Capturing and modifying the perceived traits of all possible faces

Joshua C. Peterson[1]*, Stefan Uddenberg[2,3], Thomas L. Griffiths[1,4], Alexander Todorov[2,4], Jordan W. Suchow[5]

[1]Department of Computer Science, Princeton University

[2]Booth School of Business, University of Chicago

[3]Princeton Neuroscience Institute, Princeton University

[4]Department of Psychology, Princeton University

[5]School of Business, Stevens Institute of Technology

*Corresponding author

Email: joshuacp@princeton.edu

Keywords: Face Perception, Generative Neural Networks

Abstract

The diversity in appearance of human faces and their naturalistic viewing conditions give rise to an expansive stimulus space over which humans perceive numerous psychological traits (*e.g.,* perceived trustworthiness). Current scientific models characterize only few of these traits, and over only a tiny fraction of possible faces. Here we show that generative image models from machine learning combined with over 1 million human judgments can capture more than 30 traits over a near-infinite set of face stimuli. This makes it possible to then seamlessly infer and manipulate the psychological traits of arbitrary face photograph inputs and generate infinite synthetic photorealistic face stimuli along those dimensions. The predictive accuracy of our model approaches human inter-rater reliability, which our simulations suggest would not have been possible with previous datasets having fewer faces, fewer trait ratings, or using low-dimensional feature representations.

Capturing and modifying the perceived traits of all possible faces

Faces are among the most important stimuli that people encounter—they are recognized by infants long before other objects in their environment (Farzin, Hou, & Norcia, 2012), recruit specialized circuits in the brain (Kanwisher, McDermott, & Chun, 1997), and are fundamental to social interaction (Frith, 2009). Central to our experience with faces are the psychological traits which we assign to them, often implicitly. These include traits that are "read off", describing largely objective aspects of faces (e.g., age, adiposity), and those that are "read into", such as how trustworthy a person seems (Oosterhof & Todorov, 2008). Though the inferences of the latter traits are more subjective and generally inaccurate, they are similarly psychologically consistent across people (Oosterhof & Todorov, 2008; C. A. Sutherland et al., 2013; Zebrowitz, 2017) around the globe (C. A. M. Sutherland et al., 2018; Todorov & Oh, 2021) and have important consequences (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015) ranging from electoral success (Little, Burriss, Jones, & Roberts, 2007; Todorov, Mandisodza, Goren, & Hall, 2005) to sentencing decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). Because any face can be judged with respect to such traits, these psychological dimensions are universal in that they are implicitly defined over the space of nearly all possible faces, contexts, and observation conditions, a highly diverse landscape of stimuli. For this reason, capturing this psychological content in its entirety, which both forms the basis of scientific models of face perception and defines the scope of downstream applications such as training people to overcome stereotypes (Bohil, Kleider-Offutt, Killingsworth, & Meacham, 2020), is a challenging task.

Given the importance of face trait perception, numerous techniques for scientific modeling of faces have proliferated, which can be broadly organized into two basic approaches. The first includes those that extrapolate from face photographs, often related via landmark annotations (Tiddeman, Stirrat, & Perrett, 2006; Turk & Pentland, 1991). The second includes those that employ artificial faces using parametric 3D face meshes (Blanz & Vetter, 1999). Photographs offer greater realism, but are

limited by small datasets of face stimuli that serve as the basis for interpolation, as well as the interpolation algorithms themselves. Artificially generated faces are not subject to these limitations, but incur a high cost in diversity and realism. Neither of these approaches captures the richness and diversity of human faces. In recent years, machine learning methods have emerged which learn to model faces from massive databases of face photographs scraped from photo repositories online (Karras, Laine, & Aila, 2018; Karras et al., 2020; Parkhi, Vedaldi, & Zisserman, 2015). These methods provide 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 extremely high-dimensional vectors produced via black-box optimization algorithms (O'Toole, Castillo, Parde, Hill, & Chellappa, 2018).

