A VIEW FROM SOMEWHERE: HUMAN-CENTRIC FACE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biases in human-centric computer vision models are often attributed to a lack of sufficient data diversity, with many demographics insufficiently represented. However, auditing datasets for diversity can be difficult, due to an absence of ground-truth labels of relevant features. Few datasets contain self-identified demographic information, inferring demographic information risks introducing additional biases, and collecting and storing data on sensitive attributes can carry legal risks. Moreover, categorical demographic labels do not necessarily capture all the relevant dimensions of human diversity that are important for developing fair and robust models. We propose to *implicitly* learn a set of continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. We uncover the dimensions by learning on a novel dataset of 638,180 human judgments of face similarity (FAX). We demonstrate the utility of our learned embedding space for predicting face similarity judgments, collecting continuous face attribute values, attribute classification, and comparative dataset diversity auditing. Moreover, using a novel conditional framework, we show that an annotator's demographics influences the *importance* they place on different attributes when judging similarity, underscoring the *need* for diverse annotator groups to avoid biases.

1 INTRODUCTION

Machine learning models depend on the data that they are trained on (Torralba & Efros, 2011). Dataset bias can result in models that are discriminatory (Hendricks et al., 2018; Buolamwini & Gebru, 2018) or depend on spurious correlations (Sagawa et al., 2019; Geirhos et al., 2020). For instance, a face image restoration model (Menon et al., 2020) may *hallucinate* features correlated with oversampled subgroups, resulting in the inadvertent erasure of minority groups (Vincent, 2020). Thus, prior to learning, human-centric image dataset users should evaluate the *diversity* of the data.

The canonical approach is to categorize people using demographic attributes labels. Many equate diversity with parity across the subgroup distributions (Keswani & Celis, 2021; Kay et al., 2015; Bolukbasi et al., 2016; Schwemmer et al., 2020), presupposing access to demographically labeled samples. However, most human-centric image datasets are web-scraped, lacking ground-truth information about the data subjects. Moreover, data protection legislation considers demographic attributes to be personal information and limits their collection and use (Andrus et al., 2021; 2020). Even when demographic labels are known, evaluating diversity by examining counts across subgroups fails to reflect the continuous nature of human phenotypic diversity (e.g., skin tone is often reduced to `light` vs. `dark`). Furthermore, such an approach often denies multi-group membership (Rothbart & Taylor, 1992) (e.g., erasing multi-ethnic and intersex individuals).

When labels are unknown, researchers typically choose certain attributes they consider to be *relevant* for human diversity and use human annotators to infer them (Kärkkäinen & Joo, 2019; Schumann et al., 2021; Wang et al., 2019). In practice, this is difficult for ill-defined and highly changeable social constructs such as race and gender (Hanna et al., 2020; Keyes, 2018; Benthall & Haynes, 2019). Beyond the inclusion of derogatory and offensive categories (Koch et al., 2021; Birhane & Prabhu, 2021; Hanley et al., 2020; Crawford & Paglen, 2019), categorical schemas risk encoding, reifying, and propagating *stereotypes* beyond "their cultural context" (Khan & Fu, 2021). Furthermore, discrepancies between, for example, observed and self-identified race, gender, or other sensitive

attributes can induce psychological distress (Campbell & Troyer, 2007) by invalidating an individual's self-image and identity (Roth, 2016).

Contributions. As an image subject can exhibit certain traits to a greater or lesser extent than others, even within the same subpopulation (Becerra-Riera et al., 2019; Carcagnì et al., 2015; Feliciano, 2016), the goal of our work is to learn a *similarity function*, which measures the similarity between two faces that is *aligned with human perception*. In particular, as similarity is inversely connected to diversity (Leinster & Cobbold, 2012), our model (i.e., similarity function) provides dataset users with a new type of diversity measure that coincides with human perception. Concretely, we present a model of conditional decision-making for learning continuous, human-interpretable face embeddings directly from human judgments of face similarity. Underlying our approach is FAX, a novel dataset of 638,180 judgments over 4,921 faces, where each judgment corresponds to the odd-one-out (i.e., least similar) face in a triplet of faces. Such *contextual* similarity judgments have been shown to offer a *window* into the factors of variation of object categories (Hebart et al., 2020; Zheng et al., 2018), human actions (Dima et al., 2022), and reachspace environments (Josephs et al., 2021) in the human mind. Distinct from other computer vision datasets, each judgment in FAX is accompanied by the identifier and demographic attributes of the annotator who made the judgment.

Our model concurrently learns an embedding space and the *importance* distinct annotators place on each embedding dimension when determining similarity. By learning importance scores, we induce subspaces that encode each annotator's notion of similarity, which increases predictive accuracy. Significantly, we find that annotator subspaces are not interchangeable, but can be *grouped* wrt the sociocultural backgrounds (e.g., nationality, ancestry) of the annotators. Our findings underscore the need for diverse annotator groups to avoid biases. In addition, we show that our embeddings are highly correlated with the human mental representational space of faces, as compared to face embeddings induced by learning to predict face identity or face attributes. Importantly, we find that the individual embedding dimensions are: related to concepts of gender, race, age, as well as face and hair morphology; and, positively correlated with human ratings of face prototypicality. We demonstrate the utility of such dimensions for: collecting continuous (as opposed to discrete) face attribute values for novel faces from annotators; binary attribute classification; and, estimating attribute disparity in datasets.

1.1 Related Work

Disentangled Representation Learning. Our approach is similar in spirit to wholly unsupervised generative approaches (Härkönen et al., 2020; Peebles et al., 2020; Shen & Zhou, 2021; Wei et al., 2021). However, rather than modeling the data generating distribution, our representations are trained to predict human similarity judgments, reflecting the human mental representational space of faces. Our baselines primarily center on embeddings induced by training on facial analysis datasets, since in the face domain, post-hoc interpretation of latent dimensions (Shen et al., 2020; Nitzan et al., 2022; Wu et al., 2021; Locatello et al., 2019) predominantly rely on CelebA (Liu et al., 2015).

Psychological Embeddings. The human mind is conjectured by cognitive psychologists to have "a considerable investment in similarity" (Medin et al., 1993). Significantly, when two entities are compared they mutually constrain the set of features that are activated or inferred in the human mind (Markman, 1996)—i.e., similarity is dynamic, where features are discovered and aligned based on what is being compared. Multidimensional scaling (MDS) is often used to learn psychological embeddings from pairwise similarity judgments (Torgerson, 1958; Zheng et al., 2018; Hebart et al., 2020; Roads & Love, 2021; Van Der Maaten & Weinberger, 2012; Josephs et al., 2021). As MDS cannot embed images outside of the training set, researchers have used pretrained models to predict MDS representations (Sanders & Nosofsky, 2020) or directly infer similarity judgments (Peterson et al., 2018; Attarian et al., 2020). Rather than relying on pretrained models which can introduce implicit biases (Krishnakumar et al., 2021; Steed & Caliskan, 2021), we perform end-to-end learning from scratch. Further, unlike recent work (Hebart et al., 2020; Josephs et al., 2021; Dima et al., 2022), we focus on human perception of naturalistic face images outside of a face recognition setting (McCauley et al., 2021; Sadovnik et al., 2018; Somai & Hancock, 2021), evidenced to be processed differently to objects in the human visual system (Sinha et al., 2006).

Annotator Bias. Categorization by observation depends not only on the subject being categorized, but also on the annotator's sociocultural background, perception of the subject, and contextual cues (Segall

et al., 1966; Balaresque & King, 2016; Hill, 2002; Roth, 2016; Freeman et al., 2011). Despite this, annotator positionality has only recently entered into discourse in computer vision (Chen & Joo, 2021; Zhao et al., 2021), albeit predominantly in the context of natural language processing (Al Kuwatly et al., 2020; Waseem, 2016; Wich et al., 2020; Binns et al., 2017; Salminen et al., 2018). Recent empirical studies confirm this (Denton et al., 2021; Geiger et al., 2020), in particular "only five publications [from 113 surveyed] provided any demographic information regarding who annotated the data instances" (Scheuerman et al., 2021). In order to mitigate bias, one must first measure bias (Le Quy et al., 2022). Addressing the potential for annotator bias, our conditional model elucidates the *importance* distinct annotators place on different attributes when judging similarity. Veit et al. (2017) employ a similar procedure to learn subspaces which encode different notions of similarity (e.g., font style, character type).

2 METHOD

2.1 DATA

To learn a human-interpretable similarity function aligned with human perception, we collect a large-scale dataset of odd-one-out similarity judgments, i.e., FAX. An odd-one-out judgment corresponds to the least similar face in a triplet of face images, representing a *three alternative forced choice task*. FAX contains 638,180 quality-controlled triplets over 4,921 faces, representing 0.003% of all possible triplets. Each triplet is labeled with a judgment and the identifier of the annotator who made the judgment. Quality was controlled by excluding judgments obtained from annotators who provided overly fast, deterministic, or incomplete responses. As in previous non-facial odd-one-out datasets (Josephs et al., 2021; Hebart et al., 2022; Dima et al., 2022), there is a single judgment per triplet. In total, 1,645 annotators contributed to FAX over Amazon Mechanical Turk (AMT), providing their consent, age, nationality, ancestry, and gender identity. Compensation was 15 USD per hour.

For our proof of concept, the 4,921 faces were sampled from the CC-BY licensed dataset FFHQ (Karras et al., 2019). The subset was obtained by first splitting FFHQ into 56 partitions based on estimated intersectional group labels, and then randomly sampling from each partition with equal probability.[1] All faces are near-frontal with little to no eye occlusions and an apparent age $>$ 19 years old. Additional details regarding FAX are provided in Appendix B.2.

Let $\mathbf{x} \in \mathcal{X}$ and $\mathrm{sim} : (\mathbf{x}_i, \mathbf{x}_j) \rightarrow \mathrm{sim}(i, j) \in \mathbb{R}$ denote a face image and a similarity function, resp. Our motivation for collecting FAX is fourfold. Most significantly, the odd-one-out task does not require an annotator to categorize people. Thus, we do not encode, reify, or propagate stereotypes, since we do not use categorical labels, label descriptions, or prototypical imagery. Second, for a triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, repeatedly varying $\mathbf{x}_k$ permits the identification of the *relevant* dimensions that contribute to $\mathrm{sim}(i, j)$. That is, w.l.o.g., $\mathbf{x}_k$ provides context for which $\mathrm{sim}(i, j)$ is determined, making the task easier than *explicit* pairwise similarity tasks (i.e., "Is $\mathbf{x}_i$ similar to $\mathbf{x}_j$?"). This is because it is not always apparent to an annotator which dimensions are relevant when determining $\mathrm{sim}(i, j)$, especially when $\mathbf{x}_i$ and $\mathbf{x}_j$ are perceptually different. Third, there is no need to prespecify attribute lists hypothesized as relevant for comparison (e.g., "Is $\mathbf{x}_i$ *older* than $\mathbf{x}_j$?"). The odd-one-task instead implicitly uncovers salient attributes, which do not explicitly center on normative assessments. Finally, compared to judgments for triplets composed of an anchor (i.e., reference point), $\mathbf{x}_a$, a positive, $\mathbf{x}_p$, and a negative, $\mathbf{x}_n$, odd-one-out judgments naturally provide more information. Odd-one-out triplets require an annotator to determine $\mathrm{sim}(i, j)$, $\mathrm{sim}(i, k)$, and $\mathrm{sim}(j, k)$, whereas triplets with anchors only necessitate the evaluation of $\mathrm{sim}(a, p)$ and $\mathrm{sim}(a, n)$.

2.2 LEARNING HUMAN-CENTRIC FACE REPRESENTATIONS

Background: Multidimensional Scaling. Zheng et al. (2018) developed an MDS approach for learning psychological object embeddings from odd-one-out judgments. The approach is based on three assumptions. First, embeddings can be learned solely from odd-one-judgments, where representations

[1] Ethnicity was estimated using a FairFace model (Kärkkäinen & Joo, 2019); and, binary gender expression and age group were obtained from FFHQ-Aging crowdsourced human annotations (Or-El et al., 2020). See Appendix C.2 for more details.

are constrained to be continuous, non-negative, and sparse. Such properties support interpretability such that dimensions indicate both feature presence and feature magnitude. Second, odd-one-out judgments are a function of $\mathrm{sim}(i, j)$, $\mathrm{sim}(i, k)$, and $\mathrm{sim}(j, k)$. Third, odd-one-out judgments are stochastic, where the probability of selecting x_k as the odd-one-out is $p(k) \propto \exp(\mathrm{sim}(i, j))$.

Model of Conditional Decision-Making. At a high-level, we want to apply the MDS approach outlined above to learn face embeddings. However, MDS cannot embed data outside of the training set, limiting its utility. Moreover, MDS pools all judgments, disregarding intra- and inter-annotator stochasticity. To resolve these issues, we propose to learn a *conditional* convolutional neural network (CNN).

