AI and Ethics (2021) 1:249–260
https://doi.org/10.1007/s43681-020-00035-y

ORIGINAL RESEARCH

A set of distinct facial traits learned by machines is not predictive of appearance bias in the wild

Ryan Steed[1] · Aylin Caliskan[2]

Received: 14 October 2020 / Accepted: 3 December 2020 / Published online: 12 January 2021

Abstract

Research in social psychology has shown that people's biased, subjective judgments about another's personality based solely on their appearance are not predictive of their actual personality traits. But researchers and companies often utilize computer vision models to predict similarly subjective personality attributes such as "employability". We seek to determine whether state-of-the-art, black box face processing technology can learn human-like appearance biases. With features extracted with FaceNet, a widely used face recognition framework, we train a transfer learning model on human subjects' first impressions of personality traits in other faces as measured by social psychologists. We find that features extracted with FaceNet can be used to predict human appearance bias scores for deliberately manipulated faces but not for randomly generated faces scored by humans. Additionally, in contrast to work with human biases in social psychology, the model does not find a significant signal correlating politicians' vote shares with perceived competence bias. With Local Interpretable Model-Agnostic Explanations (LIME), we provide several explanations for this discrepancy. Our results suggest that some signals of appearance bias documented in social psychology are not embedded by the machine learning techniques we investigate. We shed light on the ways in which appearance bias could be embedded in face processing technology and cast further doubt on the practice of predicting subjective traits based on appearances.

Keywords Appearance bias · Face recognition · Computer vision · Machine learning

1 Introduction

Researchers and the public have raised concerns about the use of face detection, face recognition, and other facial processing technology (FPT)[1] in applications such as police surveillance and job candidate screening due to its potential for bias and social harm [6, 29, 34, 39].[2] For example, HireVue's automated recruiting technology uses a candidate's appearance and facial expression to judge their fitness for employment [15], and the company 8 and above uses video interviews to construct candidate "blueprints," which include estimated personality traits such as openness, warmness, and enthusiasm [33]. If a surveillance or hiring algorithm learns subjective human biases from training data, it may systematically discriminate against individuals with certain facial features. We investigate whether industry-standard face recognition algorithms can learn to make biased, stereotypical trait judgments about faces based on human participants' perception of personality traits from faces. Quick trait inferences should not affect important, deliberate decisions [45], but humans do display first impression trait biases [40] and those inferences could affect human judgments of other subjective traits like "employability" or "attractiveness" that algorithms are actively designed to mimic [15]. If off-the-shelf FPT can learn biased trait inferences from faces and their labels, then application domains using FPT to make decisions are at risk of propagating harmful prejudices.

✉ Ryan Steed
ryansteed@cmu.edu

Aylin Caliskan
aylin@gwu.edu

[1] Heinz College of Information Systems & Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA

[2] Department of Computer Science, George Washington University, Washington, DC, USA

[1] Following [34], we use the term facial processing technology (FPT) to refer to a broad class of applications that rely on representations of individual's facial characteristics, including face detection, face analysis, and face recognition.

[2] In response to these results and major activist efforts, Amazon, Microsoft, and IBM recently placed temporary moratoriums on some FPT products for government surveillance [14].

Because the predictions made by machine learning models depend on both the training data and the annotations used to label them, systematic biases in either source of data could result in biased predictions. For instance, a dataset on employment information designed to predict which job candidates will be successful in the future might contain data regarding mainly European American men. If such a dataset reflects historical injustices, it is likely to unfairly disadvantage African American job candidates. Moreover, annotators could introduce human bias to the dataset by labeling items according to their implicit biases. If annotators for a computer vision task are presented with a photo of two employees, they might label a woman as the employee and the man standing next to her as the employer or boss. Such embedded implicit or sociocultural bias leads to biased and potentially prejudiced outcomes in decision-making systems.

In computer vision, models used in face detection or self-driving cars have been proven biased against genders and races [7, 46]. Some examples of racial and gender biases include gender classifications made by automated captioning systems and contextual cues used incorrectly by visual question answering systems [17, 50, 25]. These algorithms are actively used in self-driving cars [9], surveillance [23], anomaly detection [24], military drones [30], and cancer detection [4]. But while many biases are explicit and easily detected with error analysis, some "implicit" biases are consciously disavowed and are much more difficult to measure and counteract. Often, these biases take effect a split second after perception in human judgment. These biases are often quantified by implicit association tests [10, 11]. Computer vision models do embed historical racial or gender biases, but can they also embed these first-impression appearance biases documented in social psychology [43]?

In this study, we investigate whether biases formed during the first impression of a human face can be learned by industry-standard face recognition models. Like implicit biases, "first impression" appearance biases are split-second trait inferences drawn from other people's facial structure and expression [40, 45]. Todorov et al. [41] characterize first impression bias as unreflective and sometimes unconscious. We consider six types of subjective personality trait inferences drawn from faces, each measured in controlled laboratory experiments [43, 45]: attractiveness, competence, extroversion, dominance, likeability, and trustworthiness.[3] In a rational world, these physiognomic stereotypes [16], may seem unlikely to influence deliberate decisions, but appearance biases have been shown to predict numerous external outcomes, including election results [3, 41], income [13], economic decisions [35, 47], and military rank [27]. Notably, appearance bias is not known to be predictive of any objective measure of ability, performance, or personality [41], and empirically they are often wrong about the people they stereotype [21, 27, 49].

Despite the fact that appearance biases are neither causally linked to nor predictive of actual personality traits, researchers have built machine learning models to predict appearance bias from faces [37, 48]. Likewise, HireVue and other companies still advertise predictive models for other subjective attributes such as "employability" trained on historical data with historical biases [15]. We seek to determine whether general, industry-standard face representations can be used to accurately predict subjective, human trait inferences. If so, then face processing technology is at risk of propagating trait inferences embedded in labeled training data. If not, the practice of predicting subject trait inferences and other related personality attributes is even more dubious.

In this paper, we make several contributions towards understanding appearance bias in FPT. First, we design a transfer learning method for extracting general-purpose face representations suitable for state-of-the-art face processing applications. Second, we train our model on computer-generated faces manipulated to display certain personality traits, including not only Caucasian but also Asian and Black faces. Third, we find that while our model is quite good at predicting perceived trait scores for faces produced by [43]'s computational model of appearance bias, it fails to consistently predict perceived trait scores for randomly generated faces. Additionally, while it has been shown that the human perceptions of the competence of political candidates are correlated with election outcomes [3, 41], our model's competence scores do not achieve the same predictive validity. Our experimental results and additional interpretability analysis suggest that generalized representations for face recognition are not suitable for learning subjective biases.