We show that the key to unlocking the scientific potential of these models, as well as downstream applications, are large-scale datasets of human behavior not previously attainable using traditional laboratory experiments. In particular, such large datasets provide sufficient evidence to determine the robust mapping between expressive high-dimensional representations from machine learning models and human mental representations of face traits. This mapping could be computed for any meaningful psychological trait. We focus on three types of traits: relatively objective characteristics such as age and adiposity, subjective characteristics such as perceived "trustworthiness", and even more subjective characteristics such as "familiarity." We exploit this mapping to manipulate psychological perception of these traits for arbitrary face images, allowing us for example to increase or decrease the likely perceived "trustworthiness" of a person's image.

To this end, we used online crowdsourcing to obtain (perceived) trait ratings for just over 1,000 naturalistic face stimuli for 34 traits (at least 30 ratings from unique participants per trait and per stimulus), for a total of 1,020,000 human judgments. These perceived traits, especially the more subjective ones, have no necessary correspondence to the actual identities, attitudes, or competencies of the persons

depicted (*e.g.,* a trustworthy person may be wrongly assumed to be untrustworthy because of their appearance). Instead, our measurements capture systematic biases and stereotypes about traits shared by the population of raters. A detailed summary of these ratings and inter-trait relationships is provided in the supplement.

To explore the structure of the overall trait space, we first computed the correlation between the mean face ratings for each pair of traits, the results of which are shown in Figure 2. Many traits were highly correlated, including *happy-outgoing* ($r = .93$) and *dominant-trustworthy* ($r = -.81$), while others are largely unrelated, including *smart-attractive* ($r = .01$), *smart-trustworthy* ($r = .02$), *smart-skinny/fat* ($r = -.02$), and *skinny/fat-trustworthy* ($r = -.02$). Although some of these correlations are consistent with prior findings, many are not (Todorov & Oh, 2021). For example, though judgments of trustworthiness and dominance tend to be negatively correlated, the magnitude of the correlation is generally small (Oh, Dotsch, Porter, & Todorov, 2020). Similarly, judgments of smartness or competence tend to be highly positively correlated with judgments of attractiveness and trustworthiness. This apparent discrepancy is due to the fact that most other face datasets include only adult faces. In fact, the correlational structure of judgments of children's faces is different from the structure of judgments of adult faces (Collova, Sutherland, & Rhodes, 2019). However, we find only limited support for this hypothesis (see Fig. S10 in the supplement).

To model each trait, we start with the high-dimensional representation vectors $\mathbf{z}_i = \{z_1, \ldots, z_d\}$ assigned to each face i in our stimulus set using a state-of-the-art generative adversarial network (GAN) (Goodfellow et al., 2014; Karras et al., 2018, 2020), a model that has learned a mapping from each such vector to an image through extensive training on a large database of faces. We then model each psychological trait, measured via average trait ratings y_i as a linear combination of features: $y_i = w_0 + w_1 z_1 + \ldots + w_d z_d$, where the vector of weights $\mathbf{w}_k = \{w_1, \ldots, w_d\}$ represents trait k as a linear dimension cross-cutting through the overall representation space and is fit using cross-validated, $L2$-regularized linear regression. Average cross-validated (i.e., out-of-sample) model performance for each trait is reported in Figure 1. Prediction

for most traits was reasonably successful, with most R^2 values ranging from above 0.5 to almost 0.8, with traits *typical* and *familiar* being the exceptions.

Because participants partly disagree in their appraisals of each face (Martinez, Funk, & Todorov, 2020), perfect prediction is not possible. To estimate the resulting prediction ceiling, and thus better understand the performance of our models, we also computed reliability scores for each trait by estimating the extent to which human participants agree. In particular, we compute the split-half reliability for each trait by averaging the squared correlations between the averages of 100 random splits of the ratings for each image. In all cases, we found these reliability scores to be close to the prediction performance of our models, albeit higher, indicating that better prediction is possible. Interestingly, our model of *familiarity* showed the smallest gap between performance and reliability, indicating that the small R^2 value is not due to a poor model or input features. Instead, it seems more likely that *familiarity* more than other traits is based on both a shared concept or experience (determining what can be systematically predicted across different participants), and a much larger personal concept or experience (which can only be predicted at the level of individuals). This is corroborated by the similar effect for the trait *looks like you*, which can only be predicted at the aggregate level to the extent that our participant pool shares broad facial features.