Let $\{(\{\mathbf{x}_{i\ell}, \mathbf{x}_{j\ell}, \mathbf{x}_{k\ell}\}, k\ell, a)\}_{\ell=1}^{n}$ denote a training set of n (triplet, judgment, annotator) tuples, where $a \in \mathcal{A}$. To simplify notation, we assume that judgments always correspond to the index $k\ell$. Suppose $f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{w} \in \mathbb{R}^d$ is a CNN, parametrized by $\mathbf{\Theta}$, where $\mathbf{w} \in \mathbb{R}^d$ is an embedding of $\mathbf{x}$. We model the probability of annotator a selecting $k\ell$ as $p(k\ell \mid a) \propto \exp(\mathrm{sim}_a(i\ell, j\ell))$, where $\mathrm{sim}_a(\cdot, \cdot)$ denotes the *internal* similarity function of a. Given two images $\mathbf{x}_i$ and $\mathbf{x}_j$, we define their similarity according to a as: $\mathrm{sim}_a(i, j) = (\sigma(\boldsymbol{\phi}_a) \odot \mathrm{ReLU}(\mathbf{w}_i))^\top \cdot (\sigma(\boldsymbol{\phi}_a) \odot \mathrm{ReLU}(\mathbf{w}_j))$, where $\sigma(\cdot)$ and $\mathrm{ReLU}(\cdot)$ denote the sigmoid and ReLU functions, resp. Note that $\sigma(\boldsymbol{\phi}_a) \in [0, 1]^d$ is a *mask* associated with a. Each mask plays the role of an element-wise gating function, encoding the *importance* a places on each of the possible d embedding dimensions when determining similarity. By conditioning prediction on annotator identifiers, we induce subspaces that encode each annotator's notion of similarity. This allows us to study whether per-dimensional importance scores differ between annotators.

Let $\mathbf{\Phi} = [\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_{|\mathcal{A}|}^\top] \in \mathbb{R}^{d \times |\mathcal{A}|}$ denote a trainable weight matrix, where each column vector $\boldsymbol{\phi}_a^\top$ corresponds to annotator a's mask prior to applying $\sigma(\cdot)$. In our conditional framework, we jointly optimize $\mathbf{\Theta}$ and $\mathbf{\Phi}$ by minimizing:

$$-\sum_{\ell} \log\left[\hat{p}(k\ell \mid a)\right] + \lambda_1 \sum_{i} \|\mathrm{ReLU}(\mathbf{w}_i)\|_1 + \lambda_2 \sum_{ij} \mathrm{ReLU}(-\mathbf{w}_i)_j + \lambda_3 \sum_{a} \|\boldsymbol{\phi}_a\|_2^2, \quad (1)$$

where $\hat{p}(k\ell \mid a) = \frac{\exp(\mathrm{sim}_a(i\ell, j\ell))}{\exp(\mathrm{sim}_a(i\ell, j\ell)) + \exp(\mathrm{sim}_a(i\ell, k\ell)) + \exp(\mathrm{sim}_a(j\ell, k\ell))}$ is the predicted probability that $k\ell$ is the odd-one-out conditioned on a. The first term in Eq. 1 encourages similar face pairs to result in large dot products. The second term (modulated by $\lambda_1 \in \mathbb{R}$) promotes sparsity. The third term (modulated by $\lambda_2 \in \mathbb{R}$) supports interpretability by penalizing negative values. The last term (modulated by $\lambda_3 \in \mathbb{R}$) penalizes large weights.

When $\lambda_3 = 0$ and $\sigma(\boldsymbol{\phi}_a) = [1, \ldots, 1]$ ($\forall a$), Eq. 1 corresponds to the *unconditional* MDS objective proposed by Zheng et al. (2018). In what follows, we refer to unconditional and conditional models trained on FAX as `FAX-U` and `FAX-C`, resp. In addition, we refer to conditional models whose masks are learned post hoc as `FAX-C-PH`. A `FAX-C-PH` model uses a fixed unconditional model (i.e., `FAX-U`) to obtain face embeddings such that only $\mathbf{\Phi}$ is trainable.

Our conditional decision-making framework is generally applicable to any task that involves mapping from inputs to decisions made by humans; it only requires record-keeping during data collection such that each judgment (or annotation) is associated with the annotator who generated it.

Implementation. All models have ResNet18 (He et al., 2016) architectures and output 128-dimensional embeddings. We use the Adam (Kingma & Ba, 2014) optimizer with default parameters, reserving 10% of FAX for validation. Based on grid search, we empirically set $\lambda_1 = 5 \times 10^{-5}$ and $\lambda_2 = 1 \times 10^{-2}$. For `FAX-C` and `FAX-C-PH`, we additionally set $\lambda_3 = 1 \times 10^{-5}$. Across independent runs, we find that: (i) only a fraction of the 128 dimensions are needed; and (ii) individual dimensions are reproducible. Post-optimization, we remove dimensions with maximal values close to zero. For `FAX-U`, this results in 22 dimensions (61.9% validation accuracy). Note that we observe that 8/22 dimensions have a Pearson's r correlation > 0.9 with another dimension. Further implementation details are included in the Appendix.

3 Experiments

We demonstrate the utility of our FAX induced embedding spaces for predicting face similarity judgments, revealing annotator bias, collecting continuous face attribute values, attribute classification,

and comparative dataset diversity auditing. Throughout this paper, statistical tests correspond to Mann-Whitney U-tests (p-value < 0.05).

For comparison, we consider embeddings extracted from face recognition and facial analysis models. In particular, an ArcFace (Deng et al., 2019) face recognition model trained on CASIA-Webface (CWF) (Yi et al., 2014); and analysis models trained on CelebA (CA) (Liu et al., 2015) (40 binary face attribute labels), FairFace (FF) (Kärkkäinen & Joo, 2019) (perceived binary gender expression, age, and race labels), and FFHQ (binary gender expression, age, and race labels). All models have ResNet18 architectures and output 128-dimensional embeddings. For the analysis models, we also consider unnormalized class logit representations, as well as the logit representations converted to one-hot encodings. All representations are normalized to unit-length, and the dot product of two representations determines their similarity. Additional baselines are considered in the Appendix.

3.1 Predicting human similarity judgments

The utility of any face embedding method is typically determined by measuring whether similar faces are *closer* together in feature space than dissimilar faces. We therefore employ two evaluation protocols, evaluating odd-one-out predictive accuracy.

In **Protocol 1**, we test whether our embeddings are predictive of human judgments not observed during learning. This is a typical MDS test setting, as MDS cannot embed images outside of the training set. Using images from the stimulus set of 4,921 faces, we generate 1,000 novel triplets and collect 22–25 unique judgments on AMT per triplet (24,060 judgments). In addition to predictive accuracy, we compute Spearman's r correlation between the entropy in human- and model-generated triplet odd-one-out probabilities. Human-generated odd-one-out probabilities are of the form: $(n_i, n_j, n_k)/n$, where, w.l.o.g., n_k/n corresponds to the fraction of n odd-one-out votes for k.

In **Protocol 2**, we sample 56 novel face images from FFHQ not contained in the stimulus set. We then generate all $\binom{56}{3}$ possible triplets and collect 2–3 unique judgments on AMT per triplet (80,300 judgments). As an additional performance measure, we compute Spearman's r between the strictly upper triangular model- and human-generated similarity matrices. Entry (i, j) in the human-generated similarity matrix corresponds to the fraction of triplets containing (i, j), where neither was judged as the odd-one-out. Entry (i, j) in a model-generated similarity matrix corresponds to the mean $\hat{p}(i, j)$ over all triplets containing (i, j).

As we have 22–25 judgments per triplet in **Protocol 1**, we can reliably estimate odd-one-out probabilities. The Bayes optimal classifier accuracy corresponds to the *best* possible accuracy any model could achieve given the stochasticity in the human judgments. The classifier makes the most probable prediction, i.e., the majority judgment. Thus, its accuracy is equal to the mean majority judgment probability over the 1,000 triplets, corresponding to 65.5$\pm$1% (95% CI).

Results for **Protocol 1** and **Protocol 2** are shown in Tbl. 1. First, embeddings trained on FAX outperform the baselines across all metrics. In particular, as evidenced by the correlation tests, the human mental representational space is better represented by embedding spaces induced by learning on FAX. This even holds for our `FAX-U` model trained on *only* 72K judgments. Second, as hypothesized, transforming FAX into a dataset of standard triplets constraints with anchors results in lower performance due to a reduction in information. Third, `FAX-C` and `FAX-C-PH` have increased performance over their unconditional counterparts. This shows that the learned annotator-specific masks generalize, i.e., the encoded importance an annotator places on each dimension extends to novel judgments. Most significantly, in **Protocol 1**, our conditional models attain a predictive accuracy at or above the upper bound of the Bayes optimal classifier.

To quantify the number of dimensions required to represent a face while preserving **Protocol 2** performance, we follow Hebart et al. (2020)'s **dimension-elimination approach**. In **Protocol 2**, we iteratively zero out the lowest value per face embedding until a single nonzero dimension remains. As this is done per embedding, the same dimension is not necessarily zeroed out from all embeddings during the same iteration. To obtain 95–99% of the predictive accuracy 6–13 dimensions are required, whereas to explain 95–99% of the variance in the similarity matrix we require 15–22 dimensions. This shows that humans utilize a larger number of dimensions to represent the *global similarity structure of faces* than for predicting individual odd-one-out judgments.

Table 1: Predicting human similarity judgments. **(Protocol 1)** Accuracy over 24,060 judgments and Spearman's r between the entropy in human- and model-generated triplet odd-one-out probabilities. **(Protocol 2)** Accuracy over 80,300 judgments and Spearman's r between the entropy in human- and model-generated similarity matrices. Note: #Labels is an approximate of the number of labels collected to create a dataset accounting for consensus.

					Protocol 1		Protocol 2	
Model	Loss/Method	#Images	#Labels	Arch.	Acc.	r	Acc.	r
FF	Cross-entropy	62K	186K–372K	R18	55.9	0.41	51.9	0.67
FF (logits)	Cross-entropy	62K	186K–372K	R18	52.9	0.25	48.6	0.54
FF (1hot)	Cross-entropy	62K	186K–372K	R18	51.7	0.15	47.7	0.54
CA (embed)	Cross-entropy	178K	7.12M	R18	52.1	0.25	48.9	0.56
CA (logits)	Cross-entropy	178K	7.12M	R18	53.2	0.15	49.6	0.59
CA (1hot)	Cross-entropy	178K	7.12M	R18	47.0	0.14	45.9	0.47
FFHQ	Cross-entropy	70K	210K	R18	55.2	0.38	51.8	0.68
FFHQ (logits)	Cross-entropy	70K	210K	R18	53.5	0.31	50.5	0.62
FFHQ (1hot)	Cross-entropy	70K	210K	R18	51.0	0.23	47.3	0.54
CWF	ArcFace	404K	404K	R18	51.6	0.29	46.1	0.40
FAX-C	Conditional eq. (1)	5K	574K	R18	*67.4*	*0.68*	*61.7*	0.82
FAX-C-PH	Conditional eq. (1)	5K	574K	R18	66.5	*0.68*	61.4	0.82
FAX-U	Unconditional eq. (1)	5K	574K	R18	62.0	0.65	57.5	*0.86*
FAX-U·1/2	Unconditional eq. (1)	5K	287K	R18	61.3	0.61	55.8	0.81
FAX-U·1/4	Unconditional eq. (1)	5K	144K	R18	60.6	0.56	55.3	0.80
FAX-U·1/8	Unconditional eq. (1)	5K	72K	R18	58.6	0.47	55.0	0.79
FAX-Triplet	Triplet margin with distance swap (Balntas et al., 2016)	5K	574K	R18	60.2	0.46	52.8	0.64

Table 2: Annotator bias. Average AUC of linear SVMs trained to discriminate between annotators from different demographic groups.

Annotator attribute	Groups	# Masks	AUC	PDSS
Age group	30-39 / 40-49	393 / 121	0.59 ± 0.05	0.45
Gender identity	Male / Female	523 / 473	0.65 ± 0.05	0.50
Nationality	America / India	530 / 204	0.86 ± 0.03	0.73
Regional ancestry	Europe / Asia	407 / 243	0.86 ± 0.03	0.73
Subregional ancestry	West Europe / South Asia	173 / 107	0.88 ± 0.05	0.73

3.2 Annotator Bias

Conditioning prediction on annotator identifiers provides the best predictive accuracy, evidencing that knowledge of the annotator determining similarity assists in informing the outcome. However, annotators are often framed as interchangeable (Malevé, 2020; Chancellor et al., 2019). To test the validity of this assumption, in **Protocol 2**, we randomly swap the annotator associated with each judgment and then recompute the predictive accuracy of FAX-C. Repeating this process 100 times results in a performance drop from 61.7% to 52.8% $\pm$ 0.02% (95% CI) on average. This shows that annotator subspaces are not interchangeable.