1.1 Related work

There is a wealth of literature measuring the stereotypes perpetuated by image classifiers and other machine learning models, from search results to automated captioning [17, 19, 22]. Previous applications of unsupervised machine learning methods demonstrated the existence of social and cultural biases embedded in the statistical properties of language, but little research has been conducted with respect to the biases in transfer learning models for faces or people and even less attention has been paid to the intersection of machine learning and first appearance bias [8, 44]. [18] review the use of computer vision to anticipate personality traits. [48] use a novel long-short term memory (LSTM) approach to predict

[3] These trait inferences are neither objective nor reified. We examine people's subjective perception of personality traits in other face, which amount to biases and prejudices. Real-world outcomes associated with these subjective trait ratings are the result of social biases, not objective ability or personality.

first impressions of the Big Five personality traits after 15 s of exposure to various facial expressions.

Most notably, [37] also train a model on a subset of the computer-generated faces produced by [43]. They learn to predict subjective trustworthiness ratings with facial action units, or facial configurations such as smiling and frowning, and use their model to analyze the evolution of trustworthiness in portraiture. The authors claim that their model can be used to predict trustworthiness for selfies and historical portraits, but the correlation between their model's predictions and subjective ratings from human annotators in external datasets is low. We clarify their results with several modifications: first, we train on Black and Asian faces, in addition to Caucasian faces; second, we use transfer learning to obtain more generalized face representations; and third, we extract representations of the entire face to capture biases related to face structure and color, not just facial actions such as smiling and frowning.

There is a serious concern that face recognition and face modeling techniques may propagate cognitive and historical biases entrenched in human annotations and model design. Our study investigates part of this concern: we evaluate whether first impression appearance biases can be learned with off-the-shelf face processing technology. Can we observe the same biased effects in real-world datasets with a predictive model?

2 Data

To test whether first impression trait inferences can be learned from facial cues like the ones in Fig. 1, we experiment with datasets of computer-generated faces developed to represent appearance bias in two psychological studies (Table 1) [31, 43]. All the datasets used in our experiments can be obtained from the original authors at http://tlab.princeton.edu/databases. These data come from a series of studies in which [43] argue that computational models are the best tools for identifying the source of first impressions of facial features. In each study, human participants are shown a face for less than a second and then asked to rate the degree to which it exhibits a given trait (trustworthiness, competence, etc.) on a 9-point scale. Each face has a neutral expression, is hairless and is centered on a black background. The faces were generated with FaceGen, which uses a database of laser-scanned male and female human faces to create new, unique faces.[4]

Together, these two sets provide a labeled benchmark for first impression, appearance-based evaluations of personality traits by human participants. One drawback to this

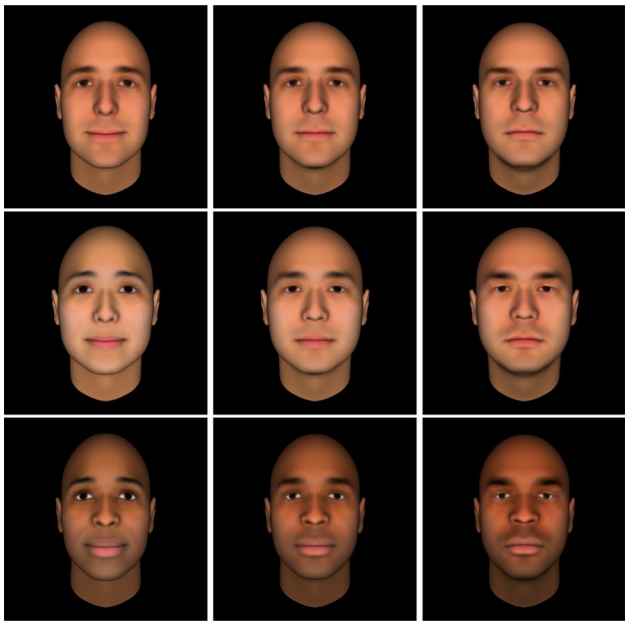


Fig. 1 Faces (center) manipulated to appear 3SD more (left) and 3SD less (right) trustworthy than the average face [43]

training set is that because the computer-generated faces do not have hair and other accessories like make-up and glasses, our results may not generalize to images of real people on the web. Unfortunately, there are no large, publicly available datasets of experimentally validated trait judgments of real faces.

2.1 Randomly generated faces

The first dataset (300 Random Faces) includes 300 computer-generated, emotionally-neutral Caucasian faces created with FaceGen, a face generation software (Fig. 1). Though the face structures are gender-neutral, participants may still perceive bald faces as male [43]. In a controlled laboratory setting, [43] asked 75 Princeton University undergraduates to judge each face from this dataset on attractiveness, competence, extroversion, dominance, likeability, and trustworthiness [31, 42]. Here, the ground-truth labels are the trait scores provided by the study participants. Ideally, we would train on ground-truth labels for a larger number of randomly generated faces and for non-Caucasian faces, but none are available. Using only 300 randomly generated faces and 75 base faces may limit generalization to different types of faces. We leave additional data collection to future work.

2.2 Faces manipulated along trait dimensions

For the second dataset (Maximally Distinct Faces) [43] select 75 "maximally distinct" faces from a random sample of 1000 randomly generated Caucasian, East Asian,

[4] https://facegen.com/.