Next, we investigated the relationship between the number of faces rated and predictive performance (Fig. S11 in the supplement, left panel). Performance curves were generated by fitting models for each of 30 random samples of images with sizes ranging from 100 to 1,000. Interestingly, significantly fewer images than most traits were required to best capture the trait *feminine/masculine*. For all other traits, adding additional images always improves performance, indicating that at least 1,000 images are required to capture most traits for naturalistic faces. Next, we investigated the relationship between the number of ratings (*i.e.,* unique participants) obtained for each face image and predictive performance (Fig. S11 in the supplement, center panel). Performance curves were generated by fitting models for each of 30 random samples of

unique ratings for each image with sizes ranging from 5 to 30. Aside from traits *feminine/masculine* and *age*, which elicit less disagreement, performance increases as the number of ratings increases for all traits. This indicates that at least 30 ratings are required to best capture most traits. Gains due to the number of ratings are diminishing, but at a slower rate than gains due to the number of faces (Fig. S11 in the supplement, left panel). Finally, we investigated the relationship between the number of image features (512 total) and predictive performance. Results are shown in the the right panel of Fig. S11 in the supplement. 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—around 100 principal dimensions—but is improved marginally with more dimensions in some cases. This indicates that at least 100 dimensions of the latent feature space are required to adequately capture psychological traits for faces.

Next, we apply our model to the manipulation of the psychological traits of input faces. Since the learned trait vectors correspond to linear dimensions, we can manipulate an arbitrary face represented by features $\mathbf{z}_i$ with respect to trait k using vector arithmetic: $\mathbf{z}_i + \beta \times \mathbf{w}_k$, where β is a scalar controlling the positive or negative modulation of the trait. We apply a symmetric range of β around 0 to each trait vector to manipulate a series of base face representations in both the negative and positive directions, and decode the results for visualization using the decoder/generator component of the neural network that was also used to derive representations (see supplement for more details). The results are shown in Fig. 3, and reveal strikingly smooth and effective manipulations along each trait dimension. For example, modulating *trustworthiness* in the given examples increases eye contact with the camera, degree of smiling, and alters face shape and facial hair. Trait manipulation involves more than one appearance dimension. For example, increasing *smartness* may involve adding glasses and/or changing facial expression. Increasing *outgoingness* increases smiling as expected, but also gives glasses a more rounded and cartoonish appearance. Other dimensions allow for greater extrapolation. For example, faces can

be made much *skinnier* or *fatter* than any examples in our dataset, yet still maintain a realistic appearance. Faces with strongly manipulated *happiness* also resemble convincing caricatures.

We set out to develop a comprehensive model of trait perception that can psychologically interpret and manipulate nearly any possible face image. With no explicit featurization or interpolation algorithm, we were able to accomplish this in a fully data-driven manner with relatively high accuracy and generalization. We provided strong quantitative evidence that large datasets of both face stimuli and intra-stimulus ratings are necessary to achieve this. Our qualitative results speak for themselves, resulting in convincing psychological trait manipulations of realistic face photos using simple vector arithmetic. Moreover, our pipeline provides a general formula for capturing any psychological trait that can be measured via image annotations. Further, because the models of traits are in the same multi-dimensional space, their similarity is immediately given, allowing for testing of specific hypotheses about the relation between psychological traits, predicting novel traits based on their relationships with models of existing traits, and controlling for shared variance between traits.

Importantly, while the primary goal of this work is to support scientific modeling and productive application, the model developed here and many possible extensions of it introduce a new class of ethical concerns. In particular, the manipulation of arbitrary faces, especially along dimensions such as perceived trustworthiness, has the potential for malicious use, and it is precisely the innovations we offer in this work that drastically simplify such efforts. 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.

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 model that we explore in this paper can in turn be

used to broaden the range of behavioral data we can collect, as it defines an infinite set of realistic and psychologically controlled stimuli for a new generation of behavioral experiments. It is our hope that the scientific community can leverage this new resource and more general method to further increase the scope of computational psychology.

Methods

Stimuli

Our face stimuli were 1,004 synthetic yet photorealistic images of highly diverse and naturalistic faces curated from a larger set that was generated using StyleGAN2 (Karras et al., 2018, 2020), a state-of-the-art Generative Adversarial Network (GAN). In particular, the model we utilized was pretrained by its authors on the Flickr-Faces-HQ Dataset (Karras et al., 2018), comprising 70,000 high-quality images at a resolution of 1024×1024 pixels. Images generated by this model are rendered at the same resolution, and largely reflect its diversity. Additional details of these stimuli and our curation protocol are provided in the supplement.

