PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Deep models of superficial face judgments

Joshua C. Peterson[a,1], Stefan Uddenberg[b], Thomas L. Griffiths[a,c], Alexander Todorov[b], and Jordan W. Suchow[d]

Edited by Winrich Freiwald, The Rockefeller University, New York, NY; received August 17, 2021; accepted March 7, 2022, by Editorial Board Member Charles D. Gilbert

The diversity of human faces and the contexts in which they appear gives rise to an expansive stimulus space over which people infer psychological traits (e.g., trustworthiness or alertness) and other attributes (e.g., age or adiposity). Machine learning methods, in particular deep neural networks, provide expressive feature representations of face stimuli, but the correspondence between these representations and various human attribute inferences is difficult to determine because the former are high-dimensional vectors produced via black-box optimization algorithms. Here we combine deep generative image models with over 1 million judgments to model inferences of more than 30 attributes over a comprehensive latent face space. The predictive accuracy of our model approaches human interrater reliability, which simulations suggest would not have been possible with fewer faces, fewer judgments, or lower-dimensional feature representations. Our model can be used to predict and manipulate inferences with respect to arbitrary face photographs or to generate synthetic photorealistic face stimuli that evoke impressions tuned along the modeled attributes.

face perception | social traits | computational models

Significance

We quickly and irresistibly form impressions of what other people are like based solely on how their faces look. These impressions have real-life consequences ranging from hiring decisions to sentencing decisions. We model and visualize the perceptual bases of facial impressions in the most comprehensive fashion to date, producing photorealistic models of 34 perceived social and physical attributes (e.g., trustworthiness and age). These models leverage and demonstrate the utility of deep learning in face evaluation, allowing for 1) generation of an infinite number of faces that vary along these perceived attribute dimensions, 2) manipulation of any face photograph along these dimensions, and 3) prediction of the impressions any face image may evoke in the general (mostly White, North American) population.

Faces are among the most important stimuli that people encounter—they are recognized by infants long before other objects in their environment (1), recruit specialized circuits in the brain (2), and are fundamental to social interaction (3). Central to our experience with faces are the attributes that we assign to them, often implicitly. These include attributes that are read off, describing largely objective attributes of faces (e.g., age and adiposity), and those that are read into, such as how trustworthy a person is (4). Although the inferences of the latter attributes are more subjective and generally inaccurate, they are similarly psychologically consistent across people (4–6) around the globe (7–9) and have important consequences (10) ranging from electoral success (11, 12) to sentencing decisions (13, 14). Because any face can be judged with respect to such attributes, these psychological dimensions are universal in that they are implicitly defined over the space of nearly all possible faces, contexts, and observational conditions. These factors combine to form a diverse landscape of stimuli that makes it challenging to capture the corresponding psychological content in its entirety. Such content forms the basis of scientific models of face perception and defines the scope of downstream applications such as training people to overcome stereotypes (15).

The importance of face attribute inferences has led to the proliferation of techniques for scientific modeling of faces, which can be organized broadly into two approaches. The first extrapolates from face photographs, often related via landmark annotations (16, 17). The second generates artificial faces using parametric three-dimensional face meshes (18). Photographs offer greater realism but are limited to available datasets of face stimuli that serve as the basis for interpolation and by the interpolation algorithms themselves, which often require high-quality landmark annotations unattainable without costly manual work (19). Artificially generated faces are not subject to these limitations but lack diversity and realism. Neither approach provides workable models that express the full richness and diversity of human faces.

Machine learning methods, in particular deep neural networks such as generative adversarial networks (GANs), can learn to model faces from massive collections of photographs scraped from image-sharing websites (20–23). These methods present a third option for developing scientific models of faces, providing expressive feature representations for arbitrary realistic face images. However, relating these representations to human perception is difficult because they are high-dimensional vectors produced via black-box optimization algorithms (24).

We show that the keys to unlocking the scientific potential of these models and their downstream applications are large-scale datasets of human behavior unattainable using traditional laboratory experiments. In particular, such large datasets provide sufficient evidence to determine a robust mapping between expressive high-dimensional representations from machine learning models and human mental representations of faces.

Author contributions: J.C.P., S.U., T.L.G., A.T., and J.W.S. designed research; J.C.P. and S.U. performed research; J.C.P., S.U., T.L.G., A.T., and J.W.S. contributed new reagents/analytic tools; J.C.P., S.U., and J.W.S. analyzed data; and J.C.P., S.U., T.L.G., A.T., and J.W.S. wrote the paper.

Competing interest statement: All authors are listed as inventors on a related patent (US Patent no. 11,250,245, "Data-driven, photorealistic social face-trait encoding, prediction, and manipulation using deep neural networks").

This article is a PNAS Direct Submission. W.F. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: joshuacp@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115228119/-/DCSupplemental.

Published April 21, 2022.

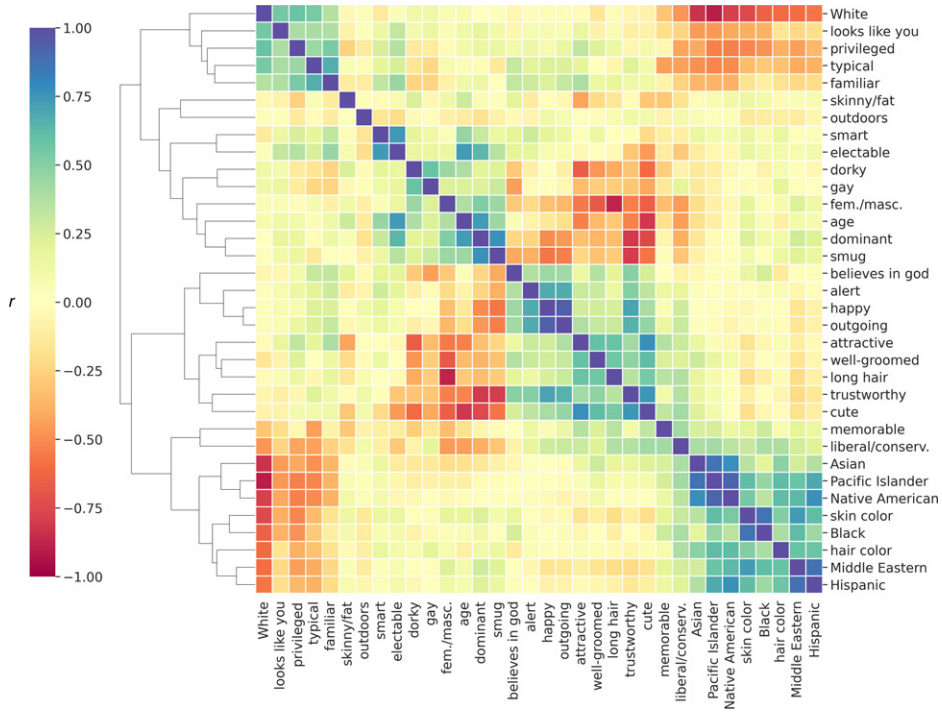

Fig. 1. Correlation matrix for 34 average attribute ratings for each of 1,000 faces. Rows and columns are arranged according to a hierarchical clustering of the correlation values.