Table 1 Sets of computer-generated faces with subjective human trait judgments

Name	Description	Race(s)	# Faces	Source
300 Random Faces	Randomly generated faces.	Caucasian	300	[31, 43]
Maximally Distinct Faces	Faces manipulated to exhibit a certain personality trait, according to subjective human judgments.	Caucasian, East Asian, Black	1875 (5 traits × 5 degrees × 75 identities)	[31, 43]
Politicians	Faces of US Senate, House, and Gubernatorial candidates, 1995–2008.	Any	543 (246 Gubernatorial, 297 Senate)	[3, 41]

and Black faces. From this random sample of base faces, additional faces with maximally distinct perceived appearance bias were constructed as follows: using principal components analysis, the authors reduced the 3D FaceGen polygonal model that represents each base face to a 50-dimensional Euclidean vector space. Specifically, each component in the shape vector corresponds to a linear change in the positions of the vertices that structure the face [31]. Oosterhof and Todorov [31] then find the best linear fit of the mean empirical trait judgments from 300 Random Faces as a function of this shape vector. If $F \in \mathbb{R}^{50\times300}$ is a matrix of the shape vectors representing 300 Random Faces with trustworthiness judgments $r \in \mathbb{R}^{300}$, then the optimal trustworthy vector is simply $t = F \cdot r$. Then a face with shape vector α can be manipulated to appear δ SD more trustworthy with the new vector $\alpha' = \alpha + \delta \cdot \hat{t}$, where $\hat{t}$ is the normalized trustworthiness gradient vector. This method leverages the ground-truth subjective trait judgments to compute the optimal direction in which to alter the subjective trustworthiness - or another trait - of a randomly generated face in FaceGen.

Todorov et al. [43] use this method to generate faces that vary along each trait dimension to produce a set of faces to elicit a trait inference $-3, -2, -1, 0, 1, 2$, and 3 SD from the mean -25 variations in total for each of the 75 faces, resulting in a total 1875 labelled faces for each trait (Fig. 1). These manipulations are not necessarily related to facial expression [31]. Though the perturbations themselves are not psychologically meaningful and do not deliberately correspond to any particular facial features or expressions, these manipulations tend to produce faces that vary noticeably along the trait dimensions (Fig. 1). Each face was presented to 15 different Princeton university students for subjective scoring on the same 9-point scale used in Random Faces. These scores were validated for interrater reliability (using Cronbach's α for all average trait ratings) and explained variance when regressed on the standard deviation scores targeted by the face-generation model; studies with human participants confirm that the manipulated faces do on average alter the subjective appearance of a given face by δ SD [43]. Faces produced with the maximally distinct method are reliable indicators of human trait judgments.

Since the authors show that there is a high degree of correlation between the average human trait score and the target SD, and validation scores are not available for every image in the dataset, we use the target SD scores as labels for training. So that the ground-truth labels for both Random Faces and Maximally Distinct Faces are scaled identically and can be trained on simultaneously, we convert the raw 9-point scale used for the Random Faces labels to standard (z-) scores such that both sets of labels are measured in terms of standard deviation from the mean.

2.3 Real faces

We also validate our model on a small set of real (not computer-generated) faces. The Politicians dataset includes the faces of US Senate, House, and Gubernatorial candidates from 1995 to 2008, evaluated by human participants on the basis of apparent competence [41]. Todorov et al. [41] and Ballew and Todorov [3] show that 1-s inferences of politicians' competence based only on facial appearance are linearly related to their margin of victory. Participants were presented with the winner and runner-up of each election and asked to judge which person was more competent (without knowing the result), in binary and on the same 9-point scale used in 300 Random Faces. Participants were more likely to choose the winner [3, 41]. Other trait judgments are also included in the dataset, but there is no significant effect for traits other than competence [41].

3 Approach

We construct a transfer learning model to leverage face representations extracted from a pre-trained, state-of-the-art face recognition model (Fig. 2). Since we are testing whether standard industry methods are capable of learning trait inferences, we use popular industry techniques for pre-processing and modeling. First, according to the state-of-the-art, we

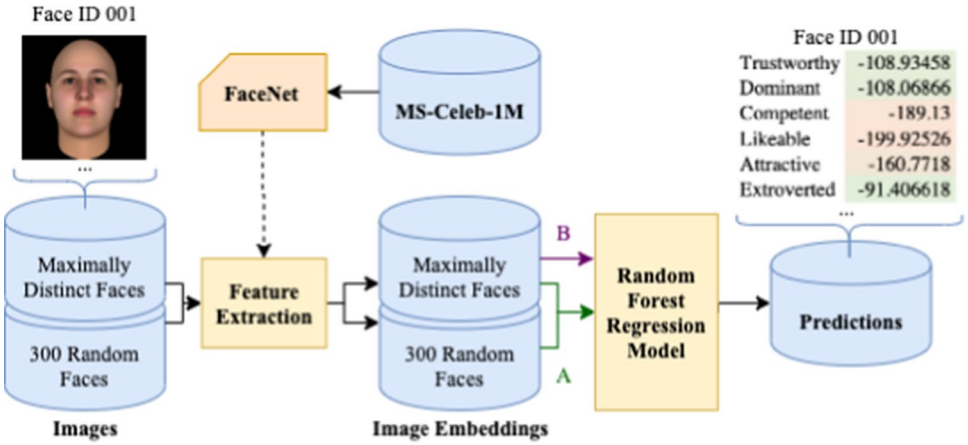


Fig. 2 A transfer learning model trained on subjective trait scores. FaceNet, pre-trained on the MS-Celeb-1M benchmark dataset, extracts embeddings for each face. In **Experiment A**, a random forest regression model is trained on feature embeddings from the set of faces manipulated to be maximally distinct and the set of randomly generated faces with human scores. **Experiment B** compares these two sets of training images with a regression trained only on the randomly generated faces

crop and align every face with pose estimation to ensure the faces have similar size, shape, and rotation [20]. By cropping out the bald head, we also make the computer-generated images seem more gender-neutral. Then, from the final layer of FaceNet, a popular open-source Inception-ResNet-V1 deep learning architecture, we extract a standard 128-dimensional feature vector from the pixels of each transformed image [38]. For thousands of images, extraction takes minutes. Rather than train FaceNet from scratch, we utilize a model with weights pre-trained on the MS-Celeb-1M dataset, a common face recognition benchmark [12], downloaded from https://github.com/davidsandberg/facenet. We chose this transfer learning approach to mitigate the fact that our dataset is entirely computer-generated: by using a model pre-trained on real faces, we can extract features more similar to features from images in the wild. MS-Celeb-1M contains 10 million images of one million celebrities and was one of the largest publicly available face recognition benchmark datasets, making it a popular choice for transfer learning face recognition models, before Microsoft took it down in 2019 [32]. Notably, MS-Celeb-1M was reportedly used to train controversial mass surveillance algorithms in China [28]. By using a pre-trained model for feature extraction, we imitate feature processing techniques used commonly in black box industry models. The FaceNet model (over 10 thousand stars on Github), and similar architectures such as OpenFace (over 13 thousand stars on Github), are used by software developers, researchers, and industry groups [2, 38].