An interesting question relates to whether an annotator's sociocultural background influences their decision making. To evaluate this, we create datasets $\{(\sigma(v_a), y)\}$, where $\sigma(v_a)$ and $y \in \mathcal{Y}$ are annotator a's learned mask and self-identified demographic attribute, resp. For a particular annotator attribute (e.g., nationality), we limit the dataset to annotators who contributed ≥ 200 judgments and belong to one of the two largest groups wrt the attribute. Using 10-fold cross validation, we train linear support vector machines (SVMs) (Hearst et al., 1998) with balanced class weights to predict y from $\sigma(v_a)$. Tbl. 2 shows the average area under the receiver operating characteristic (AUC) for each attribute and the proportion of dimensions with a statistically significant difference (PDSS). Most significantly, none of the AUC confidence intervals include chance performance. Moreover, the linear SVMs are able to discriminate between binary groups wrt nationality, regional ancestry and subregional ancestry with high probability (86–88%).

3.3 Interpretability

Since we aim to replace the *explicit* collection of problematic categorical labels, we evaluate whether the individual dimensions are human-interpretable through a **qualitative dimension labeling task** and a **quantitative dimension rating task** (Hebart et al., 2020; Josephs et al., 2021). In what follows, let x_i and $\mathrm{ReLU}(w_i)$ denote an image from the stimulus set of 4,921 faces and its corresponding 22-dimensional embedding, resp. Further, let $\%\mathrm{tile}_j^q$ denote the q-th percentile of the dimension

Figure 1: **Dimension labeling.** The top 8 faces from the stimulus set with the highest dimension j embedding values for 6-of-the-22 FAX-model dimensions. From left to right, found dimensions coincide with commonly defined demographic groups `South Asian` ($j = 1$), `Female` ($j = 4$), `Smiling Expression` ($j = 5$), `Wide Face` ($j = 10$), `Elderly` ($j = 11$), and `Balding` ($j = 18$). Word clouds generated from annotator labels (transformed into topics) are shown above each set of 8 faces. Word clouds generated using CelebA and FairFace models are shown below each set of 8 faces. The CelebA and FairFace labels were obtained by labeling the entire stimulus set (using the CelebA and FairFace models) and determining the attributes with the highest AUC.

j embedding values. We define grid_j as a 5×100 grid of faces. Column $q \in [100, \ldots, 1]$ of grid_j corresponds to faces from the stimulus set of 4,921 faces with the top 5 highest dimension j embedding values, satisfying $\%\text{tile}_j^{q-1} \leq \text{ReLU}(w)_j < \%\text{tile}_j^q$. Thus, from left (high) to right (low), grid_j shows example faces from each percentile for dimension j.

For the **dimension labeling task**, we task annotators with writing 1–3 visual characteristics that describe grid_j. For each grid_j, we collect 25–62 labels. As the task is open-ended, we convert the labels into 35 broad *topics*. Note that annotator descriptions were sometimes derogatory and offensive.

Fig. 1 provides evidence of the interpretability for a random subset of `FAX-U` dimensions. There is clear relationship between the dimension topics obtained from annotator descriptions and dimension labels generated using CelebA and FairFace models. Across the 22 dimensions, we note the materialization of individual dimensions coinciding with commonly defined demographic groups, i.e., `Male`, `Female`, `Black`, `White`, `East Asian`, `South Asian`, and `Elderly`. In addition, separate dimensions surfaced for face and hair morphology, i.e., `Wide Face`, `Long Face`, `Smiling Expression`, `Neutral Expression`, `Balding`, `Facial Hair`, and `Dyed Hair`. This is achieved without ever asking an annotator to explicitly categorize a person.

Let μ_j^q denote the mean value of `FAX-U` dimension j embedding values of *all* 4,921 faces from the stimulus set, satisfying $\%\text{tile}_j^{q-1} \leq \text{ReLU}(w)_j < \%\text{tile}_j^q$. For the **dimension rating task**, annotators must place novel faces (e.g., x) above a single column q of grid_j. Placement is based on the similarity between x and the faces contained in each column q of grid_j. Annotators are not primed with the meaning of grid_j. If x is accurately placed on grid_j, then this indicates that the ordering of dimension j is visually coherent. To test this, we sample 20 novel faces from FFHQ, i.e., the faces are not contained in the stimulus set. $\forall(x, \text{grid}_j)$, we collect 20 unique judgments (8,800 judgments). From the judgments, for each x, we create 22-dimensional human-generated embeddings of the form: $[\frac{1}{20}\sum_{n=1}^{20} \mu_1^{a_n}, \ldots, \frac{1}{20}\sum_{n=1}^{20} \mu_{22}^{a_n}]$, where $a_n \in [100, \ldots, 1]$ denotes the n-th annotator's column choice q. Next, we generate all $\binom{20}{3}$ possible triplets to create a human-generated similarity matrix $M \in \mathbb{R}^{20\times 20}$. Entry (i, j) in M corresponds to the mean $\hat{p}(i, j)$ over all triplets containing (i, j) using the human-generated embeddings. We create a model-generated matrix in the same way, but this time using the model-generated embeddings.

Spearman's r correlation between the strictly upper triangular model- and human-generated similarity matrices is 0.83 and 0.86 for `FAX-U` and `FAX-C-PH`, resp. This shows that: (i) dimensional values correspond to the feature magnitude; and (ii) image grids can be used to *directly* collect continuous attribute values for faces, sidestepping the limits of categorical definitions (Keyes, 2018; Benthall & Haynes, 2019; Khan & Fu, 2021). Note that the use of an image grid is not restricted to any particular embedding method, as long as the dimension is human-interpretable.

3.4 PROTOTYPICALITY

To test whether dimensional values represent the typicality of faces, we use face images from Chicago Face Database (CFD) (Ma et al., 2015; Lakshmi et al., 2021; Ma et al., 2021) labeled with *prototypicality* ratings. The ratings (obtained from human annotators) correspond to the average prototypicality of a face wrt a race category from one (less typical) to five (very typical), considering skin color, hair, eyes, nose, cheeks, lips, and other physical features. For gender expression, ratings correspond to the typicality of the face relative to others of the same race and gender in the United States from one (not at all typical) to seven (extremely typical).

Table 3: Prototypicality. Spearman's r between human ratings of face attribute prototypicality and model-generated dimensional values.

	Attribute				
Model	Masculine	Feminine	East Asian	Black	White
FAX-U	0.813	0.840	0.747	0.683	0.644
CA	0.858	0.855	—	—	0.123
FF	0.836	0.827	0.806	0.653	0.645
FFHQ	0.812	0.829	0.794	0.587	0.688

Tbl. 3 shows that relevant FAX-U dimensions are positively correlated with the typicality ratings according to Spearman's r. Although the odd-one-task does not require an annotator to explicitly categorize any person, category typicality appears to manifest from the similarity judgments for concepts related to race and gender.

3.5 BINARY ATTRIBUTE CLASSIFICATION

Facial analysis models are trained for classification, therefore we test the utility of FAX-U dimensions for discriminating between binary-valued attributes. For labeled face data, we use COCO (Lin et al., 2014), OpenImages MIAP (MIAP) (Schumann et al., 2021), CFD, FFHQ, CelebA (CA) (Liu et al., 2015), and Casual Conversations (CC) (Hazirbas et al., 2021).

Tbl. 4 shows that FAX-U dimensions are competitive with the baselines, even in challenging unconstrained settings as represented by COCO and MIAP. Therefore, training on FAX results in interpretable dimensions that can serve as classifiers. Notably, FAX-U dimensions have the highest AUC for self-identified CFD attributes Black, East Asian, and Indian.

Table 4: Binary attribute classification. Using attribute values from relevant model dimensions to perform binary attribute classification.

		Model AUC			
Dataset	Attribute	FAX-U	CA	FF	FFHQ
CC	> 70 y.o.*	0.800	0.936	0.959	0.962
FFHQ	> 70 y.o.	0.905	0.959	*0.971*	*0.980*
CFD	Male*	0.991	0.997	0.996	*0.998*
CC	Male*	0.971	0.986	*0.990*	0.981
CA	Male	0.990	*0.999*	0.994	*0.995*
COCO	Male	0.893	0.926	*0.963*	0.958
MIAP	Male	0.924	0.942	*0.945*	0.938
FFHQ	Male	0.933	0.959	*0.988*	*0.996*
CA	Smiling	*0.895*	*0.982*	—	—
CFD	Happy open mouth	0.969	*0.992*	—	—
CFD	Neutral	*0.731*	—	—	—
CFD	East Asian*	*0.969*	—	0.955	0.938
CFD	Black*	*0.992*	—	*0.992*	0.988
CFD	White*	0.972	0.889	*0.990*	0.988
CFD	Indian*	*0.960*	—	0.848	0.883
CC	Light skin	0.930	0.830	*0.965*	0.960
COCO	Light skin	0.889	0.771	0.927	*0.931*
CA	Balding	*0.963*	*0.995*	—	—

3.6 COMPARATIVE DATASET DIVERSITY AUDITING

Drawing inspiration from biodiversity measures (Leinster & Cobbold, 2012), we propose to use our learned dimensions for comparative dataset diversity auditing. Given a set of candidate datasets $\{c_k\}_{k=1}^n$, we aim to find $c_{k^\star} = \arg\min_{c_k} \mathrm{sim}(c_k)$, where $\mathrm{sim}(c_k)$ measures the similarity of the faces in $x \sim c_k$. We denote by $q_k \in [0, 0.01, \ldots, 0.99, 1]$ and $1 - q_k$ the proportion of images in c_k from $D_0 = \{(x_i, y_0)\}_i$ and $D_1 = \{(x_j, y_1)\}_j$, resp., where $D_0 \cap D_1 = \emptyset$. Here $y_0, y_1 \in \{0, 1\}$ are sensitive attribute labels. Let $\hat{y} \in \mathbb{R}$ denote a *proxy* continuous attribute value. We define the similarity between $(x_i, x_j) \sim c_k$ as $z_{ij} = \mathrm{abs}(\hat{y}_i - \hat{y}_j)^{-1}$. Further, let $Z_k = (z_{ij}) \in \mathbb{R}^{|c_k| \times |c_k|}$ denote c_k's similarity matrix. In biodiversity terms, the average *ordinariness* of faces in c_k is $\propto \sum_i \sum_j z_{ij}$ $(\forall i)$. This quantity is large when most faces in c_k are *concentrated* into a few very similar faces. Importantly, concentration is inversely connected to diversity. Therefore, we can interpret the mean of Z_k as a diversity score, which we denote by $\mathrm{divscore}(Z_k)$. Note that $\mathrm{divscore}(Z_k) \to 0$ for homogeneous sets of faces.

For face data, we sample from CC and CFD, as both datasets have self-identified labels. To find $c_{k^\star}$, wrt a label, we use relevant attribute values from FAX-U and relevant unnormalized logits from CA, FF, and FFHQ. Note that we do not compare candidates from multiple datasets, i.e., we sample each $c_k \in \{c_k\}_{k=1}^n$ from a single dataset (e.g., CFD). We bootstrap each c_k 100 times, where $|c_k| = 100$

($\forall k$). To measure performance, we compute: (i) the attribute disparity in $c_{k^\star}$, i.e., $\Delta = \text{abs}(2q_k - 1)$; and (ii) Spearman's r correlation between the k bootstrapped average diversity scores and the k ground-truth disparity scores.

Results are shown in Tbl. 5. On CFD, `FAX-U` has the lowest Δ for `Male` and `Indian`. On CC, `FAX-U` results in the lowest Δ for `Light skin`. In all other cases, `FAX-U` is competitive with the baselines. In particular, Spearman's r evidences that `FAX-U` disparity scores are highly correlated with ground truth disparity based on labels.

Table 5: Attribute disparity. Using attribute values from relevant model dimensions to find the most diverse candidate dataset wrt an attribute of interest.