We quantify an upper bound on the robustness of the mapping in terms of the reliability of the underlying attribute inferences and determine how that robustness scales as a function of the number of faces rated, the number of ratings per face, and the dimensionality of the deep feature space. We then use this mapping to predict and manipulate inferences over arbitrary face images, enabling us, for example, to adjust a photograph so as to increase or decrease the perceived trustworthiness of its subject to match a target rating.

Such a mapping can be computed for any psychologically meaningful attribute inference. We focus on three classes of such inferences. First, there are inferences defined by subjective impressions of relatively objective properties (e.g., age and adiposity). These more objective properties, which also include hair styling, presence of accessories (e.g., glasses), gaze, and facial expression, are commonly studied in computer vision, where they are referred to as "attributes" (25) or "soft biometrics" (26). Next, there are inferences of subjective and socially constructed attributes, such as *trustworthy* and *masculine/feminine*, the conventional targets of social scientific study (4). Finally, there are inferences of fully subjective attributes such as *familiar*, where the observer is the only arbiter of truth (27). For ease of presentation, we refer to inferences of all three classes as "attribute inferences" and the underlying attributes as "attributes," drawing distinctions between the classes in the text as necessary. Please note that these attribute inferences, especially those of the more subjective or socially constructed attributes, have no necessary correspondence to the actual identities, attitudes, or competencies of people whom the images resemble or depict (e.g., a trustworthy person may be wrongly assumed to be untrustworthy on the basis of appearance). Rather, these inferences, and in turn our measurements, reflect systematic biases and stereotypes about attributes shared by the population of raters. Nevertheless, these inferences are driven by (combinations of) physical cues present in the faces themselves—for example, faces judged to look more trustworthy may have more neotenous features (e.g., large eyes) or upturned lips, as in a smile (4).

We used online crowdsourcing to obtain attribute inference ratings for just over 1,000 synthetic (although highly naturalistic) face stimuli for 34 attributes, with ratings by at least 30 unique participants per attribute–stimulus pair, for a total of 1,020,000 human judgments (*Materials and Methods*). We call this collection of face stimuli and the corresponding behavioral data the One Million Impressions dataset. A detailed summary of these ratings and interattribute relationships can be found in *SI Appendix*.

Results

The Structure of Attribute Inferences. To explore the structure of attribute inferences, we first computed the correlation between the mean face ratings for each pair of attributes (Fig. 1). Many attributes were highly correlated, including *happy–outgoing* ($r = 0.93$) and *dominant–trustworthy* ($r = -0.81$), while others were largely unrelated, including *smart–attractive* ($r = 0.01$), *smart–trustworthy* ($r = 0.02$), *liberal/conservative–believes in god* ($r = 0.08$), and *electable–attractive* ($r = 0.05$).

Although some of these correlations are consistent with previous findings (8), others are not. First, although past work has found that judgments of trustworthiness and dominance are often negatively correlated, the correlation is generally small (on the order of –0.2), whereas the correlation observed here (–0.81) was much stronger (4, 28). Second, judgments of smartness or competence have been found to be highly positively correlated with judgments of attractiveness and trustworthiness (with values as high as ≈ 0.8), whereas we found only marginal correlations between those attribute inferences (8). One explanation for these discrepancies is that the face stimuli used here are more diverse than in the comparison studies, especially with respect to age; the present study includes (simulated) childrens' faces. This explanation is plausible given that the correlational structure of judgments of children's faces is different from the structure of judgments of adult faces (29). To probe this hypothesis, we recomputed interattribute correlations on subsets of the data with restricted age

ranges (*SI Appendix*, Fig. S9). We found that the inclusion of the children's faces partially explains some discrepancies (e.g., *smart–attractive*) and does not explain others (*trustworthy–dominant*). Third, *memorable* faces were more attractive, as observed in the positive correlation between the respective ratings (Fig. 1). This finding is inconsistent with work showing that actual memorability of faces is negatively correlated with attractiveness to the extent that predictions of memorability are veridical (30). Last, familiar faces were seen as more attractive and average-looking, consistent with the finding that average faces tend to be perceived as more attractive (ref. 31, but see ref. 32).

The attribute *outdoors* (whether the photo appeared to be taken outdoors or indoors) was included to assess potential confounds in using naturalistic face photos. It was found to be the least correlated with other attributes, having the lowest per-attribute maximum absolute correlation (*outdoors–electable*, $r = 0.20$). In comparison, the attribute with the next-lowest maximum was *skinny/fat* (*skinny/fat–attractive*, $r = 0.43$), which despite having twice the magnitude was one of the easier attributes to predict (Fig. 2). Furthermore, that *outdoors* had the lowest mean absolute correlation with all other attributes ($r = 0.08$) indicates minimal contribution of contextual effects due to naturalistic backgrounds and lighting.

Predicting Attribute Inferences. To model an attribute, we start with the high-dimensional representation vectors $\mathbf{z}_i = \{z_1, \ldots z_d\}$ assigned to each synthetic face i in our stimulus set by a pretrained state-of-the-art GAN (21, 22, 33). The GAN has learned a mapping from each such vector to an image through extensive training on a large database of real, nonsynthetic face photographs (*Methods and Materials*). We then model each psychological attribute, measured via average ratings y_i, as a linear combination of features: $y_i = w_0 + w_1 z_1 + \ldots + w_d z_d$. The vector of weights $\mathbf{w}_k = \{w_1, \ldots w_d\}$ represents the attribute as a linear dimension cross-cutting the representational space and is fit using cross-validated, $L2$-regularized linear regression. A diagram summarizing the modeling pipeline for predicting attribute inferences is provided in *SI Appendix*, Fig. S1.

Average cross-validated (i.e., out-of-sample) model performance for each attribute is reported in Fig. 2. Prediction for most attributes was reasonably successful, with most R^2 values ranging from above 0.5 to almost 0.8, with attributes *typical*, *familiar*, and *gay* being the exceptions.

Because participants partly disagree in their appraisals (34), perfect prediction is impossible. To better understand the prediction ceiling imposed by limited interrater reliability, we computed the split-half reliability for each attribute, averaging the squared correlations between the averages of 100 random splits of the ratings for each image. These prediction ceilings vary across attributes and are plotted in Fig. 2 alongside the corresponding model shortfalls they imply. (See *Factors Influencing Prediction Performance* for a detailed characterization.)