After feature extraction, we train six random forest regression models[5] to predict appearance bias for each of the six traits measured: attractiveness, competence, dominance, extroversion, likeability, and trustworthiness. The human participants' trait scores, multiplied by 100 for readability, serve as the ground-truth labels. The random forest includes 100 weak learners with no maximum depth, a minimum split size of two, and mean-squared-error split criterion. Data and code used to produce the figures, tables, and pre-trained model (Fig. 2) in this work are available at https://github.com/anonymous/repo.

4 Experiments and results

4.1 Learning appearance bias

We validate our model's ability to learn human appearance bias scores under several different experimental conditions. In general, we find that our model is capable of learning appearance bias from manipulated faces with a high degree of accuracy, but fails to make accurate predictions for randomly generated faces.

Experiment A To test how well the random forest regression model learns appearance bias from both sets of labeled faces (randomly generated and computer-generated), we shuffle the image embeddings extracted with FaceNet such that the 300 random faces and maximally distinct faces are mixed. The target labels are the original appearance bias measurements provided by human participants. Splitting the training data into ten equal folds, we do the following for each fold: (1) train the regressor on the other nine partitions; (2) record and plot appearance bias predictions for the current partition. Once all ten partitions are processed, each image has a corresponding vector of predicted appearance bias scores, one for each trait measured. Table 2 displays goodness-of-fit and correlation statistics from the cross-validations for regressions on all six traits measured. For reference, [37], who train a model on a subset of Maximally Distinct Faces to predict trustworthy ratings, report

[5] We also tested an SVM model (with RBF kernel) and a logistic regression model (l2 penalty, $C = 1.0$). The SVM performed about three points worse in cross validation than the logistic regression or random forest models. We chose the random forest model because (1) with bootstrap aggregating and random feature selection it tends not to overfit [5], and (2) parallelization for quicker execution.

(a) Model A, trained on both datasets.

(b) Model B, trained on only the maximally distinct faces and tested on the 300 random faces.

Fig. 3 Fit line and scatter plot of actual "trustworthiness" impressions against 10-fold cross-validated predictions for models trained on both sets of computer-generated faces (**a**) or just the randomly generated faces (**b**)

Table 2 Correlation of actual and predicted appearance biases

#	Traits	Attractive	Competent	Dominant	Extroverted	Likeable	Trust
A	ρ	0.99	0.99	0.99	0.98	0.99	0.98
	p value	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$
	RMSE	30.3	33.3	27.6	36.6	33.4	35.9
B	ρ	0.30	0.29	0.72	0.25	0.17	0.13
	p value	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-16}$	$< 10^{-4}$	$< 10^{-2}$	0.028
	RMSE	134.2	152.2	141.0	105.9	134.5	144.8

Pearson's correlation coefficient ρ and root mean square error (RMSE) for regression predictions. In Experiment **A**, a random forest regression is fitted on both sets of faces and predictions are produced by 10-fold cross validation; in **B**, the regression is fitted on maximally distinct faces and tested on randomly generated faces. p values are from the correlation t test of $H_0 : \rho = 0$

significant correlation coefficients of $\rho = 0.85$ for trustworthiness and $\rho = 0.86$ for dominance for cross-validation on a held-out test set of maximally distinct faces. In contrast, with 10-fold cross validation, we achieve significant correlation coefficients of $\rho = 0.98$ and $\rho = 0.99$, respectively. Likely, the higher accuracy is a result of our larger training set (Safraet al. [37] train on only the Caucasian maximally distinct faces).

Notably, our approach learns appearance bias to a high degree of precision for the maximally distinct faces ($\rho = 0.99$), but the accuracy drops on when predicting human trait scores for randomly generated faces (Fig. 3). The model performs poorly on randomly generated faces even when randomly generated faces are included in the training set, suggesting there is either no consistent signal in trait scores of random faces, that the signals are too complex to be learned by this model, or that the transfer learned face representations used for training do not contain useful information for predicting trait inferences. The former explanation seems unlikely; human participants tended to agree on trait scores for the random faces: the interrater reliability for Random Faces was $\alpha = 0.84$, roughly the same as for Maximally Distinct Faces. In other words, human participants tended to predict the judgments of other human participants, so there is some signal to be modelled). For the remainder of the paper, we will explore the second two explanations for the low correlation between our model's predictions and human appearance biases.

Experiment B To better assess our model's performance and investigate the disparity in predictive performance on the maximally distinct faces and the randomly generated faces, we train the regression model on only the maximally distinct

faces and test on only the randomly generated faces. Though a 10-fold cross validation of the maximally distinct model has an average explained variance of 97% and an average prediction correlation of 99%, prediction on the randomly generated faces is much less accurate than in Experiment A ($\bar{\rho} = 0.32$).[6] Like the human participants, our model tends to agree more about judgments of deliberately manipulated faces than about judgments of randomly generated faces. Our approach learns subjective scores of appearance bias, generated in a controlled experiment, more accurately with respect to judgments of dominance than judgments of other traits, perhaps because dominance has been shown to be less correlated with facial cues than other traits [45]. The standard deviation in dominance scores on the original 9-point scale is 1.14, much higher than the average 0.72 standard deviation for the other traits.

For reference, Safra et al. [37] report significant correlation coefficients of $\rho = 0.22$ for trustworthiness and $\rho = 0.16$ for dominance when validating on four external databases of real faces with human-annotated bias scores. Compared to the high correlation scores for the computer-generated faces, both our results and those of Safra et al. [37] convey a fairly large generalization gap, suggesting both models struggle to generalize to non-maximally distinct faces. Some of this gap may be attributed to differences in the two separate groups of study participants from which the two sets of ground-truth labels were sourced, but like Todorovet al.[43], we assume that the participants were selected from the same population and that there is no systematic difference in the biases of the two groups. That our model's performance drops significantly for randomly generated faces suggests the generalization gap is due not only to the uncanny differences between computer-generated faces and real faces but also to overfitting on the particular feature dimensions that were manipulated during maximally distinct face generation.