Participants

We recruited a total of 12,043 workers from Amazon Mechanical Turk, 11,655 of which (approximately 97%) met our criteria for inclusion (see Data Quality section). Participants identified their gender as “female” (5,675) or “male” (6,230). The remaining participants either preferred not to say (104) or did not have their gender listed as an option (34). The mean age was approximately 41 years old. Participants identified their race/ethnicity as either “White” (8,681),“Black/African American” (974), “Latinx/a/o or Hispanic” (475), “East Asian” (641), “Southeast Asian” (245), “South Asian” (255), “Native American/American Indian” (58), “Middle Eastern” (24), “Native Hawaiian or Other Pacific Islander” (6), or some combination of 2 or more races/ethnicities (576). The remaining participants either preferred not to say (80) or did not have their race/ethnicity listed as an option (28).

Procedure

We used a between-subjects design where participants evaluated faces with respect to each trait. Participants first consented and were additionally required to complete a pre-instruction agreement to answer open-ended questions at the end of the study. They were then given 25 examples of face images to provide a sense of the diversity they would encounter during the experiment, and were instructed to rate a series of faces on a continuous slider scale where extremes were bipolar descriptors such as "trustworthy" to "not trustworthy". We did not supply definitions of each trait to participants, and instead relied on participants' intuitive notions for each.

Each participant then completed 120 trials for the single attribute to which they were assigned. 100 of these trials comprised random unique images from the full set, and the remaining 20 trials were repeats of earlier trials (selected randomly from the 100 unique trials) which we used to assess intra-rater reliability. We ensured that each unique stimulus in the full set was judged by at least 30 unique participants.

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

References

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Bohil, C. J., Kleider-Offutt, H. M., Killingsworth, C., & Meacham, A. M. (2020). Training away face-type bias: perception and decisions about emotional expression in stereotypically black faces. *Psychological Research*, 1–15.

Collova, J. R., Sutherland, C. A., & Rhodes, G. (2019). Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology*, *117*(5), 900.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*(5), 383–386.

Farzin, F., Hou, C., & Norcia, A. M. (2012). Piecing it together: infants' neural responses to face and object structure. *Journal of Vision*, *12*(13), 6–6.

Frith, C. (2009). Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3453–3458.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks, 2019 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396–4405).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020).

Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110–8119).

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18–27.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, 1–17.

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General*, *149*(2), 323.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, *22*(9), 794–809. doi: 10.1016/j.tics.2018.06.006

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.

Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial First Impressions Across Culture: Data-Driven Modeling of Chinese and British Perceivers' Unconstrained Facial Impressions. *Personality & Social Psychology Bulletin*, *44*(4), 521–537. doi: 10.1177/0146167217744194

Tiddeman, B., Stirrat, M., & Perrett, D. (2006). Towards realism in facial prototyping: results of a wavelet mrf method. In *Proc. theory and practice of computer graphics* (Vol. 1, pp. 20–30).

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments

from faces. *Advances in Experimental Social Psychology*, *63*, 189–245. doi: 10.1016/bs.aesp.2020.11.004

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. doi: 10.1146/annurev-psych-113011-143831

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

Zebrowitz, L. A. (2017). First Impressions From Faces. *Current Directions in Psychological Science*, *26*(3), 237–242. doi: 10.1177/0963721416683996