Interestingly, the models of *familiar* and *looks like you* showed the smallest gaps between performance and reliability, indicating that their unpredictability is not due to poor model quality or lack of useful input features. Rather, it seems likely that *familiar* more so than other attributes is based on both a shared concept or experience and a much larger personal concept or experience; only the former can be predicted for participants in aggregate. This is corroborated by a similar effect for the attribute *looks like you*, which can be predicted only at the aggregate level to the extent that our participant pool has a shared representation of their respective facial features, which may be a byproduct of the participant pool having less diversity in appearance than does the stimulus set.

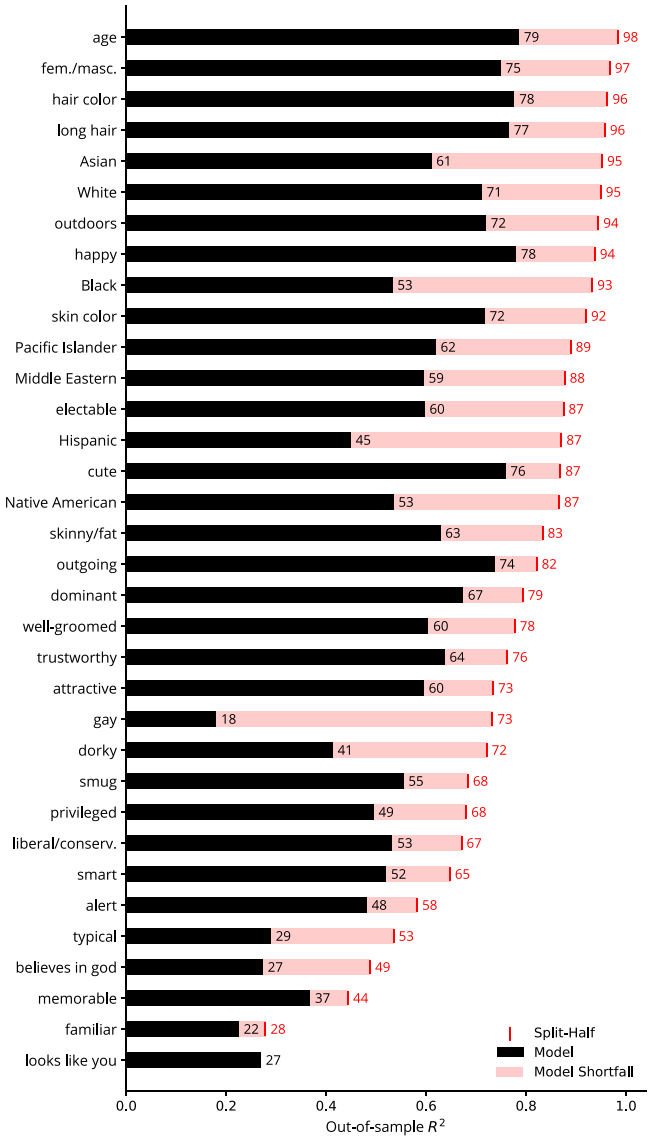

Fig. 2. Average cross-validated model performance (black bars) compared to intersubject reliability (red markers).

Attributes corresponding to some racial or ethnic social categories, such as "Black," exhibited a larger gap between reliability and model performance than did other attributes. One possible reason for this gap is a sampling bias in the stimulus generator. Indeed, rating-distribution violin plots provide some indication that, for example, Black faces were undersampled (*SI Appendix*, Fig. S3). A second possible reason for the gap is that the degree of undersampling of participants who report membership in a racially or ethnically minoritized group might correlate with the content of the aggregate attribute inferences. However, we observe no such correlation: the second most common participant self-identifier was Black, which had one of the largest gaps between reliability and model performance, whereas attributes corresponding to even less commonly self-reported racial and ethnic social categories had a smaller gap. A third possible reason is that the predominantly White participant pool might have similar stereotypes of all racially or ethnically minoritized groups, leading to comparable predictability across the relevant attributes. However, this explanation is inconsistent with the lack of strong correlation observed across those attributes (Fig. 1). Taken together, this suggests that the stimulus selection contributes more

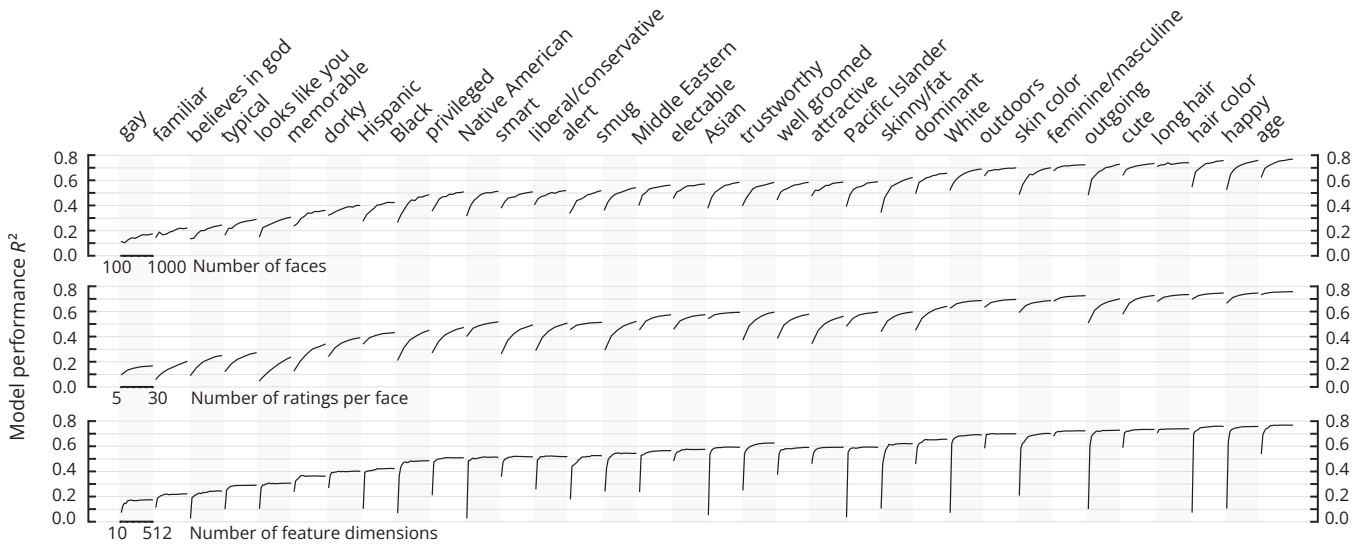

Fig. 3. Model performance (R^2) for each attribute as a function of the number of face examples (*Top*), the number of participant ratings for each face example (*Middle*), and the number of image feature dimensions (*Bottom*). Attributes are ordered by the maximum model performance observed in *Top*.

to the observed gap than does the composition of the participant pool. Even so, no firm conclusion can be drawn because we cannot rule out limitations in the representational capacity of the neural network features.