Experiment C Do these results hold if the problem of learning appearance bias is treated as a classification problem? We binarize the ground-truth trait judgments into "positive" and "negative" classes (e.g. "Trustworthy" and "Not Trustworthy") and train a random forest classifier, instead of a random forest regressor, on the class-labeled face embeddings. Again, the model performs well when tested on maximally distinct faces, with 95% accuracy for the trustworthy trait in 10-fold cross-validation, but poorly when tested on random faces (46% accuracy). If the model is trained only on the random faces, it achieves a 10-fold cross-validation accuracy of only 43%. For both models, false negatives and false positives occur at roughly the same rate.

Experiment D Though the model performs no better than chance on the randomly generated faces, perhaps the bias judgments learned from the maximally distinct faces will predict bias in real, human faces. We train a model on the computer-generated Maximally Distinct Faces dataset to predict competence scores for the Politicians faces. There is no significant correlation between the predicted competence scores and the competence scores collected from human participants, according to a Pearson's product-moment correlation t-test. The RMSE for this model is 98.8, significantly higher than in Experiment B but much worse than in Experiment A. Further, [3] find that competency judgments predict 2006 Gubernatorial and Senator election winners at an average rate of 68.6% ($p < 0.008$) and 72.4% ($p < 0.016$) against chance, respectively, according to a 1-sample chi-square test of proportion. Our predicted competency judgments only predict winners at an average rate of 45.7% ($p = 0.61$) in Gubernatorial races and 67.9% ($p < 0.1$) in Senate races. For comparison, random chance would predict the correct winner 50% of the time. Neither result differs significantly from chance at more than 95% confidence. There is no significant correlation between predicted competence and vote difference ($p = 0.20$), but there is a slight correlation ($\rho = 0.21$, $p < 10^{-3}$) between the difference in predicted competence scores between two candidates in a race and the vote spread. In summary, a model trained on random faces and maximally distinct faces also fails to generalize to real-world faces. Perhaps a model trained on more randomly generated faces would generalize better than a model trained solely on maximally distinct faces, but there are not enough ground-truth labels. We leave this to future work.

4.2 Feature analysis

Why does our model perform well on the maximally distinct faces, but poorly in the wild? Equally poor performance on the computer-generated random faces (Experiment B) suggests that generalization from computer-generated faces to real faces is not the only challenge in learning appearance bias.

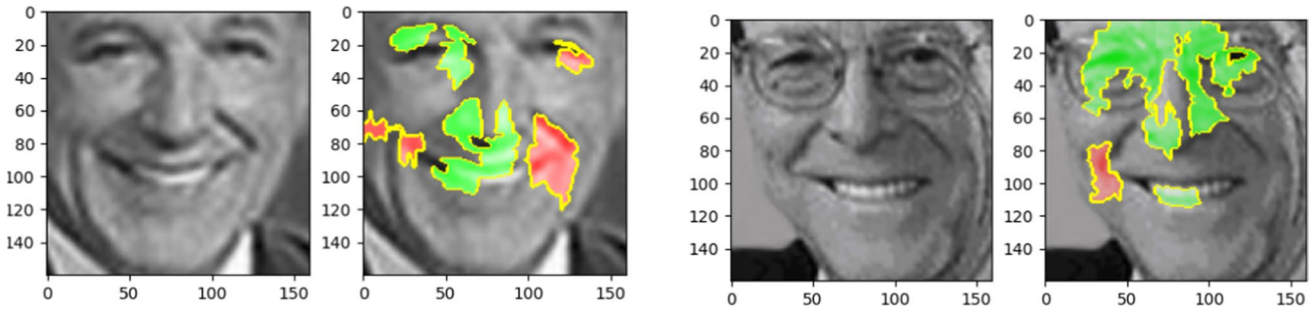
Face embeddings Though FaceNet embeddings clearly differentiate each individual face in the dataset, they are not designed to represent the facial features relevant to trait judgments. Using Uniform Manifold Approximation and Projection (UMAP), we cluster embeddings of both the Maximally Distinct Faces and the Random Faces (Fig. 4). UMAP is a popular unsupervised clustering algorithm for image data, good for efficiently capturing the global structure of high-dimensional data [26]. Recall that the maximally distinct faces are created by manipulating each of 75 random faces into 175 different faces spread along a trait dimension. In 3D space, each maximally distinct face tends to be clustered with its manipulated siblings despite variation across the trait axis. Likewise, there is no pattern of trait clustering in the distribution of random

[6] For all traits, the errors are distributed approximately normally.

(a) Features from MAXIMALLY DISTINCT FACES, shaded by face identity (which randomly generated face was manipulated to produce this face). UMAP tends to cluster maximally distinct faces that were created by manipulating the same base face.

(b) Features from 300 RANDOM FACES, shaded by perceived trustworthiness (in SD from the mean, multiplied by 100).

Fig. 4 2D UMAP projection of face features for both datasets with 15 features, minimum distance of 0.1, and 2 components. We use a high number of features to avoid spurious clustering [26]



(a) Face with greatest prediction error. From 1996 Gubernatorial race, UT.