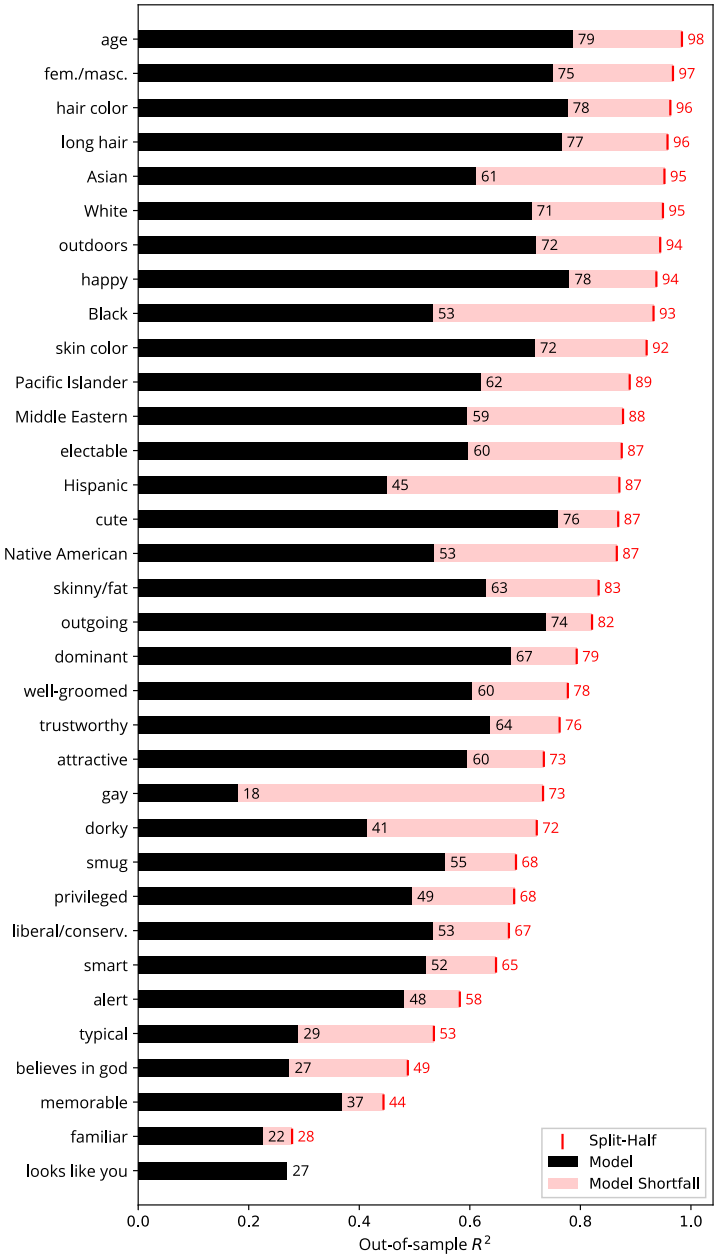

Figure 1. Model performance (R^2) for each trait compared to inter-subject reliability. Black bars show average 10-fold cross-validation performance models for each trait. Red markers show split-half reliability: average squared correlations between 100 random splits of the rating data for each trait. Red bars show the shortfall of each of our models.