Factors Influencing Prediction Performance. To characterize the factors influencing prediction performance, we first investigated the effect of the number of faces rated on predictive performance (Fig. 3, *Top*). Performance curves were generated by fitting models for each of 30 random samples of images with sizes ranging from 100 to 1,000. Most attributes benefit from increases in the number of faces rated, with significant variation across them with respect to how performance scales with the number of unique faces that were rated. Interestingly, fewer images were needed to saturate model performance for the attribute *feminine/masculine* than for most other attributes. For all other attributes, adding additional images improved performance through the full range.

Next, we investigated the relationship between the number of ratings by unique participants obtained for each face stimulus and predictive performance (Fig. 3, *Middle*). Performance curves were generated by fitting models on down-sampled datasets with sizes ranging from 5 to 30 unique ratings per image, with 30 datasets sampled per size. Aside from attributes *feminine/masculine* and *age*, which elicit less disagreement, performance increases considerably as the number of ratings increases for all attributes. Gains due to the number of ratings diminish with increases in the number of unique ratings but at a slower rate than gains due to the number of faces (Fig. 3, *Top*). It remains to be seen the extent to which scaling the number of rated faces and the number of ratings per face beyond the range explored here accounts for the gap between model performance and the ceiling imposed by interrater reliability.

Finally, we investigated the relationship between the number of image features (512 total) and predictive performance (Fig. 3, *Bottom*). Performance curves were generated by fitting models using reduced feature sets obtained via principal components analysis, varying the dimensionality between 10 and 512. In all cases, performance saturates quickly but is improved marginally with a greater number of dimensions in some cases. The various profiles of saturation indicate that as few as 10 dimensions of this latent feature space may be enough to account for the bulk of variance in attribute inferences, with a subset of attributes benefiting considerably from higher-dimensional feature representations.

It is possible that the quality of the learned representation from the particular deep neural network we employed was a limiting factor in predictive performance. Factors beyond predictive performance (specifically, the ability to generate images in addition to representing them) guided network selection for the present study. Other architectures with other forms of supervision may offer improvement in predictive performance. For example, there is evidence that identity-supervised models provide representations that are highly predictive of diverse attribute information (26, 35, 36).

Manipulating Attribute Inferences. Because the learned attribute vectors correspond to linear dimensions, we can manipulate an arbitrary face represented by features $\mathbf{z}_i$ with respect to attribute k using vector arithmetic: $\mathbf{z}_i + \beta\mathbf{w}_k$, where β is a scalar that controls the positive or negative modulation of the attributes. We apply a symmetric range of β around 0 to each attribute vector to manipulate a series of base face representations in both the negative and positive directions and decode the results for visualization using the same decoder/generator component of the neural network that was used to derive representations (see *SI Appendix* for more details).

The results of these transformations for six sample attribute inferences are shown in Fig. 4. The manipulations are strikingly smooth and effective along each attribute dimension. For example, modulating *trustworthiness* increases features associated with perceptions of trustworthiness, such as eye gaze, degree of smiling, face shape, and facial femininity (4, 37). Manipulations of attribute inferences may affect more than one dimension of appearance. For example, increasing *smartness* may add glasses or change the facial expression. Increasing *outgoingness* may increase smiling, as expected, but also give glasses a more rounded and cartoonish appearance. Other dimensions allow for greater levels of extrapolation. For example, faces can be made considerably *skinnier* or *fatter* than any examples in the dataset, yet still maintain a realistic appearance. Faces with strongly manipulated *happiness* also resemble convincing caricatures.

It is possible to manipulate one dimension s (e.g., *smartness*) while controlling for another t (e.g., *trustworthiness*) by creating an orthogonal vector, subtracting the projected component of the dimension to be controlled for:

$$s - t\frac{s \cdot t}{||t||^2}. \quad [1]$$

Fig. 4. (*A*) The faces judged on average to have the highest and lowest ratings along six sample perceived attribute dimensions. (*B*) Model-based manipulations of two sample base faces along the sample dimensions, demonstrating smooth and effective manipulations along each attribute.

An example of such transformations controlling for trustworthiness on a given exemplar face can be seen in *SI Appendix*, Fig. S11.

Note that the attribute-inference manipulations can affect both internal facial features and external features. When only internal face features are altered, it is not because the GAN manipulates only internal features but because the external features are orthogonal or irrelevant to that attribute inference in the region of the manipulated face.

Validating Models of Attribute Inferences. Do the attribute models generated above reliably change participants' impressions of faces transformed with them? To answer this question, we ran a series of 20 preregistered experiments with over 1,000 participants to verify that our models can indeed manipulate attribute impressions in observers. Each of the experiments paired one of two face image types (artificial vs. real) with one of 10 different attribute dimensions, chosen to represent a wide range of different model performances and levels of objectivity/subjectivity (*age*, *feminine/masculine*, *skinny/fat*, *trustworthy*, *attractive*, *dominant*, *smart*, *outgoing*, *memorable*, and *familiar*). Like in the attribute-modeling experiments, for the artificial face experiments we generated 50 unique synthetic faces at random using StyleGAN2 (21, 22), a state-of-the-art GAN architecture (SG2). However, for the real-face experiments, we encoded into our model 50 unique face photograph, chosen from a commonly used database of real faces from the psychological literature (38). On each trial, participants were shown a single face and asked to rate it. Critically, each face image was transformed by one (experimentally assigned) perceived attribute model to evoke one of three levels of that impression; faces could be set to the mean observed value of the perceived attribute or at $\pm$ 0.5 SD from that mean. Every face was shown at every level of the assigned attribute (50 identities $\times$ 3 transformation levels $=$ 150 unique faces), and once again, 20% of trials were repeated to measure test–retest reliability (in order to exclude subjects with negative such reliability) for a total of 180 trials. If the attribute model transformations indeed change participants' impressions of the faces, then we should observe that participants' ratings of the faces increase with increasing levels of the manipulation. We found just that: repeated measures ANOVAs revealed that all of the manipulations yielded highly significant results [all $F(2, 98)$s $>$ 14.13, all values of $P < 0.000005$, all values of $\eta^2 > 0.223$]. Critically, all of the experiments' data showed a strongly significant positive linear trend [all values of $t(98) > 2.69$, all values of $P \leq 0.008$, with the exception of real faces manipulated along

the *familiar* attribute dimension, $t(98) = -1.60$, $P = 0.112$; see *SI Appendix*, Fig. S12 for a swarmplot of all the data].*

General Discussion

We set out to develop a comprehensive model of attribute perception that can predict human attribute inferences from face images and manipulate them along psychologically meaningful dimensions. With no explicit featurization or interpolation algorithm, the model accomplishes this in a fully data-driven manner with relatively high accuracy and generalization. Large datasets (with respect to both the number of face stimuli and the number of ratings per face) are necessary to achieve this. Qualitative results and validation experiments demonstrate that psychological attribute manipulations of realistic face photos can be accomplished using simple vector arithmetic. Moreover, our pipeline provides a general formula for modeling inferences of any attributes that can be measured via image annotations. Because the models of attributes are expressed in the same multidimensional space, their similarity is immediately given, enabling testing of specific hypotheses about the relation between psychological attributes, predicting novel attributes based on their relationships with models of existing attributes, and controlling for shared variance between attributes.