Predicted competence:	+2.27SD
Ground-truth competence:	-1.61SD

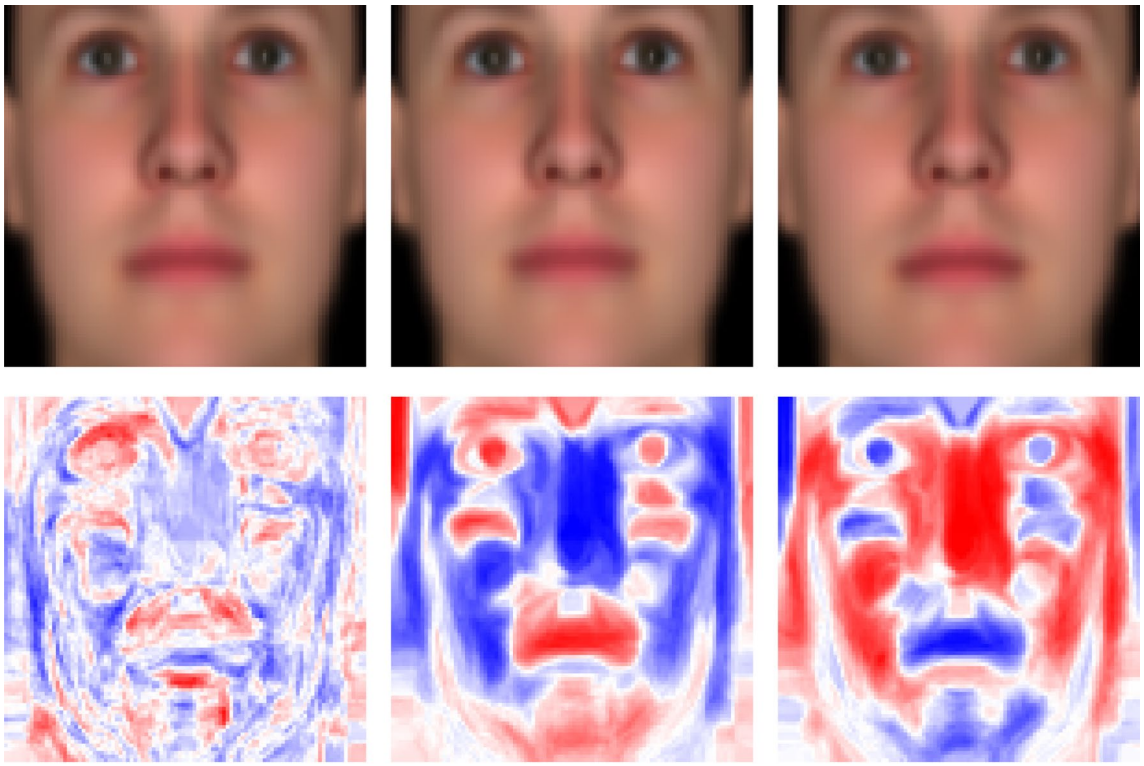
(b) Face with least prediction error. From 2002 Senate race, NM.

Predicted competence:	+0.72SD
Ground-truth competence:	+0.72SD

Fig. 5 LIME explanations for predicted competence judgments of politician faces. The ten most important superpixels are shaded. Green indicates agreement with the ultimate prediction; red indicates disagreement (color figure online)

faces. Despite computer manipulation of trait appearances in the maximally distinct faces, an unsupervised projection of FaceNet embeddings emphasizes differences between individual faces, not differences in features that contribute to appearance biases. The unsupervised industry-standard face embedding model we use in this study is designed to embed features that distinguish individual faces in a variety of poses and expressions, allowing face recognition classifiers trained on these embeddings to more easily generalize to new settings. But evidently, these unsupervised embeddings do not automatically distinguish faces according to subjective traits. As a result, the final, supervised classifier struggles to generalize to real-world datasets.

Feature importance What facial features is our model using to make trait judgments? We generated Local Interpretable Model-Agnostic Explanations (LIME) for each face in RANDOM FACES and POLITICIANS [36]. LIME is a popular black box interpretability tool that approximates an interpretable local model for the classification version of this problem (Experiment C). Taking our model and a test sample as input, LIME perturbs the superpixels of the sample face and measures the corresponding changes in our model's prediction. These changes indicate which groups of pixels or most important to our model's prediction and whether they agree with or contradict the final prediction. Images are segmented into 300 superpixels



(a) All faces. **(b)** Predicted trustworthy. **(c)** Predicted untrustworthy.

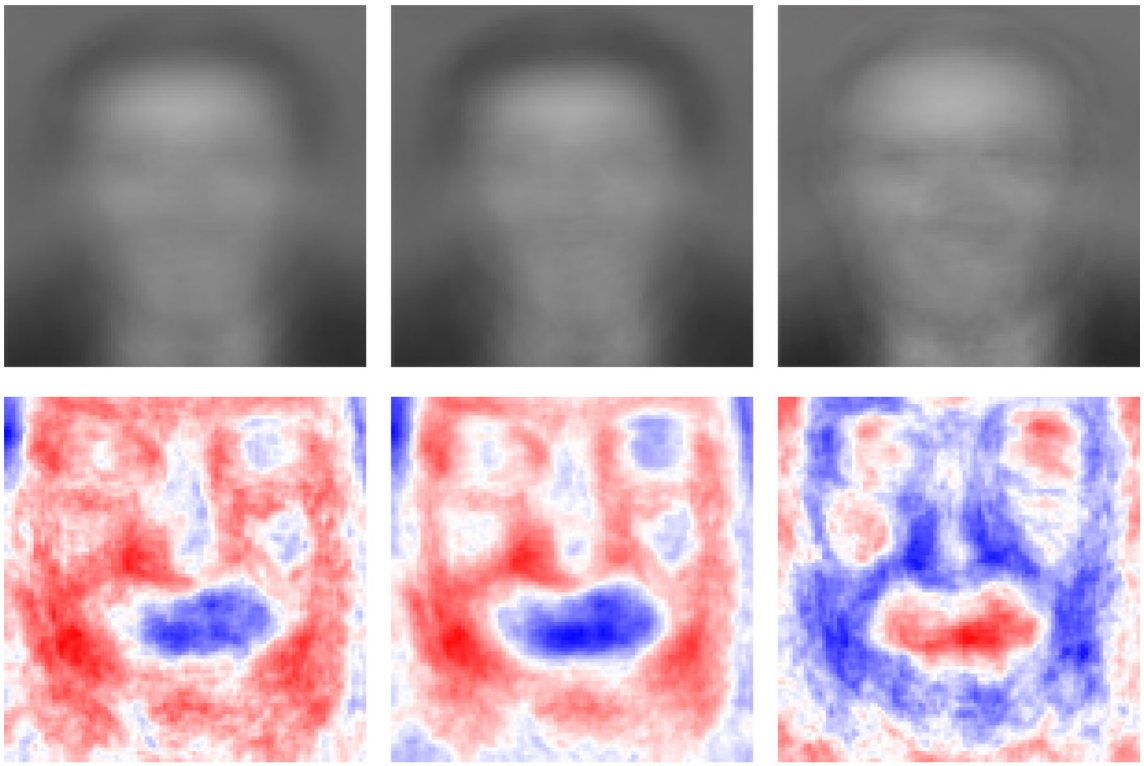
Fig. 6 Average face (top) and corresponding heatmap of average LIME explanation across all predicted trustworthy judgments of 300 Random Faces. Blue indicates agreement with the ultimate prediction; red indicates disagreement (color figure online)

with Simple Linear Iterative Clustering (SLIC), enough to capture facial features as small as the pupil [1]. We generate explanations in a neighborhood of 5000 samples; large samples tend to reduce variability in the outputted feature weights [36]. Figure 5 depicts two example explanations for the greatest and least prediction errors in the Politicians out-of-sample test set.