		Model (Disparity Δ / Spearman's r)			
Dataset	Attribute	`FAX-U`	`CA`	`FF`	`FFHQ`
CC	$>$ 70 y.o.*	0.22 / 0.96	*0.06* / *0.99*	*0.06* / 0.95	0.08 / *0.99*
CC	Male*	0.06 / 0.97	*0.02* / 0.99	*0.02* / *1.00*	0.04 / 0.99
CC	Light skin	*0.06* / 0.94	0.3 / 0.65	0.18 / 0.85	*0.06* / *0.99*
CFD	Happy open mouth	0.12 / *0.99*	0.06 / *0.95*	—	—
CFD	Male*	*0.00* / *1.00*	0.08 / 0.99	0.02 / *1.00*	0.02 / *1.00*
CFD	East Asian*	0.06 / *1.00*	—	*0.02* / 0.98	*0.02* / *1.00*
CFD	Black*	0.10 / 0.98	—	0.14 / 0.97	*0.00* / *1.00*
CFD	White*	0.04 / *1.00*	*0.02* / 0.99	0.04 / 0.97	0.04 / 0.99
CFD	Indian*	*0.08* / *1.00*	—	0.30 / 0.73	*0.08* / 0.87

4 Discussion

We proposed a method for *implicitly* learning continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. We uncovered the face embedding space by learning on a novel dataset of human judgments of face similarity (FAX). We showed that the individual dimensions are human-interpretable and related to concepts of gender, race, age, as well as face and hair morphology categories. We demonstrated the utility of our learned embedding space for predicting face similarity judgments, collecting continuous face attribute values, attribute classification, and comparative dataset diversity auditing. Moreover, using a novel conditional framework, we showed that an annotator's demographics influences the *importance* they place on different attributes when judging similarity, underscoring the *need* for diverse annotator groups to avoid biases. As our conditional decision-making framework is generally applicable to any task that involves mapping from inputs to decisions made by humans, it would be interesting to study the behavior of annotators in other visual tasks.

However, our work is not without its limitations. First, we assume that the stimuli set is sufficiently diverse. Therefore, our current proposal is limited to the proof of concept that our approach can only uncover factors that vary in the data and are salient to human perception of face similarity. For instance, if skin color does not vary among data instances, then skin color cannot possibly influence human judgments, rendering it impossible to learn skin color as a face dimension. Second, the number of (near) nonzero dimensions depends on the $L1$ sparsity penalty, which must be carefully chosen. Too high of a penalty will result in the entanglement of distinct factors of variation, whereas too low of a penalty will result in repeated dimensions. We erred on the side of caution (lower penalty), since it is more difficult to localize the source of bias when factors are merged. Third, we found that participants did not always follow our instructions in full. Dimensions emerged corresponding to facial expression, which participants were instructed to ignore. Finally, our participant pool was demographically imbalanced, which is an unfortunate consequence of using AMT.

Despite these limitations, we have shown that similarity judgments are a valuable means by which to reveal the human mental representations of faces. Of particular note is the fact this was achieved by learning on 0.003% of all possible triplets from 4,921 faces. We hope that our work inspires others to pursue unorthodox tasks for learning the dimensions of human diversity, which do not encode, reify, and propagate stereotypes, or invalidate the self-image and identity of image subjects. Moreover, we aim to increase discourse on annotator positionality. As the philosopher Thomas Nagel suggested, it is impossible to take a "view from *nowhere*" (Nagel, 1989). Therefore, we need to make datasets more inclusive by integrating a diverse set of perspectives from their inception.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 184–190, 2020.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

McKane Andrus, Elena Spitzer, and Alice Xiang. Working to address algorithmic bias? don't overlook the role of demographic data. *Partnership on AI. Retrieved from https://www. partnershiponai. org/demographic-data*, 2020.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 249–260, 2021.

Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *arXiv preprint arXiv:2109.13228*, 2021.

Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*, 2020.

Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *European Conference on Computer Vision*, pp. 547–563. Springer, 2020.

P Balaresque and TE King. Human phenotypic diversity: An evolutionary perspective. *Current topics in developmental biology*, 119:349–390, 2016.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, pp. 3, 2016.

Lawrence Barsalou. The instability of graded structure: Implications for the nature of concepts. 1987.

Fabiola Becerra-Riera, Annette Morales-González, and Heydi Méndez-Vázquez. A survey on facial soft biometrics for video surveillance and forensic applications. *Artificial Intelligence Review*, 52 (2):1155–1187, 2019.

Mostapha Benhenda. Can ai reproduce observed chemical diversity? *bioRxiv*, pp. 292177, 2018.

Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 289–298, 2019.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pp. 405–415. Springer, 2017.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Mary E Campbell and Lisa Troyer. The implications of racial misclassification by observers. *American Sociological Review*, 72(5):750–765, 2007.

Pierluigi Carcagnì, Marco Del Coco, Dario Cazzato, Marco Leo, and Cosimo Distante. A study on different experimental configurations for age, race, and gender estimation problems. *EURASIP Journal on Image and Video Processing*, 2015(1):1–22, 2015.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14980–14991, 2021.

Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *AI and Society*, 2019.

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.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1962–1981, 2022.

Diana C Dima, Martin N Hebart, and Leyla Isik. A data-driven investigation of human action representations. *bioRxiv*, 2022.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2021.

Cynthia Feliciano. Shades of race: How phenotype and observer characteristics shape racial classification. *American Behavioral Scientist*, 60(4):390–419, 2016.

Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 325–336, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pp. 22–34. PMLR, 2021.

Nelson Goodman. Seven strictures on similarity. 1972.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. An ethical highlighter for people-centric dataset creation. *arXiv preprint arXiv:2011.13583*, 2020.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512, 2020.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

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.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data: A multimodal collection of large-scale datasets for investigating object representations in brain and behavior. *bioRxiv*, 2022.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, 2018.

Mark E Hill. Race of the interviewer and perception of skin color: Evidence from the multi-city study of urban inequality. *American Sociological Review*, pp. 99–108, 2002.

Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 743–756, 2022.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Emilie L Josephs, Martin N Hebart, and Talia Konkle. Emergent dimensions underlying human perception of the reachable world. *Journal of Vision*, 21(9):2154–2154, 2021.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828, 2015.

Vijay Keswani and L Elisa Celis. Auditing for diversity using representative examples. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 860–870, 2021.

Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 587–597, 2021.

Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758, 2009.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, volume 1, pp. 3, 2021.

Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12:161, 2021.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 693–702, 2021.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1452, 2022.

Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14970–14979, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

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

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Bradley C Love and Brett D Roads. Similarity as a window on the dimensions of object representation. *Trends in Cognitive Sciences*, 25(2):94–96, 2021.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

Debbie S Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53(3):1289–1300, 2021.

Nicolas Malevé. On the data set's ruins. *AI & SOCIETY*, pp. 1–15, 2020.

Arthur B Markman. Structural alignment in similarity and difference judgments. *Psychonomic Bulletin & Review*, 3(2):227–230, 1996.

Arthur B Markman and Dedre Gentner. Nonintentional similarity processing. *The new unconscious*, pp. 107–137, 2005.

John McCauley, Sobhan Soleymani, Brady Williams, John Dando, Nasser Nasrabadi, and Jeremy Dawson. Identical twins as a facial similarity benchmark for human facial recognition. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2021.

Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Thomas Nagel. *The view from nowhere*. oxford university press, 1989.

Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19239–19249, 2022.

Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision*, pp. 739–755. Springer, 2020.

William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *European Conference on Computer Vision*, pp. 581–597. Springer, 2020.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42 (8):2648–2669, 2018.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*, 2019.

Brett D Roads and Bradley C Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3547–3557, 2021.

Emilie M Roth and Edward J Shoben. The effect of context on the structure of categories. *Cognitive psychology*, 15(3):346–378, 1983.

Wendy D Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016.

Myron Rothbart and Marjorie Taylor. Category labels and social reality: Do we view social categories as natural kinds? 1992.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Amir Sadovnik, Wassim Gharbi, Thanh Vu, and Andrew Gallagher. Finding your lookalike: Measuring face similarity rather than face identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2345–2353, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 88–94. IEEE, 2018.

Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

Craig A Sanders and Robert M Nosofsky. Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 3(3):229–251, 2020.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Rebecca Pantofaru. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

Marshall H Segall, Donald Thomas Campbell, and Melville Jean Herskovits. *The influence of culture on visual perception*. Bobbs-Merrill Indianapolis, 1966.

Benny Shanon. On the similarity of features. *New ideas in psychology*, 6(3):307–321, 1988.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94 (11):1948–1962, 2006.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Rosyl S Somai and Peter JB Hancock. Exploring perceived face similarity and its relation to image-based spaces: an effect of familiarity. *Journal of Vision*, 21(9):2149–2149, 2021.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 701–713, 2021.

L Edward Thorndike. Human learning. *The Journal of Nervous and Mental Disease*, 75(5):589, 1932.

Warren S Torgerson. Theory and methods of scaling. 1958.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 830–838, 2017.

James Vincent. What a machine learning tool that turns obama white can (and can't) tell us about ai bias. https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias, 2020. Accessed May 16, 2022.

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, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 692–702, 2019.

Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pp. 138–142, 2016.

Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6721–6730, 2021.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 191–199, 2020.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021.

Charles Y Zheng, Francisco Pereira, Chris I Baker, and Martin N Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, 2018.

A EXTENDED RELATED WORK

Similarity. Similarity represents a central theoretical construct in cognitive psychology (Medin et al., 1993; Markman & Gentner, 2005; Goodman, 1972), in particular as the human mind is conjectured by many scholars to have "a considerable investment in similarity" (Medin et al., 1993). This is because, similarity, a type of comparison, permits people to make inferences on the basis of incomplete information, i.e., transfer learning (Thorndike, 1932). The process of comparing similarity hinges on the alignment of the attributes of the objects being compared (Markman, 1996). Alignment is contended to be weighted, where humans multiplicatively integrate over both *similar* and *dissimilar* attributes to arrive at a judgment (Medin et al., 1993). This involves evaluating the attributes present in one object and measuring their fit in the other object. In other words, when two objects are compared they mutually constrain the set of features that are activated or inferred in the human mind. For example, determining whether an object is red can depend on the object it is being compared to. As another example, comparing face A to face B will most likely involve different sets of features than when comparing face A to face C. That is, face A may be associated with different features when we vary the face it is being compared to. The process of alignment is contended to be central to similarity comparisons. Shanon (1988) goes as far as arguing that similarity is not a function of features, but that the features themselves are a function of the similarity comparison. Thus similarity is dynamic, where features are discovered and aligned based on what is being compared.

Contextual Similarity. Human perception of similarity can vary with respect to context (Barsalou, 1987; Roth & Shoben, 1983). For example, Barsalou (1987) found that a snake and raccoon were judged to be less similar when participants were not given any context. However, when the comparison was contextualized, i.e., by placing snakes and raccoons in the context of *pets* their similarity increased. The odd-one-out similarity task Dima et al. (2022); Josephs et al. (2021); Zheng et al. (2018); Hebart et al. (2020) also provides context. For instance, it requires annotators to perform pairwise comparisons in order to determine the odd-one-out. By varying the context, e.g., face C in the triplet (A, B, C), we uncover different sets of features that contribute to the similarity and dissimilarity between A and B. This is important as there are an uncountable number of ways in which two objects may be similar or dissimilar (Love & Roads, 2021). Odd-one-out similarity judgments (Zheng et al., 2018; Hebart et al., 2020; Josephs et al., 2021; Dima et al., 2022), or more generally contextual similarity comparisons, can be used to uncover the salient features that contribute to pairwise similarity. As highlighted in several cognitive psychology works (Medin et al., 1993; Markman & Gentner, 2005; Goodman, 1972), context makes salient: (i) context-related properties; and (ii) the extent to which objects being compared share context-related properties.

Data Protection and Privacy. Based on the principle of data minimization (i.e., only collect what you need), privacy and legal counsel typically advise against collecting demographic information even if when it is for fairness purposes, so as to avoid potential privacy violations. Therefore, many practitioners avoid collecting such information (Andrus et al., 2021), in particular special, and/or sensitive, category data (e.g., race), which is often subject to stronger restrictions. However, if we are unable to collect demographic information, as well as other attributes, for developing fairer and more ethical AI systems, then we require creative solutions to ensure that datasets have some level of diversity and are representative.

While labels can be inferred using attribute classifiers, there are concerns with this, which have been voiced by many practitioners (Andrus et al., 2021). Data access request rights (e.g., via GDPR, CCPA, PIPL) may force data holders into revealing information held about data subjects, including inferred information. Due to this, practitioners are more cautious about inferring labels about people (Andrus et al., 2021), especially as mismatches between inferred and self-identified race, gender, or other sensitive and nonsensitive attributes can induce psychological distress (Campbell & Troyer, 2007) by invalidating an individual's self-image and identity (Roth, 2016). Our method instead implicitly learns face-varying axes from similarity judgments, altogether avoiding any explicit labeling of faces. We show that the axes are meaningful and human-interpretable through the dimension labeling and dimension rating human experiments. Our model (i.e., similarity function) provides dataset users with a new type of diversity measure that coincides with human perception, which can be plugged into any metric (Friedman & Dieng, 2022; Benhenda, 2018).

Together, our dataset, FAX, and model offer a creative solution to the issues arising from data protection and privacy legislation: (i) FAX does not contain any attribute labels; and (2) our model does not explicitly label faces.