The model broadly characterizes inferences about diverse faces in their everyday contexts and viewing conditions. For example, in any particular image, one can generally discern whether the photograph is candid or contrived, environment conditions (e.g., outside in direct sunlight near vegetation versus inside of a building with warm lighting), the subject's pose and gaze, grooming habits, and even hints of their culture or tastes based on partially visible clothing, necklines, head wear, jewelry, glasses, etc. Other factors of variation include viewing angle, head pose, photo quality, focal length, and depth of field, among others. While capturing behavior in a way that generalizes across these variations (and includes their effects) is the primary goal, it has the considerable disadvantage of making interpretation more challenging. Consider, for example, when one face is inferred to be more trustworthy than another. Is it because of furrowed brows and a wide jaw or because of a highly atypical hat and dark lighting? Although our stimuli are synthetic, they are not highly controlled, more comparable to randomly sampled and weakly curated photographs. Thus, understanding the bases of attribute inferences will require significant additional lower-level attribute annotations (e.g., hair color). Fig. 1, for example, implies that *skinny/fat*, *long hair*, and *hair color* (hair darkness) are not particularly explanatory of attribute inferences, with some exceptions (e.g., *skinny/fat–attractive*). Qualitative inspection revealed that transformations did not appear to frequently or significantly alter nonface features (with the notable exception of spawning glasses when increasing *smart*-ness).

Another notable consideration when interpreting the current work is that the diversity of the faces used in the experiment will almost certainly influence and may even obfuscate the meaning of some attributes. For example, the semantics of the attribute *attractive* may differ when rating images of children versus adults. It is not enough to simply analyze subsets of faces in the data, because the context of the experiment may induce order effects or serial dependence (39) that influences participant ratings of all faces. It is also difficult to manipulate the context of the experiment because so many different contexts are possible. Future work could consider two possible solutions. The first is to exploit the fact that our experiment sampled faces and their ordering, and thus contexts, at random and relatively densely. Because some of these contexts will cluster into, e.g., many-children/few-children groups, this provides one possible avenue for probing relevant effects. The second is to model attributes as multimodal, wherein attributes are not single linear factors (linear combinations of features) but many potentially correlated but not wholly colinear factors that cluster in different regions of the underlying representation space of the attribute model. This may also explain part of the current gap between our predictive models and the corresponding estimated upper bounds based on intersubject reliability.

Last, it is unclear whether synthetic stimuli generated by architectures like SG2, despite being generally convincingly realistic, are in fact different from real faces in ways that could bias conclusions that make use of them. For this reason, researchers making use of these stimuli to draw conclusions about human perception should take care to validate findings derived from them using photos of real faces as appropriate.

Ethical Implications. Importantly, while the primary goal of this work is to support scientific modeling, the framework developed here adds significantly to the ethical concerns that already enshroud image manipulation software. In contrast to traditional photo editing, which may be limited in effectiveness by the intuitions of a particular artist, the current method may be more accurate and at the least is faster and more efficient through its automation. Further, in contrast to other methods making use of deep neural networks such as DeepFakes (40), which can affect the social perception of an individual by placing them in an unwanted or compromising context (e.g., superimposed on a body in an arbitrary target image), our model can induce (perceived) changes within the individual's face itself and may be difficult to detect when applied subtly enough. We argue that such methods (as well as their implementations and supporting data) should be made transparent from the start, such that the community can develop robust detection and defense protocols to accompany the technology, as they have done, for example, in developing highly accurate image forensics techniques to detect synthetic faces generated by SG2 (41, 42). More generally, to the extent that improper use of the image manipulation techniques described here is not covered by existing defamation law (43, 44), it is appropriate to consider ways to limit use of these technologies through regulatory frameworks proposed in the broader context of face-recognition technologies (45, 46).

There is also potential for our data and models to perpetuate the biases they measure, which are first impressions of the population under study and have no necessary correspondence to the actual identities, attitudes, or competencies of people whom the images resemble or depict. While the bias in our sample of raters comes from the same population as most crowdsourced studies, it may be particularly important to understand in the context of social attribute perception, given that it consists of primarily White participants from the United States. Further, we found that the generative model that synthesized our stimuli, while highly diverse, nonetheless undersamples faces of Black people and other minoritized groups. Applications of the model will thereby produce face images that are more closely aligned with this bias than are the original inputs. We quantify part of this bias beyond participant demographics primarily using the *White*, *familiar*, and *looks like you* attributes, all of which are moderately correlated (Fig. 1).

*The fact that *familiar* performed so poorly is not unexpected, as this dimension was specifically chosen to represent a poorly performing attribute model from the model-generation studies.

Conclusion. Modern data-driven methods from machine learning provide new tools for representing and manipulating complex, naturalistic stimuli but are not explicitly designed to model or explain human mental representations. However, applying the same "big data" philosophy to behavioral experiments allows us to align these powerful models with human perception. The deep models of superficial face judgments that we explore in this paper can in turn be used to broaden the range of behavioral data we can collect because they define an infinite set of realistic and psychologically controlled stimuli for a new generation of behavioral experiments.

Materials and Methods

Stimuli. Our experiments make use of 1,004 synthetic yet photorealistic images of faces generated using SG2. The generator network component of SG2 models the distribution of face images conditioned on a 512-dimensional, unit-variance, multivariate normal latent variable. When a vector is sampled from this distribution and passed through the network, it is mapped to a second, intermediate 512-dimensional representation (for which the distribution is unknown), which is in turn fed through multiple layers and ultimately mapped to an output image resembling those from the dataset on which the model was trained. Thus, either of the two 512-dimensional representations can be used for our modeling applications, each associating one fully descriptive (latent) feature vector with each face. We used the latter representation throughout because it yielded superior results in all analyses. Specifically, we use these representations from a pretrained model that was trained on the Flickr-Faces-HQ Dataset (21), containing 70,000 high-quality images at a resolution of 1,024 × 1,024 pixels. Images generated by this model are rendered at the same resolution.