In the 300 Random Faces dataset, the features which contribute most (positively or negatively) to the final prediction are consistently clustered around the eyes, nose cheekbones, mouth, and upper lip (Fig. 6). Occasionally, particularly for lighter faces, features are scattered all across the face, but usually not in the background. These observations hold for photos of real people: though there is additional variance in face position, features clustered around the average position of the mouth, cheekbones, and eyes contribute most to the final competence prediction for Politicians dataset (Fig. 7). There do not appear to be any differences in allotment of feature importance between trustworthiness, competence, and other traits. This result may be surprising, but the models used to generate the training data (Maximally Distinct Faces) do not explicitly manipulate particular facial features [31]. In both datasets, our model appears to rely on the same facial features to classify faces for multiple traits; according to our results, these features are not predictive of human appearance biases.

5 Discussion and conclusions

Though our model can learn appearance bias from a small set of maximally distinct, computer-manipulated faces, it fails to make similar trait judgments out-of-sample and does not exhibit the same biases in the wild as people do. This result casts doubt on the use of computationally manipulated features to learn appearance bias: our model is trained on the same data as Safra et al. [37], who achieve similarly low correlation scores on out-of-sample faces. With clustering and interpretability analyses, we identify two explanations for this phenomenon. First, there is insufficient overlap between state-of-the-art embeddings for face recognition and the features required to identify appearance biases in real and random faces, if they exist at all. Second, though (1) the trait dimensions identified and manipulated by Todorov et al. [43] to produce maximally distinct faces match human appearance biases and (2) similar features can be used to explain our model's predictions, these features are not predictive of subjective



(a) All faces. **(b)** Predicted trustworthy. **(c)** Predicted untrustworthy.

Fig. 7 Average face (top) and corresponding heatmap of average LIME explanation across all predicted competence judgments of Politicians. Blue indicates agreement with the ultimate prediction; red indicates disagreement (color figure online)

human judgments for real or even randomly generated faces. In short, industry-standard face recognition is not sufficient to learn subjective human judgments from this computational model of trait perception.

If, as Todorov et al. [43] claim, "computational models are the best available tools for identifying the source of [trait] impressions," then more research is needed to construct externally valid representations of appearance bias. For example, the ground-truth measures of subjective trait judgments currently available are sourced from largely white, young Princeton students, whose appearance biases are not globally representative. Further, predictions from transfer learning models trained on maximally distinct, computer-generated features provide neither an objective measure of personality traits (they represent subjective biases) nor a good measure of subjective bias itself, as we show. However, our results do not rule out the possibility that appearance bias could be embedded from a larger training set of real faces with labels from a more representative set of participants: future work should investigate whether human appearance bias manifests in large-scale datasets in the wild.

Acknowledgements We thank two anonymous reviewers for their insightful comments and helpful feedback; and in particular, we thank one anonymous reviewer for inspiring Figs. 6 and 7.

Funding Funding was provided by the George Washington University's Department of Computer Science. This material is based on research partially supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB20D212. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the NIST.

Availability of data and material All training and test data are available on request from the original authors at http://tlab.princeton.edu.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. To preserve anonymity, other declarations are made on the unblinded title page.

Code availability Auxiliary data and code used to produce the figures, tables, and machine learning model in this work are available at https://github.com/ryansteed/learning-appearance-bias.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2281 (2012). https://doi.org/10.1109/TPAMI.2012.120
2. Amos, B., Ludwiczuk, B., Satyanarayanan, M. OpenFace: a general-purpose face recognition library with mobile applications. CMU-CS-16-118, CMU School of Computer Science, Tech. Rep. (2016). http://cmusatyalab.github.io/openface/
3. Ballew, C.C., Todorov, A.: Predicting political elections from rapid and unreflective face judgments. Proc. Natl. Acad. Sci. USA **104**(46), 17948–17953 (2007). https://doi.org/10.1073/pnas.0705435104. www.pnas.org/cgi/doi/10.1073/pnas.0705435104
4. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., Van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.J., Heng, P.A., Haß, C., Bruni, E., Wong, Q., Halici, U., Oner, M.A., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Phoulady, H.A., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA J. Am. Med. Assoc. **318**(22), 2199–2210 (2017). https://doi.org/10.1001/jama.2017.14585
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Buolamwini, J.: Opinion: when the robot doesn't see dark skin. In: New york times (2018). https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html
7. Buolamwini, J., Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A.,Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, vol. 81, pp. 77–91 New York, NY, USA (2018). http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
8. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Tech. Rep. 6334 Sci. (2017). https://doi.org/10.1126/science.aal4230. https://science.sciencemag.org/content/356/6334/183/tab-pdf
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2012). https://doi.org/10.1109/CVPR.2012.6248074
10. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. J. Personal. Social Psychol. **74**(6), 1464–80 (1998). http://www.ncbi.nlm.nih.gov/pubmed/9654756
11. Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., Banaji, M.R.: Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. J. Personal. Social Psychol. **97**(1), 17 (2009)
12. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: European conference on computer vision, Springer, 87–102 (2016). http://arxiv.org/abs/1607.08221
13. Hamermesh, D.S., Biddle, J.E.: Beauty and the labor market. Am. Econ. Rev. **84**(5), 1174–1194 (1994). http://www.jstor.org/stable/2117767
14. Hao, K.: The two-year fight to stop Amazon from selling face recognition to the police. MIT. Tech. Rev. (2020). https://www.technologyreview.com/2020/06/12/1003482/amazon-stopped-selling-police-face-recognition-fight/
15. Harwell, D.: A face-scanning algorithm increasingly decides whether you deserve the job. In: Washington Post (2019). https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/
16. Hassin, R., Trope, Y.: Facing faces: studies on the cognitive aspects of physiognomy. J. Pers. Soc. Psychol. **78**(5), 837–852 (2000). https://doi.org/10.1037/0022-3514.78.5.837
17. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: overcoming bias in captioning models. CoRR (2018). https://doi.org/10.1007/978-3-030-01219-9fng47
18. Jacques Junior, J.C., Andujar, C., BaroBar, X., Jair Escalante, H., Guyon, I., van Gerven, M.A., van Lier, R., Escalera, S., Jair Escalanteis, H.: First impressions: a survey on computer vision-based apparent personality trait analysis. Tech. Rep. arXiv. (2018). arXiv:1804.08046https://www.theguardian.com/technology/2017/apr/13/
19. Kay, M., Matuszek, C., Munson, S.A.: 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—CHI'15, pp. 3819–3828. ACM Press, New York (2015). https://doi.org/10.1145/2702123.2702520. http://dl.acm.org/citation.cfm?doid=2702123.2702520
20. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
21. Keating, C.F., Randall, D., Kendrick, T.: Presidential physiognomies: altered images, altered perceptions. Polit. Psychol. **20**(3), 593–610 (1999). https://doi.org/10.1111/0162-895X.00158. /record/1999-11324-006
22. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. Tech. Rep. 23180 Nat. Bureau Econ. Res. (2017). https://doi.org/10.3386/w23180. http://www.nber.org/papers/w23180
23. Ko, T.: A survey on behavior analysis in video surveillance for homeland security applications. Proc. Appl. Imagery Pattern Recognit. Workshop (2008). https://doi.org/10.1109/AIPR.2008.4906450
24. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. Proc. IEEE Comput. Soc. Conf.