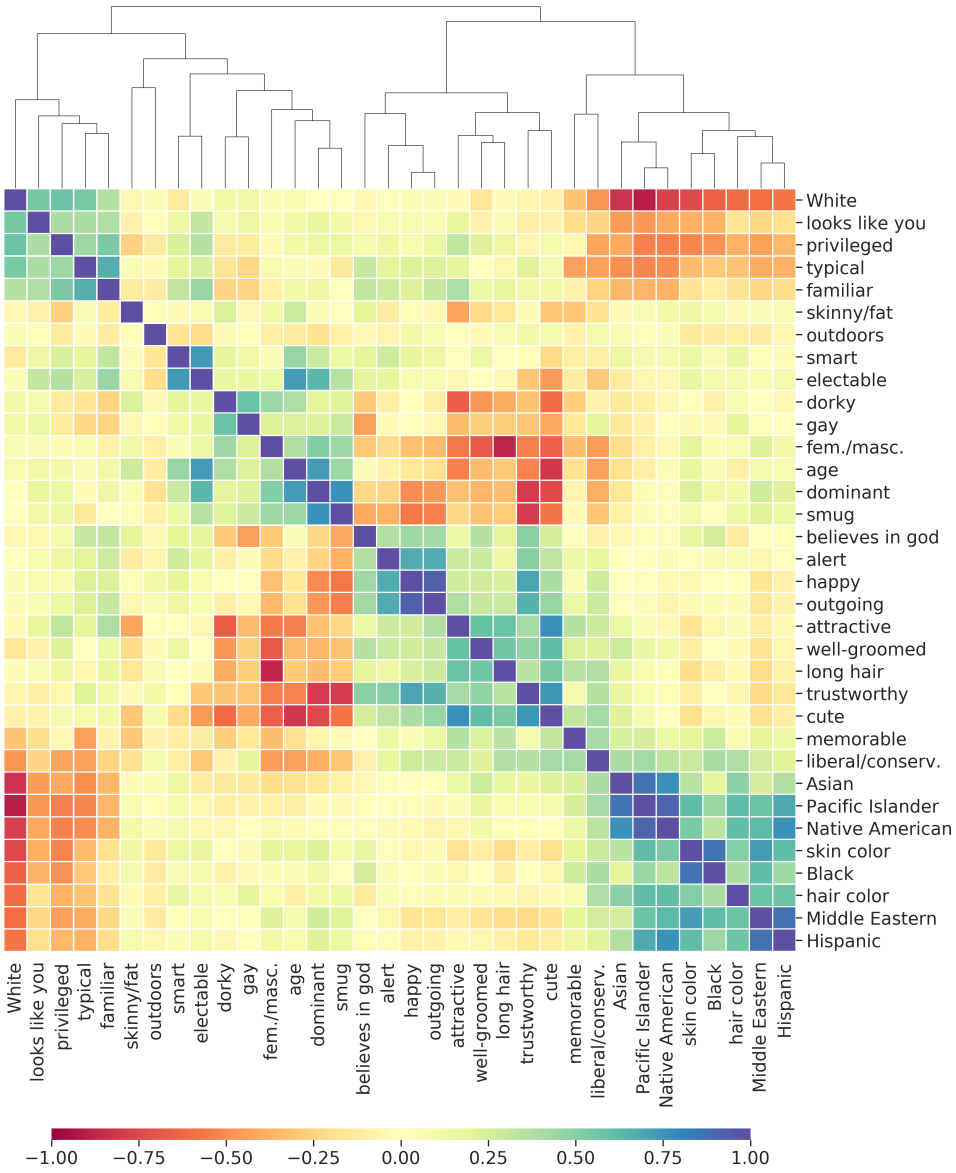

Figure 2. Correlation matrix for 34 average trait ratings for each of 1,000 faces. Rows and columns are arranged according to a hierarchical clustering of the signed correlation values.

Figure 3. Manipulation of two base faces along sample trait dimensions.

Supplementary Information for

"Capturing and modifying the perceived traits of all possible faces"

Joshua C. Peterson[1]*, Stefan Uddenberg[2,3], Thomas L. Griffiths[1,4], Alexander Todorov[2,4], Jordan W. Suchow[5]

[1]Department of Computer Science, Princeton University

[2]Booth School of Business, University of Chicago

[3]Princeton Neuroscience Institute, Princeton University

[4]Department of Psychology, Princeton University

[5]School of Business, Stevens Institute of Technology

*Corresponding author

Email: joshuacp@princeton.edu

Keywords: Face Perception, Generative Neural Networks

Supplementary Information for

"Capturing and modifying the perceived traits of all possible faces"

Supplemental Methods

Additional Stimulus Details

Our experiments make use of 1,004 synthetic yet photorealistic images of faces generated using StyleGAN2 (Karras, Laine, & Aila, 2018; Karras et al., 2020), a state-of-the-art Generative Adversarial Network (GAN) architecture, hereafter referred to as 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 to 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 use the latter representation throughout since it yielded superior results in all of our analyses. Specifically, we utilize these representations from a pretrained model supplied by the authors that was trained on Flickr-Faces-HQ Dataset (Karras et al., 2018), containing 70,000 high-quality images at a resolution of 1024×1024 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 post-training 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 approximately 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 the supplement.

Facial Trait Model Details

To broadly capture human face trait perception, we want a model that can accurately reproduce human judgments about the traits 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.,* m-dimensional vectors of raw pixel intensities) to a given psychological trait (average judgment for face $\mathbf{x}_i$):

$$\phi(\mathbf{x}_i)_{PE} = y_i. \tag{1}$$

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

$$\phi(\mathbf{x})_{PE} = \phi(\phi(\mathbf{x})_F)_S, \tag{2}$$

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 explain this function in more detail later in this section.