The synthetic faces generated by SG2 are diverse and convincingly realistic in most cases but can occasionally contain visual artifacts that appear odd or even jarring. We minimized these artifacts in our dataset using two strategies. First, SG2 employs a parameter ψ for posttraining image generation that bounds the norm of each multivariate input sample and, as a result, trades off between sample diversity and sample quality. We set ψ to 0.75, which by inspection appeared to jointly maximize the criteria for our purposes. Second, we manually inspected and filtered the generated images, removing all instances that contained obviously distorted faces, multiple faces, hands, localized blotches of color, implausible headdress, or any particularly notable visual artifact. Specifically, we sampled ∼10,000 512-dimensional normal vectors, fed them through the generator network of SG2 to obtain 10,000 candidate face stimuli for our dataset, and took the first ≈ 1,000 that met the criteria for quality. Random examples from the stimulus set are provided in *SI Appendix*.

For the model-validation studies, 50 real face identities were generated by encoding 50 faces from the Chicago Face Database (CFD) (38). The CFD faces used in these real-face experiments were roughly balanced in terms of the four races and two genders available in the main stimulus set. The final set included 12 East Asian, 14 Black, 12 Latin American, and 12 White faces (with equal numbers of male and female faces in each racial group). The 50 faces used in the artificial face experiments were chosen via the same procedure as in the previous attribute modeling studies. Each unique face identity was then transformed along one of 10 perceived attribute dimensions (*age*, *feminine/masculine*, *skinny/fat*, *trustworthy*, *attractive*, *dominant*, *smart*, *outgoing*, *memorable*, and *familiar*) at three levels of the attribute (–0.5 SD, 0 SD, and +0.5 SD from the mean ratings observed in the attribute model studies). This yielded 150 unique images per model-validation study and therefore 3,000 unique images in total.

Participants. For the attribute model studies, we used Amazon Mechanical Turk to recruit a total of 4,157 participants across 10,974 sessions, of which 10,633 (≈ 97%) met our criteria for inclusion (*SI Appendix*, *Data Quality*). Participants identified their gender as female (2,065) or male (2,053), preferred not to say (21), or did not have their gender listed as an option (18). The mean age was ∼39 y old. Participants identified their race/ethnicity as either White (2,935), Black/African American (458), Latinx/a/o or Hispanic (158), East Asian (174), Southeast Asian (71), South Asian (70), Native American/American Indian (31), Middle Eastern (12), Native Hawaiian or Other Pacific Islander (3), or some combination of two or more races/ethnicities (215). The remaining participants either preferred not to say (22) or did not have their race/ethnicity listed as an option (8).

For the model-validation studies, we recruited a total of 1,022 workers from Amazon Mechanical Turk via CloudResearch (47), of which 1,000 (∼98%) met our criteria for inclusion. Of those, 18 participants were excluded for low test–retest reliability, one was excluded for participating in the experiment twice, and three were overrecruited beyond our target sample size of 1,000. Participants identified their gender as female (530) or male (484), preferred not to say (3), or did not have their gender listed as an option (5). The mean age was ∼42 y old. Participants identified their race/ethnicity as either White (781), Black/African American (77), Latinx/a/o or Hispanic (37), East Asian (37), Southeast Asian (11), South Asian (12), Native Hawaiian or Other Pacific Islander (3), or some combination of two or more races/ethnicities (57). The remaining participants either preferred not to say (3) or did not have their race/ethnicity listed as an option (4).

The Institutional Review Board at Princeton University approved both sets of studies. Participants provided informed consent before beginning the study.

Procedure. For the attribute model studies, we used a between-subjects design where participants evaluated faces with respect to each attribute. Participants first consented. Then they completed a preinstruction agreement to answer open-ended questions at the end of the study. In the instructions, participants were given 25 examples of face images in order to provide a sense of the diversity they would encounter during the experiment. Participants were instructed to rate a series of faces on a continuous slider scale where extremes were bipolar descriptors such as "trustworthy" and "not trustworthy." We did not supply definitions of each attribute to participants and instead relied on participants' intuitive notions of each.

Each participant then completed 120 trials with the single attribute to which they were assigned. One hundred of these trials displayed images randomly selected (without replacement) from the full set; the remaining 20 trials were repeats of earlier trials, selected randomly from the 100 unique trials, which we used to assess intrarater reliability. Each stimulus in the full set was judged by at least 30 unique participants.

At the end of the experiment, participants were given a survey that queried what participants believed we were assessing and asked for a self-assessment of their performance and feedback on any potential points of confusion, as well as demographic information such as age, race, and gender. Participants were given 30 min to complete the entire experiment, but most completed it in under 20 min. Each participant was paid $1.50.

The model-validation studies followed an identical procedure to that of the attribute modeling studies above, except where noted below.

Each participant completed one of 20 preregistered experiments involving a pair of one of two different face image types (artificial vs. real) with manipulations along one of 10 different attribute dimensions (*age*, *feminine/masculine*, *skinny/fat*, *trustworthy*, *attractive*, *dominant*, *smart*, *outgoing*, *memorable*, and *familiar*). In each experiment, observers rated 150 unique images (with 30 repeats) along the assigned perceived attribute dimension. Each participant was paid $2.00 due to the longer experiment duration.

Face Attribute Model. To broadly capture human face attribute perception, a model should accurately reproduce human judgments about the attributes of natural faces. More formally, we seek a function $\phi(\cdot)_{PE}$ (what we call a "psychological encoder") that maps from any possible face stimulus $\mathbf{x}_i = \{x_1, \ldots x_m\}$ (i.e., an m-dimensional vector of raw pixel intensities) to a given psychological attribute inference (average judgment for face $\mathbf{x}_i$):

$$\phi(\mathbf{x}_i)_{PE} = y_i. \quad [2]$$

We further define $\phi(\cdot)_{PE}$ as a decomposition of functions:

$$\phi(\mathbf{x})_{PE} = \phi(\phi(\mathbf{x})_F)_S, \quad [3]$$

where $\phi(\mathbf{x}_i)_F = \mathbf{z}_i = \{z_1, \ldots z_d\}$ is a rich feature representation of face stimulus $\mathbf{x}_i$ and $\phi(\cdot)_S$ maps these features to psychological dimensions of interest. This formulation allows us to leverage state-of-the-art neural networks to featurize arbitrary, complex face images.