Comput. Vis. Pattern Recognit. (2010). https://doi.org/10.1109/CVPR.2010.5539872
25. Manjunatha, V., Saini, N., Davis, L.: Explicit bias discovery in visual question answering models. 9554–9563 (2019). https://doi.org/10.1109/CVPR.2019.00979
26. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. (2018). http://arxiv.org/abs/1802.03426
27. Mueller, U., Mazur, A.: Facial dominance of west point cadets as a predictor of later military rank*. Soc. Forces **74**(3), 823–850 (1996). https://doi.org/10.1093/sf/74.3.823
28. Murgia, M.: Who's using your face? The ugly truth about facial recognition. Financial Times. https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e
29. Nagpal, S., Singh, M., Singh, R., Vatsa, M.: Deep learning for face recognition: pride or prejudiced? (2019). http://arxiv.org/abs/1904.01219
30. Nex, F., Remondino, F.: UAV for 3D mapping applications: a review. (2014). https://doi.org/10.1007/s12518-013-0120-x
31. Oosterhof, N.N., Todorov, A.: The functional basis of face evaluation. Tech. Rep. (2008). https://www.pnas.org/content/105/32/11087
32. Pearson, J.: Microsoft deleted a massive facial recognition database, but it's not dead. Vice. https://www.vice.com/en_us/article/a3x4mp/microsoft-deleted-a-facial-recognition-database-but-its-not-dead
33. Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating bias in algorithmic hiring: evaluating claims and practices. FAT* 2020 Proc. 2020 Conf. Fairness Account. Transp. Assoc. Comput. Mach. Inc. (2020). https://doi.org/10.1145/3351095.3372828
34. Raji, I.D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., Denton, E.: Saving face: investigating the ethical concerns of facial recognition auditing. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 145–151 (2020)
35. Rezlescu, C., Duchaine, B., Olivola, C.Y., Chater, N.: Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. PLoS ONE **7**(3) (2012). https://doi.org/10.1371/journal.pone.0034293. https://pubmed.ncbi.nlm.nih.gov/22470553/
36. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, vol. 13–17, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778. https://dl.acm.org/doi/10.1145/2939672.2939778
37. Safra, L., Chevallier, C., Grèzes, J., Baumard, N.: Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. Nat. Commun. **11**(1), 4728 (2020). https://doi.org/10.1038/s41467-020-18566-7. http://www.nature.com/articles/s41467-020-18566-7
38. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: IEEE CVPR, pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682. http://arxiv.org/abs/1503.03832
39. Snow, J.: Amazon's face recognition falsely matched 28 members of congress with mugshots. In: American Civil Liberties Union (2018). https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28
40. Todorov, A.: Face Value: The Irresistible Influence of First Impressions. Princeton University Press, Princeton (2017)
41. Todorov, A., Mandisodza, A.N., Goren, A., Hall, C.C.: Inferences of competence from faces predict election outcomes. Science **308**(5728), 1623–1626 (2005). https://doi.org/10.1126/science.1110589
42. Todorov, A., Dotsch, R., Wigboldus, D.H.J., Said, C.P.: Data-driven methods for modeling social perception. Social Person. Psychol. Compass **5**(10), 775–791 (2011). https://doi.org/10.1111/j.1751-9004.2011.00389.x. http://doi.wiley.com/10.1111/j.1751-9004.2011.00389.x
43. Todorov, A., Dotsch, R., Porter, J.M., Oosterhof, N.N., Falvello, V.B.: Validation of data-driven computational models of social perception of faces people instantly form impressions from facial. Emotion **13**(4), 724–738 (2013). https://doi.org/10.1037/a0032335.supp. http://tlab.princeton.edu/publication_files/TodorovDotschetalEmotion2013.pdf
44. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. CVPR, IEEE, 1521–1528 (2011)
45. Willis, J., Todorov, A.: First impressions. Psychol. Sci. **17**(7), 592–598 (2006). https://doi.org/10.1111/j.1467-9280.2006.01750.x. http://journals.sagepub.com/doi/10.1111/j.1467-9280.2006.01750.x
46. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv preprint 190211097 (2019). http://arxiv.org/abs/1902.11097
47. van't Wout, M., Sanfey, A.G.: Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. Cognition **108**(3), 796–803 (2008). https://doi.org/10.1016/j.cognition.2008.07.002. https://pubmed.ncbi.nlm.nih.gov/18721917/
48. Yang, K., Mall, S., Glaser, N.: Prediction of personality first impressions with deep bimodal LSTM. Tech. Rep. arXiv. (2017). http://cs231n.stanford.edu/reports/2017/pdfs/713.pdf
49. Zebrowitz, L.A., Andreoletti, C., Collins, M.A., Lee, S.Y., Blumenthal, J.: Bright, bad, babyfaced boys: appearance stereotypes do not always yield self-fulfilling prophecy effects. J. Personal. Social Psychol. **75**(5), 1300–1320 (1998). https://doi.org/10.1037/0022-3514.75.5.1300
50. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: reducing gender bias amplification using corpus-level constraints. EMNLP 2017 Conf. Empirical Methods Nat. Lang. Process. Proc. Assoc. Comput. Linguistics (ACL) (2017). https://doi.org/10.18653/v1/d17-1323