We then relate these features $\mathbf{z}_i$ to psychological ones, assuming that $\phi(\cdot)_S$ is a linear function, and thus implying that each psychological trait is a 512-dimensional (potentially sparse) vector in the overall feature space (we provide support for this assumption later). We learn the function $\phi(\cdot)_S$ from our human trait judgment data. In particular, given continuous-scale trait 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* trait ratings y_i:

$$y_i = \phi(\mathbf{z}_i)_S = w_0 + w_1 z_1 + ... + w_d z_d. \tag{3}$$

In both cases, weight vector $\mathbf{w}_k = \{w_1, ..., w_d\}$ represents a single trait 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 k psychological traits, each obtained by fitting separate linear models.

The above components of our model allow for predictions of traits to be made for arbitrary face stimuli, but we would also like the flexibility to manipulate these traits for a given face. Since we represent each trait 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:

$$\mathbf{z}'_i = \mathbf{z}_i + \beta \times \mathbf{w}_k, \tag{4}$$

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 trait 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:

$$\mathbf{x}'_i = \phi^{-1}(\mathbf{z}'_i)_F = \phi^{-1}(\mathbf{z}_i + \beta \times \mathbf{w}_k) = \phi^{-1}(\phi(\mathbf{x}_i)_F + \beta \times \mathbf{w}_k), \tag{5}$$

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

The success of the above formulation (*i.e.,* good prediction of human trait judgments for arbitrary faces) is highly dependent on the choice of the feature encoder $\phi(\cdot)_F$, which allows us to abstract over raw pixels and provides the basis for modeling traits. If the features are not rich enough, we will fail to make good predictions of

human trait judgments. Likewise, the ability of the inverse function $\phi^{-1}(\cdot)_F$ to generate face stimuli given their feature representations determines whether trait-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 from features back to stimuli, and attempts for inverting models after the fact often introduce artifacts (Dosovitskiy & Brox, 2016).

Instead, we start with a model that is primarily aimed at the inverse problem alone. Generative adversarial networks 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 trained correctly 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 fake ones. SG2, previously described in the stimulus generation section, is one of the most successful applications of this model structure and training paradigm to date, having implemented several key improvements that yield highly convincing results (see examples in Fig. 2).

Importantly, 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 supply their own solution to this problem, we find that it is not accurate enough for our purposes. Instead, we define our encoder function and featurizer $\phi(\mathbf{x}_i)_F$ as an optimization process which searches for the vector input to SG2 via gradient descent that produces an output image like the one we want to featurize. This likeness is defined as euclidean distance in the feature space of another external convolutional network pretrained to recognize faces (Parkhi, Vedaldi,

& Zisserman, 2015). 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.

To summarize, we model a set of psychological traits (*e.g.,* trustworthiness) over a large possible space of naturalistic face stimuli as vectors in a rich feature space provided by a deep generative neural network. This allows us to generalize trait predictions to arbitrary new face stimuli, modulate traits for any face stimulus, and generate new stimuli with estimated traits.

Model Fitting & Generalization

All linear regression models were fit using the least squares algorithm. Since 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 sub-optimal 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. In the following results, 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.

Supplemental Results

Data Quality

Intra-rater (*i.e.,* test-retest) reliability was reasonably high on average across all of the tested traits, as shown in Table 1 (keeping in mind that each observer re-rated 20% of all seen stimuli). Most individual participants showed high levels of reliability, as can

be seen in Figure 3's left skew. Participants were not included in the models if their intra-rater reliability was below 0. Due to this conservative exclusion criterion, only 4.7% of all participants tested were excluded from our trait models. The traits eliciting the lowest reliability (although still reasonably high) were familiarity and typicality, while those eliciting the highest were less subjective traits such as age and gender. All other traits had a mean reliability above 0.6.