We then relate these features $\mathbf{z}_i$ to psychological features by assuming that $\phi(\cdot)_S$ is a linear function and thereby implying that each attribute is a 512-dimensional (potentially sparse) vector in the overall feature space. The function $\phi(\cdot)_S$ is learned from human attribute judgment data. In particular, given continuous-scale attribute judgments (i.e., degree of trustworthiness on a scale from 1 to 100), we use linear regression to map 512-dimensional feature vectors $\mathbf{z}_i$ to *average* attribute ratings y_i:

$$y_i = \phi(\mathbf{z}_i)_S = w_0 + w_1 z_1 + \ldots + w_d z_d. \quad [4]$$

In both cases, weight vector $\mathbf{w}_k = \{w_1, \ldots w_d\}$ represents a single attribute k as a linear factor. Therefore, at the heart of our model is a matrix $W \in \mathbb{R}^{k \times d}$, a set of d-dimensional linear factors for each of the k attributes, each obtained by fitting a separate linear model.

The above components of the model enable predictions of attributes to be made for arbitrary face stimuli. We further desire the flexibility to manipulate these attributes for a given face. Because we represent each attribute as a vector $\mathbf{w}_k$ in the feature representation space, we can manipulate each face in this space (i.e., represented by $\mathbf{z}_i$) using vector addition:

$$\mathbf{z}_i' = \mathbf{z}_i + \beta \mathbf{w}_k, \quad [5]$$

where $\mathbf{z}_i'$ is the new transformed face and β is a scalar parameter that controls the strength of the transformation, which can be positive or negative. When $\beta = 0$, $\mathbf{z}_i' = \mathbf{z}_i$, and no transformation takes place. In other words, β scales the attribute vector that is added to the given face representation. Finally, in order to generate a new stimulus corresponding to our transformation, the inverse featurizer (i.e., decoder/generator network of SG2) $\phi^{-1}(\cdot)_F$ is employed to map from features $\mathbf{z}_i$ back to a face stimulus $\mathbf{x}_i$, such that manipulation of face images can be fully described by

$$\mathbf{x}_i' = \phi^{-1}(\mathbf{z}_i')_F = \phi^{-1}(\mathbf{z}_i + \beta \mathbf{w}_k) = \phi^{-1}(\phi(\mathbf{x}_i)_F + \beta \mathbf{w}_k), \quad [6]$$

where $\mathbf{x}_i'$ is the attribute-transformed version of input face $\mathbf{x}_i$.

The success of the above formulation (i.e., good prediction of human attribute judgments for arbitrary faces) is highly dependent on the choice of the feature encoder $\phi(\cdot)_F$, which abstracts over raw pixels and provides the basis for modeling attributes. If the features are not sufficiently expressive, the model will fail to make good predictions of human attribute judgments. Likewise, the ability of the inverse function $\phi^{-1}(\cdot)_F$ to generate face stimuli given their feature representations determines whether attribute-transformed face stimuli will successfully avoid the uncanny valley effect. There are many modern neural networks that could make for a good choice of featurizer $\phi(\cdot)_F$. For example, convolutional neural networks, which learn hierarchies of translation-invariant features, can be trained to classify faces to a high level of accuracy, and their hidden representations can be taken as a feature representation $\mathbf{z}$. However, this method does not yield an inverse function from features back to stimuli and attempts to invert models after the fact often introduce artifacts (48).

Instead, we selected a model primarily aimed at solving the inverse problem alone. GANs are a form of deep latent variable model that learn to model a distribution of images using two components: a generator network that generates images by mapping Gaussian noise to (synthetic) images and a discriminator network that discriminates between real and generated data. When properly trained in a way that balances the two components, the discriminator network forces the generator to produce realistic images, and the discriminator can no longer distinguish between real and generated images. SG2, described earlier, is one of the most successful applications of this model structure and training paradigm; it includes several key improvements that yield highly convincing results (see example faces in *SI Appendix*).

SG2 yields only the inverse function $\phi(\mathbf{x}_i)_F^{-1}$, a learned convolutional generator or decoder function which maps from features to images. In order to apply our model to arbitrary face images outside of our set of 1,004, inverting this function is required. While the authors of SG2 supply their own solution to this problem, we find that it is not accurate enough for our purposes. Instead, we define an encoder function and featurizer $\phi(\mathbf{x}_i)_F$ as an optimization process that searches via gradient descent for the vector input to SG2 that produces an output image with a good likeness to the one we wish to featurize. This likeness is defined as Euclidean distance in the feature space of another external convolutional network pretrained to recognize faces (20). Additionally, because this process is slow, we initialize the image-encoding vector using a first-pass approximation from yet another convolutional neural network that we trained to regress thousands of SG2 image samples to the output vectors that generated them. This encoder is much less accurate, but much faster, and drastically speeds convergence of the slower and more accurate decoding process outlined above.

A summary diagram of our modeling pipeline is provided in *SI Appendix*, Fig. S1.

Model Fitting and Generalization. All linear regression models were fit using the least squares algorithm. Because image feature representations (i.e., vectors of predictors in the design matrix) are high-dimensional, there is a significant risk of overfitting, which could potentially result in suboptimal or meaningless model solutions. To address this, we use ridge regression, which penalizes solutions $\mathbf{w}_k$ that have a large euclidean distance from the $\mathbf{0}$ vector. The strength of this penalty and its influence on the resulting solution is controlled by a free parameter λ. We search for the optimal value of this parameter based on the generalization performance of the model, specifically using 10-fold cross-validation. All reported model scores are averages over those for each of the 10 folds, such that we never report performance on data that was used to fit our models.

Data Availability. The One Million Impressions dataset and all behavioral judgments and synthesized images have been deposited in a GitHub repository (https://github.com/jcpeterson/omi) (49).

ACKNOWLEDGMENTS. A.T. was supported by the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business. Data collection was funded by the Innovation Fund for New Ideas in the Natural Sciences from Princeton University's Dean for Research.