Bias Detection. Several works cluster pretrained embeddings to surface systematic failure modes (Krishnakumar et al., 2021; d'Eon et al., 2022; Sohoni et al., 2020). As the clusters are often difficult to interpret, embeddings that intergrate visual and textual information have been proposed (Eyuboglu et al., 2021). Cross-modal embeddings, however, are trained on datasets that contain negative stereotypes and racist slurs (Birhane et al., 2021), causing downstreams harms (Hundt et al., 2022). Others utilize disentangled representations learned by generative models for unknown attribute bias discovery in classifiers (Li & Xu, 2021; Lang et al., 2021). Balakrishnan et al. (2020) find semantic latent space directions, via human annotators, which *explicitly* rely on normative assessments.

Disentangled Representation Learning. Learning interpretable representations through semantically labeled concepts and prototypes has recently gained attention (Alvarez Melis & Jaakkola, 2018; Koh et al., 2020; Li et al., 2018). Others reveal meaningful directions inside generative models, where directional distances estimate the magnitude of attributes (Nitzan et al., 2022). However, the identification of interpretable directions is performed post-hoc, requiring potentially biased labeled examples and/or pretrained models (Wu et al., 2021; Locatello et al., 2019), self-supervision (Plumerault et al., 2019; Jahanian et al., 2019), or paired images sharing at least one factor of variation (Locatello et al., 2020). Our approach is similar in spirit to wholly unsupervised approaches based on PCA (Härkönen et al., 2020), Hessian regularization (Peebles et al., 2020), and orthogonal Jacobian regularization (Shen & Zhou, 2021; Wei et al., 2021). Rather than modeling the data generating distribution, our representations are trained to predict human similarity judgments, reflecting the human mental representational space of faces. Our baselines primarily center on embeddings induced by training on facial analysis datasets, since in the face domain, post-hoc interpretation of latent dimensions (Shen et al., 2020; Nitzan et al., 2022; Wu et al., 2021; Locatello et al., 2019) predominantly rely on CelebA (Liu et al., 2015). Models trained on such datasets effectively map from image space to concept space, similar to encoders in concept bottleneck models (Koh et al., 2020).

B DATA

B.1 FACE IMAGE STIMULI

The stimulus set was sampled from the publicly available FFHQ dataset (Karras et al., 2019) of 70,000 permissively licensed face images. The FFHQ dataset which includes JSON metadata and documentation is available from the following repository: `https://github.com/NVlabs/ffhq-dataset`.

For face detection, as in (Karras et al., 2019), we used the dlib mmod human face detector (King, 2009) (`https://github.com/davisking/dlib-models`). To align the detected faces, we followed the same procedure as detailed in (Or-El et al., 2020), which uses the dlib 68 landmark shape predictor (King, 2009) (`https://github.com/davisking/dlib-models`).

As stated in the main text, we used FFHQ-Aging annotations (Or-El et al., 2020), obtained using Face++, for absolute head yaw angle, absolute head pitch angle, eyewear, and eye occlusion annotations to prune the dataset. The purpose of these constraints was to reduce the influence of head pose on human behavior and to better allow for eye comparisons. The stimulus set of 4,921 representative examples were then sampled from the remaining images such that each intersectional subgroup based on perceived binary gender expression, age group, and ethnicity would be included. We only considered face images with an apparent age greater than 19 years old in an attempt to exclude images of minors. Ethnicity was estimated using a FairFace (Kärkkäinen & Joo, 2019) trained model (implementation details are in Appendix C.2). Perceived binary gender expressions and apparent age groups were obtained from FFHQ-Aging crowdsourced annotations. To mitigate biasing our set

Table 6: Apparent age group, binary gender expression group, and ethnicity group counts for the stimuli of 4,921 representative examples in the face image stimuli set

Age	Count
20–29	1400
30–39	1351
40–49	1181
50–120	989

Ethnicity	Count
White	998
East Asian	801
Latino Hispanic	788
Black	729
Indian	556
Middle Eastern	538
Southeast Asian	511

Gender	Count
Male	2607
Female	2314

toward stereotypical faces, we randomly sampled images from the 56 possible intersectional subgroups rather than selecting the most confidently predicted or annotated.

Table 6 shows the group counts at the level of age, gender, and ethnicity, whereas Table 8 reports the counts at the finer intersectional subgroup level. Exact balancing was not possible due to the skewed composition of FFHQ. Ideally, we would have sampled images from a dataset with self-reported attributes. However, while such datasets do exist, they are typically small, tightly constrained, and largely unrepresentative.

B.2 Collection of human behavioral judgments

All participants were recruited on AMT and were required to have previously completed a minimum of 100 human intelligence tasks (HITs) on AMT with a 95% approval rating. Eligibility was further determined through a prescreening survey (nominal fee of 0.01 USD), which included a short multiple-choice English language proficiency test. Since we placed no restriction on participant location, in order to be eligible, we required participants to answer at least two out of three multiple-choice English language proficiency questions correctly.

Collection of human behavioral odd-one-out similarity judgments for training and validation. Based on estimated face image labels, a triplet can comprise 1–2 binary gender expression group(s), 1–3 age group(s), and 1–3 ethnicity group(s) resulting in 18 triplet *types* which were uniformly sampled. We sampled 703,300 unique triplets in total. Participants were presented with triplets alongside the following minimal instructions: "Choose the person that looks *least* similar to the two other people. Focus your judgment on the people depicted, ignoring differences in head position/direction, facial expression/emotion, lighting, accessories, and background/objects". Each HIT consisted of 20 unique triplets, where each triplet was presented separately and in succession. We collected 703,300 unique triplet similarity judgments (i.e., one judgment per triplet) on AMT from 1,887 eligible participants. Each triplet was judged once by a single participant, resulting in a single judgment per triplet. To control for quality, according to the criteria outlined in Table 7, we excluded triplet judgments obtained from any participant who provided overly fast, deterministic, or incomplete judgments. This left 638,180 judgments from 1,645 participants. Irrespective of estimated quality, all participants were compensated at a rate of 15 USD per hour. Table 9 shows the age, ancestry, gender identity, and nationality group counts of the 1,645 participants.

Collection of human behavioral odd-one-out similarity judgments for testing: Same images, novel triplets. We sampled an additional 1,000 triplets (images were sampled from the stimulus set of 4,921 images). We collected 25 unique judgments per triplet on AMT (total: 25,000 judgments). After excluding responses based on the quality control criteria outlined in Table 7, we were left with 22–25 unique judgments per triplet (total: 24,060 judgments) obtained from 355 participants. Table 10 shows the age, ancestry, gender identity, and nationality group counts of the 355 participants.

Collection of human behavioral odd-one-out similarity judgments for testing: Novel images, novel triplets. We sampled all possible triplets from 56 novel images (sampled from FFHQ, but disjoint from the stimulus set) and collected 3 unique judgments per triplet on AMT (total: 83,160 judgments). After excluding responses based on the quality control criteria outlined in Table 7, we were left with 2–3 unique judgments per triplet (total: 80,300 judgments) obtained from 632 participants. Table 11 shows the age, ancestry, gender identity, and nationality group counts of the 632 participants.

Collection of dimension labels. For each of the 22 dimensions, we constructed a scale (measurement units corresponded to dimensional values which were converted to percentiles) and tasked participants with providing up to three labels (minimum one label) using words only. Note that participants were

Table 7: Participant exclusion policies and criteria

Policy	Criteria	Min. # of judgments
Fast #1	> 25 % of judgments generated in < 0.8 s	100
Fast #2	> 50 % of judgments generated in < 1.1 s	100
Deterministic	> 40 % of judgments correspond to a single triplet position	200
Incomplete	Submission of an empty judgment	1

Table 8: The 56 intersectional subgroup counts for the stimuli of 4,921 representative examples

Age	Ethnicity	Gender	Count
50–120	White	Male	140
20–29	White	Female	135
40–49	White	Male	133
20–29	White	Male	127
30–39	Latino Hispanic	Female	126
30–39	White	Male	125
50–120	White	Female	118
20–29	East Asian	Female	118
30–39	East Asian	Female	115
30–39	White	Female	112
30–39	East Asian	Male	111
40–49	White	Female	108
20–29	East Asian	Male	107
20–29	Latino Hispanic	Male	106
40–49	Black	Male	105
40–49	Latino Hispanic	Female	105
20–29	Latino Hispanic	Female	104
30–39	Black	Male	104
40–49	East Asian	Female	102
30–39	Middle Eastern	Male	100
20–29	Black	Male	99
40–49	Latino Hispanic	Male	98
50–120	Middle Eastern	Male	97
40–49	Middle Eastern	Male	97
20–29	Middle Eastern	Male	96
30–39	Black	Female	95
20–29	Indian	Female	95
30–39	Indian	Female	94
40–49	Black	Female	93
30–39	Latino Hispanic	Male	93
20–29	Black	Female	92
40–49	East Asian	Male	91
30–39	Indian	Male	91
50–120	East Asian	Male	88
50–120	Black	Male	86
20–29	Southeast Asian	Female	86
20–29	Indian	Male	83
20–29	Southeast Asian	Male	82
50–120	Latino Hispanic	Female	81
50–120	Latino Hispanic	Male	75
30–39	Southeast Asian	Male	75
30–39	Southeast Asian	Female	72
20–29	Middle Eastern	Female	70
50–120	East Asian	Female	69
40–49	Indian	Male	66
50–120	Southeast Asian	Female	58
40–49	Indian	Female	58
50–120	Black	Female	55
40–49	Southeast Asian	Male	54
40–49	Southeast Asian	Female	47
50–120	Indian	Male	41
30–39	Middle Eastern	Female	38
50–120	Southeast Asian	Male	37
50–120	Indian	Female	28
40–49	Middle Eastern	Female	24
50–120	Middle Eastern	Female	16

not restricted to entering single word labels, they were provided with free-text boxes. Participants were presented with a scale and were asked: "Which *visual* characteristics do you *think* the people at the high end of the scale have in common compared to the people at the low end of the scale?". In addition, participants were asked to view and interact with the entire scale before providing any labels. Figure 4 shows an example of a scale and a novel face image shown to participants. After quality control, there were 25–62 labels per dimension. Quality control roughly entailed removing responses that did not correspond to words and those that were derogatory in nature. The filtered labels were then converted into 35 broad topic categories: "Age", "Age related", "Ancestry", "Ear shape", "Eye color", "Eye related", "Eye shape", "Eyebrow related", "Eyebrow shape", "Face related", "Face shape", "Facial expression", "Facial hair", "Forehead shape", "Gender expression", "Hair color", "Hair length", "Hair related", "Hair texture", "Hairstyle", "Head shape", "Lip related", "Lip shape", "Mouth related", "Mouth shape", "Neck related", "Nose related", "Nose shape", "Skin color", "Skin

Table 9: Age, ancestry, gender identity, and nationality group counts for the 1,645 eligible participants who passed our quality checks and contributed 638,180 triplet judgments

Age	Count
30–39	612
40–49	373
50–120	333
20–29	327

Ancestry	Count
Europe	711
Asia	331
Americas	256
Americas and Europe	175
Africa	50
Africa and Americas	21
Asia and Europe	21
Africa, Americas, and Europe	20
Africa, Europe	16
Americas and Asia	15
Americas, Asia, and Europe	6
Europe and Oceania	6
Africa, Americas, Asia, Europe, and Oceania	5
Africa, Americas, Asia, and Europe	3
Africa, Asia, and Europe	3
Africa and Asia	2
Oceania	2
Americas, Europe, and Oceania	1
Asia, Europe, and Oceania	1

Gender identity	Count
Male	849
Female	788
Other gender identity	8

Nationality	Count	Nationality	Count
American	981	Thai	2
Indian	260	Greek	2
Brazilian	74	Colombian	2
Canadian	44	Pakistani	2
Italian	42	Bulgarian	2
English	39	Japanese	1
British	37	Puerto Rican	1
German	21	Liberian	1
French	15	Moldovan	1
Spanish	13	South Korean	1
Australian	7	Northern Irish	1
Romanian	5	Serbian	1
Venezuelan	5	Nepalese	1
Nigerian	4	Scottish	1
Chinese	4	Lithuanian	1
Welsh	4	Slovenian	1
Filipino	4	Somali	1
Armenian	4	Russian	1
Vietnamese	4	South African	1
Polish	3	Belizean	1
Irish	3	Sri Lankan	1
Turkish	3	Cameroonian	1
Portuguese	3	Belgian	1
Dutch	3	Trinidadian	1
Mexican	3	Argentine	1
Ukrainian	3	Latvian	1
Macedonian	2	Austrian	1
Brasileiro	2	Icelandic	1
Estonian	2	Vincentian	1
Jamaican	2	Tunisian	1
Kenyan	2	Taiwanese	1
Jordanian	2	Hong Konger	1
Malaysian	2	Ethiopian	1
Indonesian	2	Afghan	1
Czech	2		

related", "Skin texture", "Teeth related", and "Weight related". Figure 3 shows the top three topics for model dimensions 1 to 22, as well as the top three labels (i.e. those resulting in the highest AUCs per dimension) using CelebA, FairFace, and FFHQ-Aging annotations. Table 12 shows the age, ancestry, gender identity, and nationality group counts of the 102 participants.

