described above are comparable between humans and machines, we look at the confidence intervals for the odds ratios of comparable models. For both tasks, recall that we observed a disparity on gender and skin type for humans but only on skin type for machines.

The skin type disparity is more pronounced in academic models than in humans. We see that the OR 95% confidence intervals do not match for the darkest skin types in all cases (identification and verification as well as 2 and 3 skin type categories). Furthermore, the academically trained machines show a larger disparity (smaller odds ratio) than humans do. In identification, we have confidence intervals for binary skin types as [0.57, 0.73] for humans and [0.21, 0.50] for academic models. For tertiary skin types, we have confidence intervals of [0.56, 0.74] for humans and [0.17, 0.47] for the darkest subjects. In verification, we have confidence intervals for binary skin types as [0.60, 0.73] for humans and [0.12, 0.36] for academic models. For tertiary skin types, we have confidence intervals of [0.63, 0.80] for humans and [0.05, 0.28] for academic models.

G Raw Performance Tables

In Tables 6-17, we report the results of the humans from each demographic group on each demographic group, broken up by the three recognizability categories. The blue colors indicate, *across columns, i.e., for each respondent demographic*, which question groups have statistically the top performance, as measured from two-sided t-tests.

		Very Famous – Identification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.653	0.703	0.612	0.636	0.715	0.707
	2	0.86	0.854	0.761	0.795	0.811	0.802
	3	0.662	0.778	0.659	0.791	0.782	0.784
	4	0.653	0.676	0.761	0.728	0.714	0.704
	5	0.664	0.603	0.63	0.777	0.722	0.745
	6	0.592	0.552	0.524	0.736	0.682	0.652

Table 7: For the very famous, identification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Very Famous – Verification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.666	0.676
	Male	0.759	0.753

Table 8: For the very famous, verification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Very Famous – Identification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.666	0.676
	Male	0.759	0.753

Table 9: For the very famous, identification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Somewhat Famous – Verification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.557	0.636	0.648	0.644	0.701	0.731
	2	0.698	0.704	0.685	0.74	0.756	0.705
	3	0.759	0.718	0.731	0.721	0.699	0.721
	4	0.694	0.608	0.604	0.674	0.72	0.636
	5	0.639	0.608	0.639	0.687	0.734	0.796
	6	0.612	0.629	0.627	0.686	0.696	0.696

Table 10: For the somewhat famous, verification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Somewhat Famous – Identification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.557	0.636	0.648	0.644	0.701	0.731
	2	0.698	0.704	0.685	0.74	0.756	0.705
	3	0.759	0.718	0.731	0.721	0.699	0.721
	4	0.694	0.608	0.604	0.674	0.72	0.636
	5	0.639	0.608	0.639	0.687	0.734	0.796
	6	0.612	0.629	0.627	0.686	0.696	0.696

Table 11: For the somewhat famous, identification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Somewhat Famous – Verification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.637	0.618
	Male	0.72	0.755

Table 12: For the somewhat famous, verification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Somewhat Famous – Identification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.637	0.618
	Male	0.72	0.755

Table 13: For the somewhat famous, identification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Less Famous – Verification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.794	0.727	0.731	0.793	0.746	0.699
	2	0.786	0.815	0.81	0.849	0.838	0.844
	3	0.778	0.81	0.827	0.816	0.794	0.804
	4	0.756	0.738	0.772	0.773	0.765	0.772
	5	0.776	0.724	0.709	0.735	0.745	0.736
	6	0.71	0.671	0.684	0.746	0.755	0.702

Table 14: For the less famous, verification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Less Famous – Identification Respondent Skin Type					
		1	2	3	4	5	6
Question Skin Type	1	0.577	0.598	0.515	0.613	0.586	0.656
	2	0.811	0.806	0.822	0.808	0.793	0.79
	3	0.814	0.796	0.706	0.769	0.753	0.746
	4	0.677	0.614	0.509	0.661	0.575	0.489
	5	0.536	0.556	0.638	0.643	0.612	0.657
	6	0.41	0.487	0.484	0.6	0.615	0.583

Table 15: For the less famous, identification questions, report the results of the humans from each skin type group on each question skin type. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Less Famous – Verification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.572	0.535
	Male	0.775	0.775

Table 16: For the less famous, verification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.

		Less Famous – Identification Respondent Gender	
		Female	Male
Question Gender Id	Female	0.572	0.535
	Male	0.775	0.775

Table 17: For the less famous, identification questions, report the results of the humans from each gender identity group on each question perceived gender. The blue colors are for each column and indicate those question types which have the highest performance within the statistical t-tests.