Author affiliations: [a]Department of Computer Science, Princeton University, Princeton, NJ 08540; [b]Booth School of Business, University of Chicago, Chicago, IL 60637; [c]Department of Psychology, Princeton University, Princeton, NJ 08540; and [d]School of Business, Stevens Institute of Technology, Hoboken, NJ 07030

1. F. Farzin, C. Hou, A. M. Norcia, Piecing it together: Infants' neural responses to face and object structure. *J. Vis.* **12**, 6–6 (2012).
2. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
3. C. Frith, Role of facial expressions in social interactions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 3453–3458 (2009).
4. N. N. Oosterhof, A. Todorov, The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11087–11092 (2008).
5. C. A. Sutherland *et al.*, Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).
6. L. A. Zebrowitz, First impressions from faces. *Curr. Dir. Psychol. Sci.* **26**, 237–242 (2017).
7. B. C. Jones *et al.*, To which world regions does the valence-dominance model of social perception apply? *Nat. Hum. Behav.* **5**, 159–169 (2021).
8. A. Todorov, D. Oh, The structure and perceptual basis of social judgments from faces. *Adv. Exp. Soc. Psychol.* **63**, 189–245 (2021).
9. C. A. M. Sutherland *et al.*, Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Pers. Soc. Psychol. Bull.* **44**, 521–537 (2018).
10. A. Todorov, C. Y. Olivola, R. Dotsch, P. Mende-Siedlecki, Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
11. A. Todorov, A. N. Mandisodza, A. Goren, C. C. Hall, Inferences of competence from faces predict election outcomes. *Science* **308**, 1623–1626 (2005).
12. A. C. Little, R. P. Burriss, B. C. Jones, S. C. Roberts, Facial appearance affects voting decisions. *Evol. Hum. Behav.* **28**, 18–27 (2007).
13. I. V. Blair, C. M. Judd, K. M. Chapleau, The influence of Afrocentric facial features in criminal sentencing. *Psychol. Sci.* **15**, 674–679 (2004).
14. J. L. Eberhardt, P. G. Davies, V. J. Purdie-Vaughns, S. L. Johnson, Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychol. Sci.* **17**, 383–386 (2006).
15. C. J. Bohil, H. M. Kleider-Offutt, C. Killingsworth, A. M. Meacham, Training away face-type bias: Perception and decisions about emotional expression in stereotypically black faces. *Psychol. Res.* **85**, 2727–2741 (2021).
16. M. Turk, A. Pentland, Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86 (1991).
17. B. Tiddeman, M. Stirrat, D. Perrett, Towards realism in facial prototyping: Results of a wavelet mrf method. *Proc. Theory Pract Comp. Graph.* **1**, 20–30 (2006).
18. V. Blanz, T. Vetter, "A morphable model for the synthesis of 3d faces" in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, A. Rockwood, Ed. (ACM Press/Addison-Wesley Publishing Co., 1999), pp. 187–194.
19. C. A. Sutherland, G. Rhodes, A. W. Young, Facial image manipulation: A tool for investigating social perception. *Soc. Psychol. Personal. Sci.* **8**, 538–551 (2017).

20. O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep face recognition" in *Proceedings of the British Machine Vision Conference (BMVC)*, X. Xi, M. W. Jones, G. K. L. Tam, Eds. (BMVA Press, 2015), pp. 41.1–41.12.
21. T. Karras, S. Laine, T. Aila, "A style-based generator architecture for generative adversarial networks" in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2018), pp. 4396–4405.
22. T. Karras *et al.*, "Analyzing and improving the image quality of StyleGAN" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 8110–8119.
23. Y. Choi *et al.*, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 8789–8797.
24. A. J. O'Toole, C. D. Castillo, C. J. Parde, M. Q. Hill, R. Chellappa, Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* **22**, 794–809 (2018).
25. Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, AttGAN: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.* **28**, 5464–5478 (2019).
26. P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *2020 IEEE International Joint Conference on Biometrics (IJCB)* (IEEE, 2020), pp. 1–10.
27. V. Bruce, A. Young, Understanding face recognition. *Br. J. Psychol.* **77**, 305–327 (1986).
28. D. Oh, R. Dotsch, J. Porter, A. Todorov, Gender biases in impressions from faces: Empirical studies and computational models. *J. Exp. Psychol. Gen.* **149**, 323–342 (2020).
29. J. R. Collova, C. A. M. Sutherland, G. Rhodes, Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *J. Pers. Soc. Psychol.* **117**, 900–924 (2019).
30. W. A. Bainbridge, P. Isola, A. Oliva, The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **142**, 1323–1334 (2013).
31. G. Rhodes, The evolutionary psychology of facial beauty. *Annu. Rev. Psychol.* **57**, 199–226 (2006).
32. C. P. Said, A. Todorov, A statistical model of facial attractiveness. *Psychol. Sci.* **22**, 1183–1190 (2011).
33. I. J. Goodfellow *et al.*, "Generative adversarial networks." in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2014), vol. 27.
34. J. E. Martinez, F. Funk, A. Todorov, Quantifying idiosyncratic and shared contributions to judgment. *Behav. Res. Methods* **52**, 1428–1444 (2020).
35. A. Song, L. Linjie, C. Atalla, G. Cottrell, "Learning to see people like people: Predicting social impressions of faces." in *39th Annual Meeting of the Cognitive Science Society* (CogSci, 2017).
36. C. J. Parde, Y. Hu, C. Castillo, S. Sankaranarayanan, A. J. O'Toole, Social trait information in deep convolutional neural networks trained for face identification. *Cogn. Sci.* **43**, e12729 (2019).
37. M. L. Willis, R. Palermo, D. Burke, Social judgments are influenced by both facial expression and direction of eye gaze. *Soc. Cogn.* **29**, 415–429 (2011).
38. D. S. Ma, J. Correll, B. Wittenbrink, The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135 (2015).
39. A. Kiyonaga, J. M. Scimeca, D. P. Bliss, D. Whitney, Serial dependence across perception, attention, and memory. *Trends Cogn. Sci.* **21**, 493–497 (2017).
40. P. Korshunov, S. Marcel, Deepfakes: A new threat to face recognition? Assessment and detection. arXiv [Preprint] (2018). https://arxiv.org/abs/1812.08685 (Accessed 20 December 2018).
41. D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art" in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021), pp. 1–6.
42. G. Tang *et al.*, Detection of GAN-synthesized image based on discrete wavelet transform. *Secur. Commun. Netw.* **2021**, 5511435 (2021).
43. R. Winick, Intellectual property, defamation and the digital alteration of visual images. *Columbia VLA J. Law Arts* **21**, 143 (1996).
44. L. B. A. Potter, Altered realities: The effect of digital imaging technology on libel and right of privacy. *Hast. Comm. Ent. LJ* **17**, 495 (1994).
45. D. Almeida, K. Shmarko, E. Lomas, The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI Ethics*, 10.1007/s43681-021-00077-w (2021).
46. E. Learned-Miller, V. Ordóñez, J. Morgenstern, J. Buolamwini, Facial recognition technologies in the wild: A call for a federal office (2020). https://www.ajl.org/federal-office-call. Accessed 6 April 2022.
47. L. Litman, J. Robinson, T. Abberbock, TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods* **49**, 433–442 (2017).
48. A. Dosovitskiy, T. Brox, "Generating images with perceptual similarity metrics based on deep networks" in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, Inc., 2016), pp. 658–666.
49. J. C. Peterson, S. Uddenberg, J. W. Suchow, One Million Impressions (OMI) Dataset. GitHub. https://github.com/jcpeterson/omi. Deposited 14 March 2022.