Collection of dimension ratings. We tasked participants with placing 20 novel face images along each of the 22 dimensions using the unlabeled scales from the dimension labeling task. We collected 8,800 judgments in total (20 per image-dimension tuple). Participants were presented with a scale and a novel face image above the scale, and were asked: "Decide where on the scale you would place the person shown based on their similarity to the people shown beneath the scale". In addition, participants were asked to view and interact with the entire scale before making their decision. Figure 5 shows an example of a scale and a novel face image shown to participants. Table 13 shows the age, ancestry, gender identity, and nationality group counts of the 164 participants.

C Implementation details

C.1 FAX models

All images were resized to 128×128 and normalized to $[-1, 1]$. Standard data augmentation was used (horizontal mirroring and 112×112 random crops). For the unconditional FAX models, guided by the validation loss, we empirically set $\lambda_1 = 5 \times 10^{-5}$ and $\lambda_2 = 1 \times 10^{-2}$. (The conditional models used the same hyperparameters with the additional loss term's weight set to $\lambda_3 = 1 \times 10^{-4}$.) All models were optimized for 40 epochs with Adam Kingma & Ba (2014) (default parameters), learning rate 1×10^{-3}, and batch size 128 on a single Tesla T4 GPU. Post-optimization, we obtained our core set of dimensions by dispensing with those whose maximal value was close to zero, resulting in a low-dimensional space. (For the FAX trained model, this resulted in 22 dimensions.) The

Table 10: Age, ancestry, gender identity, and nationality group counts for the 355 eligible participants who passed our quality checks and contributed 24,060 triplet judgments in the "same images, novel triplets" test setting

Age	Count
30–39	143
40–49	77
50–120	76
20–29	59

Ancestry	Count
Europe	131
Asia	92
America	62
America and Europe	35
Africa and Europe	8
Africa	7
America and Asia	4
Africa, America, and Europe	4
America, Asia, and Europe	3
Africa, America, Asia, Europe, and Oceania	3
Africa and America	3
Asia and Europe	2
Africa and Asia	1

Gender identity	Count
Male	184
Female	166
Other gender identity	5

Nationality	Count	Nationality	Count
American	211	Venezuelan	1
Indian	65	German	1
Brazilian	20	Swedish	1
Canadian	12	Filipino	1
Italian	8	Estonian	1
English	6	Malaysian	1
British	6	Colombian	1
French	4	Pakistani	1
Armenian	3	Cameroonian	1
Vietnamese	2	Polish	1
Brasileiro	2	Vincentian	1
Kenyan	1	South African	1
Jamaican	1	Thai	1
Welsh	1		

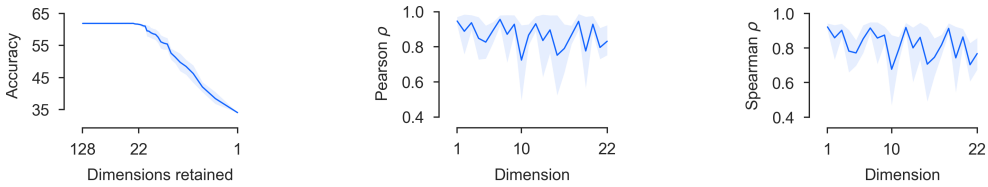

Figure 2: Left: Validation accuracy as a function of the number of dimensions retained over five independent runs (95 % CI), empirically evidencing that the judgments can be encapsulated using a limited number of dimensions. Middle and Right: Pearson and Spearman r between each of the 22 dimensions and the best "matching" dimension from each of the other four runs (95 % CI), empirically evidencing the replicability of dimensions

threshold for all models was determined based on maximizing accuracy on the validation set. Across five independent runs, the 95% confidence interval (CI) for the number of dimensions and validation accuracy were 27.4±5.8 and 61.9±0.1%, respectively. Figure 2 shows that the judgments can be encapsulated using a limited number of the available dimensions and that the 22 dimensions are approximately replicated across the independent runs.

Note that we discovered some redundancy in the dimensions, in particular 8 dimensions have a Pearson $r > 0.9$ with another dimension. The number of (near) nonzero dimensions depends on the $L1$ sparsity penalty, which must be carefully chosen: Too high of a penalty will result in the entanglement of distinct factors of variation, whereas too low of a penalty will result in repeated dimensions. We erred on the side of caution (lower penalty), since it is more difficult to localize the source of bias when factors are merged, i.e., since our main aim is auditing datasets and models.

C.2 Baseline models

All images were resized to 128×128 and normalized to $[-1, 1]$. Standard data augmentation was used (horizontal mirroring and 112×112 random crops). For preprocessing, we used the same

Table 11: Age, ancestry, gender identity, and nationality group counts for the 632 eligible participants who passed our quality checks and contributed 80,300 triplet judgments in the "novel images, novel triplets" test setting

Age	Count
30–39	252
40–49	141
20–29	126
50–120	109
18–19	4

Ancestry	Count
Europe	231
Asia	194
America	88
America and Europe	62
Africa	12
Africa, America and Europe	10
America and Asia	8
Africa and Europe	7
Africa and America	7
Asia and Europe	6
America, Asia, and Europe	2
Africa, America, Asia, and Europe	2
Asia, Europe, and Oceania	1
Africa, Asia, and Europe	1
Africa, America, Asia, Europe, and Oceania	1

Gender identity	Count
Male	350
Female	279
Other gender identity	3

Nationality	Count	Nationality	Count
American	291	Mexican	2
Indian	164	Cameroonian	1
Brazilian	51	Chinese	1
Italian	19	Filipino	1
Canadian	17	Colombian	1
English	14	Austrian	1
British	10	Estonian	1
French	8	Jamaican	1
Spanish	5	Indonesian	1
German	4	Irish	1
Vietnamese	3	Japanese	1
Venezuelan	3	Latvian	1
Kenyan	3	Lithuanian	1
Macedonian	2	Malaysian	1
Greek	2	Nigerian	1
Welsh	2	Slovenian	1
Brasileiro	2	South African	1
Bulgarian	2	Sri Lankan	1
Romanian	2	Ukrainian	1
Thai	2	Vincentian	1
Turkish	2	Afghan	1
Armenian	2		

Table 12: Age, ancestry, gender identity, and nationality group counts for the 102 eligible participants who contributed labels in the dimension naming task

Age	Count
30–39	36
40–49	26
20–29	23
50–120	16
18–19	1

Ancestry	Count
Europe	40
Asia	31
America	14
America and Europe	9
Africa	3
Asia and Europe	1
America and Asia	1
Africa and Europe	1
Africa, America, and Europe	1
Africa and America	1

Gender identity	Count
male	52
female	50

Nationality	Count	Nationality	Count
American	43	Ethiopian	1
Indian	25	Chinese	1
Brazilian	7	Vietnamese	1
English	7	Venezuelan	1
Italian	4	Icelandic	1
Spanish	2	Lithuanian	1
Canadian	2	Tunisian	1
German	2	Greek	1
French	2		

alignment procedure as in Or-El et al. (2020) to produce 128×128 face crops. Image values were normalized to $[-1, 1]$.

CelebA model. Training was performed using a batch size of 512 (across 4 Tesla T4 GPUs) with SGD for 45 epochs. The initial learning rate 0.1 was divided by 10 at epoch 15, 30, and 40. L^2 weight

Table 13: Age, ancestry, gender identity, and nationality group counts for the 164 eligible participants who contributed 8,800 triplet judgments in the dimension rating task

Age	Count
30–39	55
20–29	37
40–49	36
50–120	35
18–19	1

Ancestry	Count
Europe	67
Asia	40
America	26
America and Europe	13
Africa and America	4
Africa	4
America, Asia	3
Africa, America, and Europe	3
Europe and Oceania	2
Asia and Europe	1
Africa and Europe	1

Gender identity	Count
Female	86
Male	77
Other gender identity	1

Nationality	Count	Nationality	Count
American	70	Estonian	1
Indian	35	Vietnamese	1
Brazilian	12	Venezuelan	1
English	12	Icelandic	1
Canadian	8	Israeli	1
British	4	Japanese	1
Italian	4	Kenyan	1
German	3	Spanish	1
French	3	Thai	1
Australian	1	Trinidadian	1
Colombian	1	Greek	1

decay and SGD momentum were set to 0.0005 and 0.9, respectively. Standard data augmentation was used (horizontal mirroring and 112×112 random crops).

FairFace model. Training was performed using a batch size of 512 (across 4 Tesla T4 GPUs) with SGD for 45 epochs. The initial learning rate 0.1 was divided by 10 at epoch 15, 30, and 40. L^2 weight decay and SGD momentum were set to 0.0005 and 0.9, respectively. Standard data augmentation was used (horizontal mirroring and 112×112 random crops).

CASIA-WebFace models. We experimented with several face recognition models that differed only in terms of the loss function minimized: Softmax, ArcFace Deng et al. (2019), CosFace Wang et al. (2018), and SphereFace Liu et al. (2017). Training was performed using a batch size of 512 (across 4 Tesla T4 GPUs) with SGD for 55 epochs. The initial learning rate 0.1 was divided by 10 at epoch 15, 30, and 40. L^2 weight decay and SGD momentum were set to 0.0005 and 0.9, respectively. Standard data augmentation was used (horizontal mirroring and 112×112 random crops).

D ADDITIONAL BASELINE RESULTS

When appropriate, as baselines, we compare to supervised and self-supervised representation learning methods. Supervised models minimize a cross-entropy, ArcFace (Deng et al., 2019), CosFace (Wang et al., 2018), SphereFace (Liu et al., 2017), or triplet margin with distance swap (Balntas et al., 2016) loss. Self-supervised approaches correspond to MoCo-v2 (Chen et al., 2020), SwAV (Caron et al., 2020), or DINO (Caron et al., 2021). For data, supervised methods train on CASIA-Webface (Yi et al., 2014), CelebA (Liu et al., 2015), FairFace (Kärkkäinen & Joo, 2019), FFHQ (Karras et al., 2019; Or-El et al., 2020), or ImageNet-1K (Russakovsky et al., 2015). Self-supervised approaches learn on IG-1B (Goyal et al., 2022), ImageNet-1K (Russakovsky et al., 2015), or PASS (Asano et al., 2021). As is standard, baseline representations are extracted from the final encoder layer of a model and then normalized to unit-length. The dot product of two representations determines their similarity. Results for **Protocol 1** and **Protocol 2** are shown in Tbl. 14.

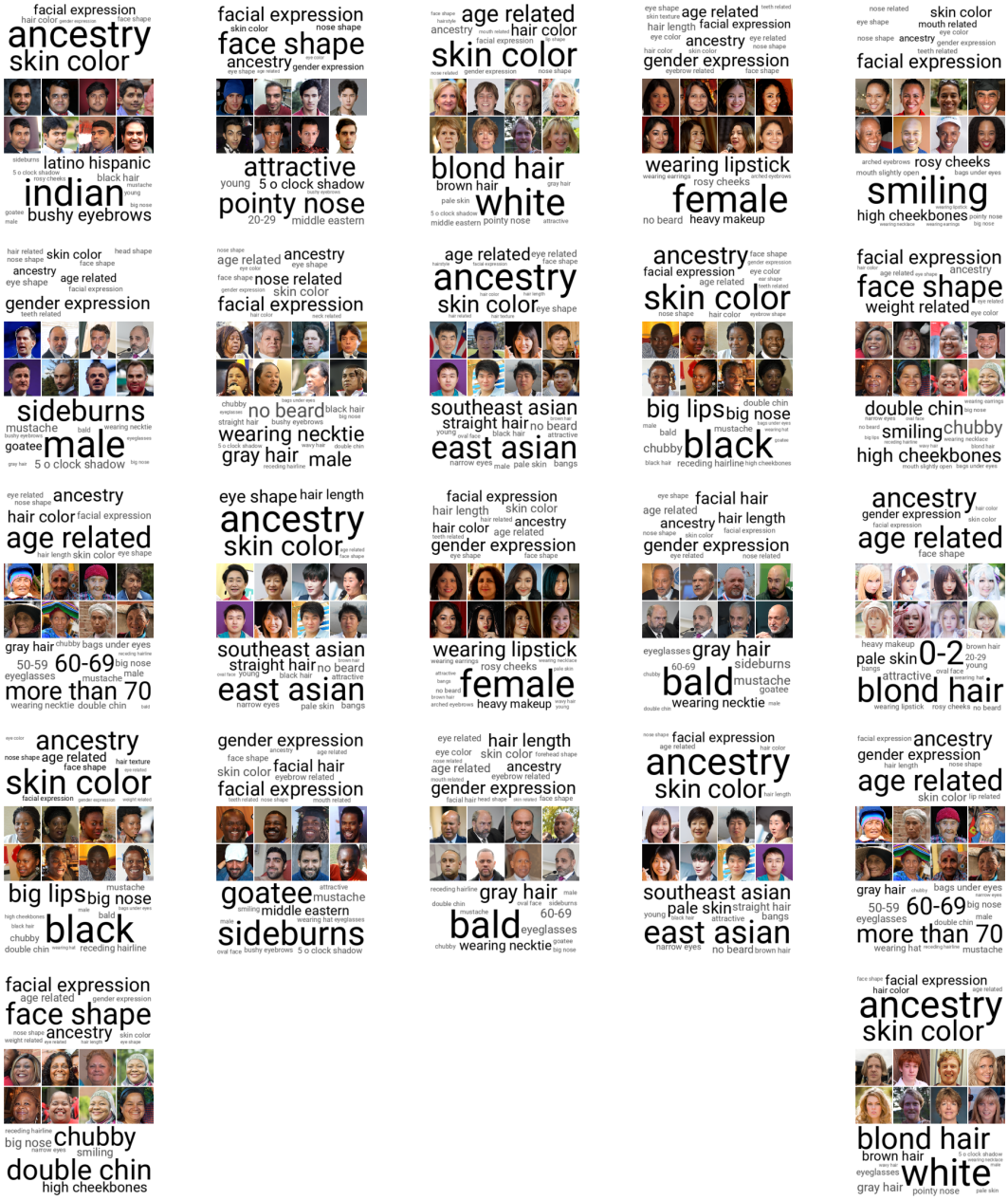

Figure 3: Dimension labeling results for the 22 dimensions. Top 8 images according to their embedding values for each dimension are shown. Above the top 8 images we show a word cloud of participant-generated topics and below the top 8 images we show a word cloud of CelebA/FairFace model-generated labels

E AUDITING MODEL BEHAVIOR

Beyond dataset auditing, as an example, we study the disparate impact of the face image restoration model PULSE (Menon et al., 2020). We select PULSE due to its much discussed racial bias (Salminen et al., 2020; Menon et al., 2020; Vincent, 2020). Let $R : x_L \mapsto R(x_L) = x'_H$ denote the PULSE model, which maps a low-resolution face image to a high-resolution face image. We denote by x_H the ground-truth high-resolution face image. Suppose z_H and z'_H denote face attribute representations of x_H and x'_H, resp. Disparate outcomes in attribute changes due to R across demographic groups may be an indicator of data and/or algorithmic bias. For face data, we use CFD (Ma et al., 2015),

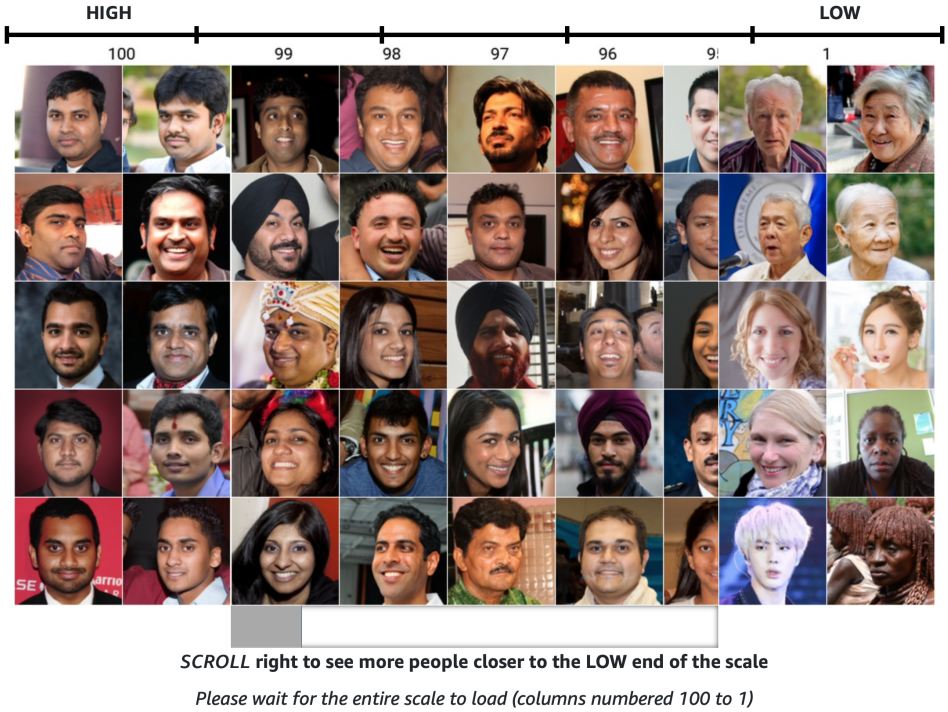

Figure 4: Dimension labeling task. Example scale shown to participants (dimension 1)

Table 14: Predicting human similarity judgments. **(Protocol 1)** Accuracy over 24,060 judgments and Spearman's r between the entropy in human- and model-generated triplet odd-one-out probabilities. **(Protocol 2)** Accuracy over 80,300 judgments and Spearman's r between the entropy in human- and model-generated similarity matrices. Note: #Labels is an approximate of the number of labels collected to create a dataset accounting for consensus.

					Protocol 1		Protocol 2	
Model	Loss/Method	#Images	#Labels	Arch.	Acc.	r	Acc.	r
ImageNet-1K	Cross-entropy	1.28M	>1.28M	RG-32Gf	46.6	0.09	41.7	0.32
FF	Cross-entropy	62K	186K–372K	R18	55.9	0.41	51.9	0.67
CA	Cross-entropy	178K	7.12M	R18	52.1	0.25	48.9	0.56
FFHQ	Cross-entropy	70K	140K	R18	50.5	0.16	47.3	0.49
CWF	ArcFace	404K	404K	R18	51.6	0.29	46.1	0.40
CWF	Cross-entropy	404K	404K	R18	48.6	0.21	43.2	0.34
CWF	SphereFace	404K	404K	R18	48.3	0.25	43.8	0.35
CWF	CosFace	404K	404K	R18	44.0	0.12	40.8	0.23
FAX-C	Conditional eq. (1)	5K	574K	R18	*67.4*	*0.68*	*61.7*	0.82
FAX-C-PH	Conditional eq. (1)	5K	574K	R18	66.5	*0.68*	61.4	0.82
FAX-U	Unconditional eq. (1)	5K	574K	R18	62.0	0.65	57.5	*0.86*
FAX-U·1/2	Unconditional eq. (1)	5K	287K	R18	61.3	0.61	55.8	0.81
FAX-U·1/4	Unconditional eq. (1)	5K	144K	R18	60.6	0.56	55.3	0.80
FAX-U·1/8	Unconditional eq. (1)	5K	72K	R18	58.6	0.47	55.0	0.79
FAX-Triplet	Triplet margin with distance swap (Balntas et al., 2016)	5K	574K	R18	60.2	0.46	52.8	0.64
ImageNet-1K	SwAV	1.28M	0	RN50-w5	43.8	0.08	41.0	0.30
IG-1B	SwAV	1B	0	RG-32Gf	47.2	0.18	44.7	0.45
IG-1B	SwAV	1B	0	RG-64Gf	48.1	0.16	44.4	0.45
IG-1B	SwAV	1B	0	RG-128Gf	46.8	0.15	42.8	0.40
IG-1B	SwAV	1B	0	RG-256Gf	47.8	0.17	43.1	0.41
PASS	MoCo-v2	1.28M	0	R50	42.3	0.09	40.5	0.27
PASS	SwAV	1.28M	0	R50	42.4	0.12	40.4	0.27
PASS	DINO	1.28M	0	ViTS-16	43.2	0.10	41.8	0.32

centering our analysis on self-identified sensitive labels: `Black`, `White`, `Indian`, and `(East) Asian`. We create low-resolution faces by downsampling from 1024×1024 to 32×32. For each group, we calculate the mean cosine similarity between each (z_H, z'_H) pair and report the min-max group ratio, i.e., *worst-case* scenario (Ghosh et al., 2021). For the CelebA and FairFace models,

Figure 5: Dimension rating task. Example scale shown to participants (dimension 1)

attribute representations correspond to unnormalized classifier predictions (logits), whereas for the FAX and CASIA-WebFace-AF models we use the face embeddings.

The min-max ratio for the FAX, CelebA, FairFace, and CASIA-WebFace-AF attribute representations are 0.78, 0.88, 0.58, and 0.98, resp. All attribute representations result in the same min-max pair, i.e., `Black`-`White`. Using FAX, CelebA, and FairFace attribute representations, we find statistically significant differences in 22/22, 36/40, and 18/18, resp., for `Black` individuals, whereas for `White` individuals, we find statistically significant differences in 16/22, 26/40, and 12/18, resp. All methods show that `Black` individuals undergo more significant attribute changes. Fig. 6 plots the mean of $z'_H - z_H$, highlighting FAX-model attribute changes per group. To ease presentation, we combined FAX-model dimensions measuring the same attribute. First, `Black` individuals have highly reduced values in the `Black` dimension and magnified values in the `White` dimension, corroborating previous qualitative findings (Vincent, 2020). Second, the impact of PULSE on `White` individuals largely centers on changes in face width and length, representing a novel insight. While neither *type* of modification is desirable, the inadvertent erasure of minority groups is extremely harmful. FAX dimensions are therefore useful as a tool for gaining insight into the behavior of face-based models.

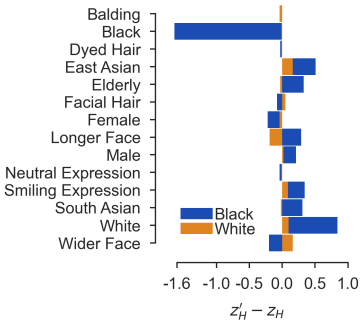

Figure 6: FAX dimensions depicting changes due to the PULSE restoration process on self-identified Black and White groups

F NAIVE PREDICTIONS

Given that previous works equate diversity with parity across demographic subgroup distributions (Keswani & Celis, 2021; Kay et al., 2015; Bolukbasi et al., 2016; Schwemmer et al., 2020), we test whether categorical labels corresponding to gender, race, and age are predictive of human judgments of similarity. We only consider triplets that are composed of two faces that share the same attribute(s), whereas the third face does not. We use a parameterless naive model that always predicts the third face as the odd-one-out. **Protocol 1** and **Protocol 2** results are shown in Fig. 7. We note that when two of the faces have at least the same gender and race attribute labels that the naive predictor starts to approach the predictive accuracy of our models. This is unsurprising, since the more attributes two faces have in common as compared to the third face, the more likely the third face will be selected as the odd-one-out. Note that this strategy cannot explain judgments for triplets composed of faces from the exact same subgroup.

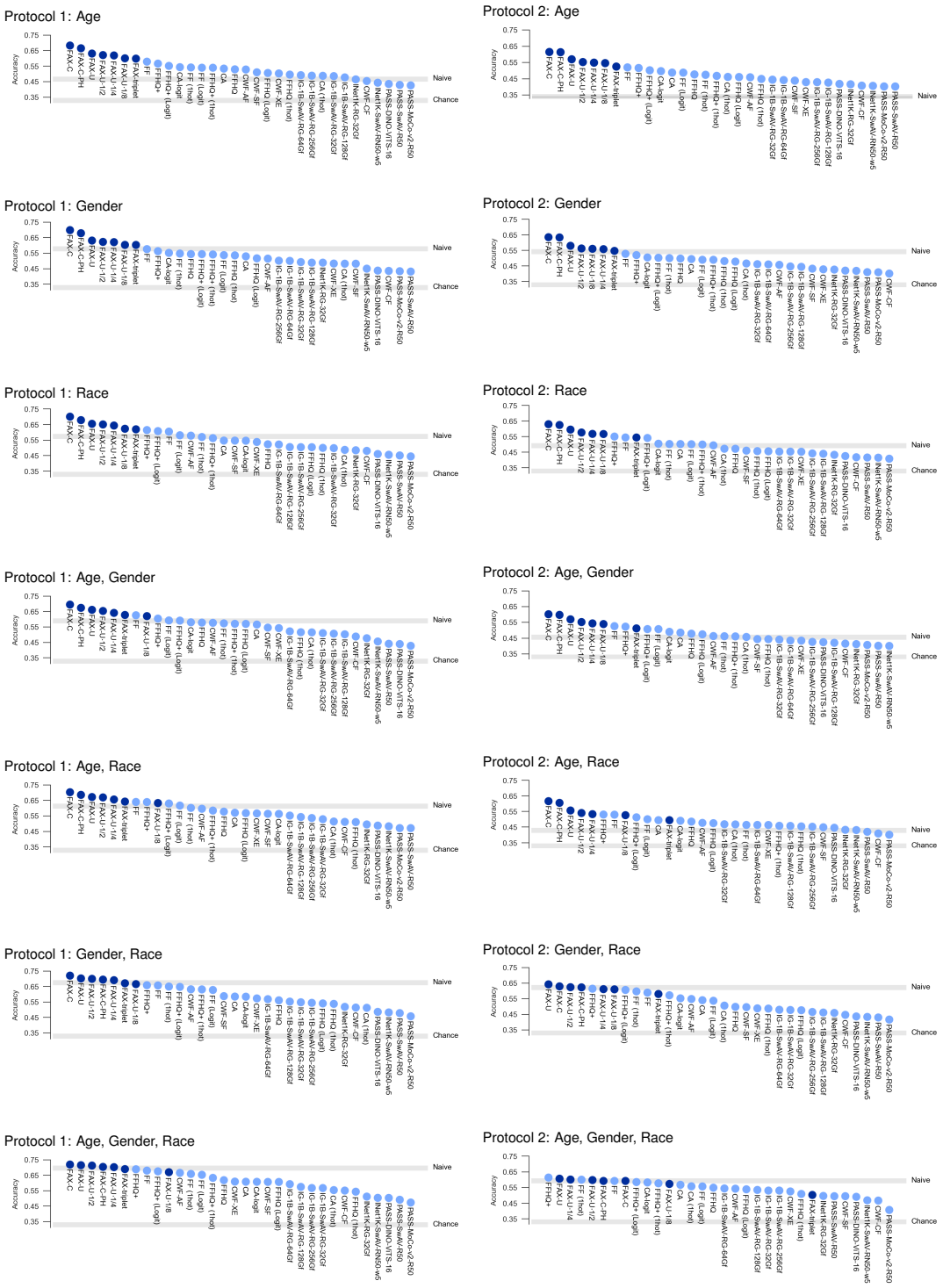

Figure 7: **Naive predictor.**

G DATASET DOCUMENTATION FOR FAX

Below we follow the datasheets for datasets documentation guidelines Gebru et al. (2018), answering questions with respect to FAX's motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance.

G.1 MOTIVATION

For what purpose was the dataset created? Biases in human-centric computer vision models are often attributed to a lack of sufficient data diversity, with many demographics insufficiently represented. However, auditing datasets for diversity can be difficult, due to an absence of ground-truth labels of relevant features. Few datasets contain self-identified demographic information, inferring demographic information risks introducing additional biases, and collecting and storing data on sensitive attributes can carry legal risks. Moreover, categorical demographic labels do not necessarily capture all the relevant dimensions of human diversity that are important for developing fair and robust models. One can instead *implicitly* uncover a set of continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. To do so, we propose a novel dataset of 638,180 human judgments of face similarity (FAX). Notably, each judgment is associated with the identifier and demographic attributes of the annotator who made the judgment. We demonstrate the utility of FAX induced embedding spaces for predicting face similarity judgments, collecting continuous face attribute values, comparative dataset diversity auditing, and studying annotator bias.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was constructed by the authors.

Who funded the creation of the dataset? The dataset was created for research purposes only and was funded entirely by the authors.

G.2 COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The dataset consists of several JSON files, containing information such as randomly generated participant identifiers, triplet image identifiers, and odd-one-out triplet judgments. JSON format data is provided for each participant and contains their demographic information.

How many instances are there in total (of each type, if appropriate)? For training and validation, FAX contains 638,180 triplet-judgment unique tuples (one unique judgment for each unique triplet). For testing generalization, we include two additional subsets: **(a)** 1,000 unique triplets, where each triplet has 22–25 judgments from distinct participants (total: 24,060 triplet-judgment tuples); **(b)** 27,720 unique triplets, where each triplet has 2–3 judgments from distinct participants (total: 80,300 triplet-judgment tuples). We provide demographic information for each of the 1,663 participants that contributed judgment responses.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The training and validation sets represent 0.003% of all possible triplets that could be sampled from 4,921 images. The 4,921 images are a subset of the 70,000 images contained in FFHQ Karras et al. (2019). We used FFHQ-Aging annotations (Or-El et al., 2020), obtained using Face++, for absolute head yaw angle, absolute head pitch angle, eyewear, and eye occlusion annotations to prune the 70,000 images. The purpose of these constraints was to reduce the influence of head pose on human behavior and to better allow for eye comparisons. The stimulus set of 4,921 representative examples were then sampled from the remaining images such that each intersectional subgroup based on perceived binary gender expression, age group, and ethnicity would be included. We only considered face images with an apparent age greater than 19 years old. Ethnicity was estimated by using a model traned on the FairFace dataset (Kärkkäinen & Joo, 2019). Perceived binary gender expression and apparent age group were obtained from FFHQ-Aging crowdsourced annotations. To mitigate biasing our set toward stereotypical faces, we randomly sampled images from the 56 possible intersectional subgroups rather than selecting the most confidently predicted or annotated. Additional images (total: 56) used for testing (disjoint from the 4,921 stimuli set images) were also sampled in the same manner. In this case, all possible triplets were sampled. Note: No images are contained in our dataset, the images are hosted on NVIDIA

Corporation's Google Drive, please see the original FFHQ dataset Karras et al. (2019) for more details: https://github.com/NVlabs/ffhq-dataset.

What data does each instance consist of? Each instance in the training and validation set consist of a randomly generated participant identifier, triplet image identifiers, an odd-one-out triplet judgment, and a training or validation split label. The test set instances consist of the same information, however they do not have training or validation split labels, since they should be used for testing only.

Is there a label or target associated with each instance? Each instance is labeled with the odd-one-out triplet judgment, as well as a randomly generated participant identifier (i.e., the participant who generated the judgment). Each participant's identifier can be mapped to their demographic information (age, nationality, ancestry, and gender identity) using the participant metadata JSON file. Note: Participant identifiers were randomly generated and do not correspond to their true Amazon Mechanical Turk worker identifiers.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? We make explicit the relationship between a judgment and the participant who generated the judgment.

Are there recommended data splits (e.g., training, development/validation, testing)? The dataset is split into training, validation, and testing splits. The training and validation data are contained in the same JSON file, where each instance has a training or validation split label. The test data subsets are provided in separate JSON files.

Are there any errors, sources of noise, or redundancies in the dataset? No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The triplet judgments pertain to images hosted on NVIDIA Corporation's Google Drive, please see the original FFHQ dataset Karras et al. (2019) for more details: https://github.com/NVlabs/ffhq-dataset.

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)? The dataset includes the demographic information (age, nationality, ancestry, and gender identity) of each participant. Participants who contributed judgments consented to the collection and use of this data as well as the information being made public.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. FAX contains judgments gathered from human participants about the similarity of human faces in images.

Does the dataset identify any subpopulations (e.g., by age, gender)? Each judgment in the dataset is explicitly linked to the participant who generated the judgment. Moreover, each participant's age, nationality, ancestry, and gender identity is available, which were provided when obtaining informed consent. Note: Participant identifiers were randomly generated and do not correspond to their true Amazon Mechanical Turk worker identifiers.

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? We replaced each participant's Amazon Mechanical Turk worker identifier with randomly generated identifiers. Therefore, it is not possible to directly identify individuals.

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)? Each instance in the dataset is associated with the participant who generated the judgment and each participant is associated with their age, nationality, ancestry, and gender identity, which were provided when we obtained informed consent.

G.3 COLLECTION PROCESS

How was the data associated with each instance acquired? All judgments were collected on Amazon Mechanical Turk directly from participants.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? The Amazon Mechanical Turk API was used to deploy human intelligence tasks and collect responses.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Based on estimated face image labels, a triplet can comprise 1–2 binary gender expression group(s), 1–3 age group(s), and 1–3 ethnicity group(s) resulting in 18 triplet *types* which were uniformly sampled. Ethnicity was estimated using a model trained on the FairFace dataset (Kärkkäinen & Joo, 2019), whereas perceived binary gender expression and apparent age group were obtained from FFHQ-Aging crowdsourced annotations Or-El et al. (2020).

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)? The authors and participants were involved in the data collection process. For the prescreening survey (which included the consent form), participants were paid a nominal fee of 0.01 USD. For all other tasks, participants were compensated at a rate of 15 USD per hour, regardless of whether their work passed our quality checks.

Over what timeframe was the data collected? Data was collected on Amazon Mechanical Turk between 6 November 2021 and 20 February 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The judgments and participant demographic information were collected directly from the participants on Amazon Mechanical Turk.

Were the individuals in question notified about the data collection? The participants who contributed judgments were provided with an "Information Sheet and Consent Form" as well as a separate "Notification and Consent for Collection and Use of Study Data Form".

Did the individuals in question consent to the collection and use of their data? Yes. Participants who contributed judgments consented to the collection and use of the following Study Data: Amazon Mechnical Turk worker identifier, age, gender identity, nationality, and ancestry, as well as any other data provided while performing human intelligence tasks. Note: Participant identifiers were randomly generated and do not correspond to their true Amazon Mechanical Turk worker identifiers.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Participants were informed that their Study Data is regulated in the European Economic Area under the General Data Protection Regulation (the "GDPR") and in the UK under the UK GDPR. GDPR gives them rights relating to their Study Data, including the right to: (a) Access, correct or withdraw their Study Data; (b) Restrict the types of activities that can be done with their Study Data; (c) Object to using their Study Data for specific types of activities; (d) Withdraw their consent to use their Study Data for the purposes outlined in the original consent form. In addition, residents in the EU, EEA or the UK were informed who to contact in the event that they wish to exercise their rights under the GDPR or UK GDPR.

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? A privacy impact assessment was conducted.

G.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)? To control for quality, according to the criteria outlined, we excluded triplet judgments obtained from any participant who provided overly fast, deterministic, or incomplete judgments.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes. The authors retain a copy of the raw data.

Is the software that was used to preprocess/clean/label the data available? No software was used per se. The quality controls used to clean the data are conditional statements.

G.5 Uses

Has the dataset been used for any tasks already? We have shown that the dataset can be used to implicitly uncover human-interpretable dimensions corresponding to perceptual face-varying axes, aligned with the human mental representational space of faces. We have also shown that the learned embedding dimensions may be used to audit the visual diversity of unlabeled datasets and the behavior of pretrained face-based models. Furthermore, since each judgment in the dataset is associated with the generating participant's identifier and their demographic attributes, we have shown that this permits the identification and study of annotator bias.

Is there a repository that links to any or all papers or systems that use the dataset? We will list these on the dataset's GitHub repository which will be made public in August 2022.

What (other) tasks could the dataset be used for? The dataset can be used to mitigate bias during model training.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The triplet judgments pertain to images from Flickr. While we made attempts to ensure the subset of images that we collected judgments for were demographically diverse, the subset is not representative of the entire population. Moreover, although the images have permissive licenses that allow free use, redistribution, and adaptation for non-commercial purposes, image subjects and/or image creators are free to revoke the use of their data. When this is the case, we will remove any affected judgments from our dataset. The same applies when a participant who contributed judgments withdraws their consent.

Our participant pool was demographically imbalanced, which is an unfortunate consequence of using Amazon Mechanical Turk. Therefore, the criteria used by our participants and the importance of each criterion when performing judgments does not necessarily generalize to demographically dissimilar participants.

Are there tasks for which the dataset should not be used? The dataset was created for research purposes only, in particular for those wishing to study the human mental representational space of faces. The dataset is not intended for, and should not be used for, development or improvement of facial recognition technologies.

G.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? Yes, the dataset will be made publicly available on Zenodo: https://zenodo.org/.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? The dataset will be hosted on Zenodo: https://zenodo.org/.

When will the dataset be distributed? The dataset will be made publicly available in August 2022 on Zenodo: https://zenodo.org/.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Creative Commons BY-NC-SA: https://creativecommons.org/licenses/by-nc-sa/4.0/.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? We are not aware of any controls or restrictions.

G.7 MAINTENANCE

Who will be supporting/hosting/maintaining the dataset? The dataset is supported by the authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Contact information removed for peer review.

Is there an erratum? In the event that any errors are found, an erratum file will be added to the dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? The dataset will be versioned.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? Participants were informed that their public Study Data (age, gender identity, nationality, and ancestry) would be kept indefinitely. However, participants are free to withdraw their consent at any time and for any reason.

Will older versions of the dataset continue to be supported/hosted/maintained? Older versions of the dataset will continue to be hosted, except when a participant withdraws their consent or when an image subject and/or image creator revokes the use of their data. In this case, older versions containing such data instances will be removed and a new version will be created.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? External parties are free to extend the dataset, or contact us directly if they wish to collaborate on extending the dataset.

G.8 AUTHOR STATEMENT OF RESPONSIBILITY

The authors confirm all responsibility in case of violation of rights and confirm the license associated with the dataset.