Trait Ratings and Inter-Trait Relationships

Figures 5 through 9 show the faces with the ten highest and ten lowest mean ratings for each perceived trait. Selections for less subjective traits such as *age*, *fat/skinny*, and *feminine/masculine* are straightforward, although there are some interesting observations. For example, the most *masculine*-looking men are not necessarily the most *dominant*-looking ones, who tend to look younger and have stronger jawlines. More subjective traits also show clear patterns. Consistent with prior findings, children's faces look more *trustworthy* (Berry & McArthur, 1985, 1986; Montepare & Zebrowitz, 1998), while straight-faced *masculine*-looking faces with sunglasses appear least trustworthy. *Feminine*-looking faces were rated as more *attractive* (Said & Todorov, 2011), while older *masculine*-looking faces wearing glasses were rated as the least *attractive*. Faces rated as especially *smart* also often wore glasses (Sutherland et al., 2013), but appeared young to middle-aged, while the most *outgoing*-looking faces were often smiling. Finally, even perceived traits with the lowest intra-rater reliability in the full set are reasonably interpretable. For example, young to middle-aged white *masculine*-looking faces were rated as more *typical*, while less *typical* faces were more diverse in terms of their race and gender. This is to be expected, given the fact that our MTurk sample reflected the demographics of the platform at large, and was therefore predominantly white. *Feminine*-looking faces were judged as looking more *familiar*, while less *familiar* faces were also more diverse. Our goal is to model the full extent of these effects, and not just what can be inferred qualitatively from inspecting such examples.

References

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, *48*(2), 312.

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: the impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*(1), 3.

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems* (pp. 658–666).

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks, 2019 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396–4405).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110–8119).

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology*, *30*, 93–161.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science*, *22*(9), 1183–1190.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.

Perceived Trait	Median Reliability	% Participants Excluded
trustworthy	0.734	5.895
attractive	0.799	3.207
dominant	0.786	3.106
smart	0.740	4.615
age	0.955	0.303
fem./masc.	0.937	4.050
skinny/fat	0.778	0.000
typical	0.656	4.969
happy	0.867	1.286
familiar	0.520	16.393
outgoing	0.782	1.905
memorable	0.691	3.115
well groomed	0.794	1.274
long hair	0.933	0.322
smug	0.746	2.160
dorky	0.740	4.334
skin color	0.874	0.629
hair color	0.918	0.625
alert	0.697	2.532
cute	0.873	0.000
privileged	0.763	3.145
liberal	0.724	2.769
Asian	0.904	0.637
Middle Eastern	0.811	0.943
Hispanic	0.807	0.000
Pacific Islander	0.847	1.558
Native American	0.812	2.950
Black	0.894	7.599
white	0.919	0.312
looks like you	0.826	7.207
gay	0.721	4.545
electable	0.869	1.572
believes in God	0.674	2.839
outdoors	0.870	0.645

Table 1

Intra-rater test-retest reliability for all participants and participant exclusion statistics for each of the collected traits.

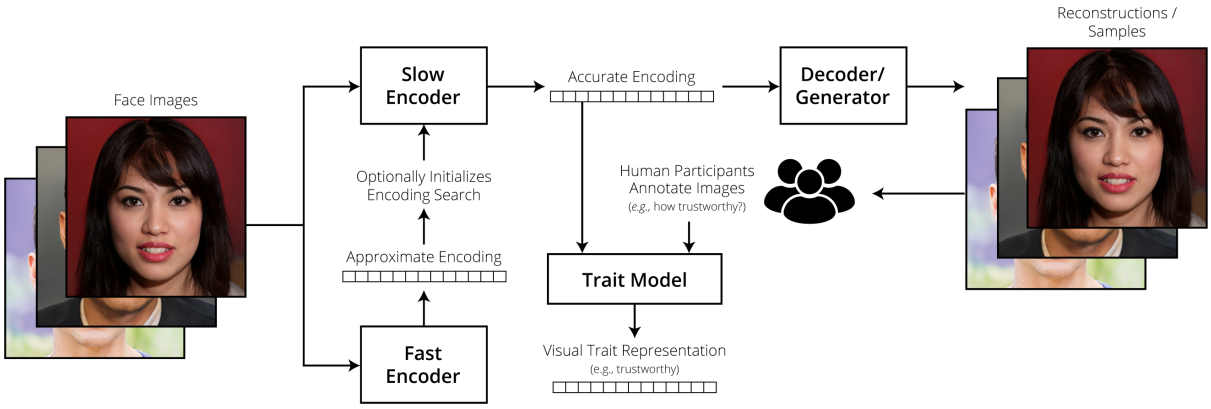

Figure 1. General pipeline for modeling arbitrary psychological traits. Encodings either for generated face images or real ones inferred via our encoding models can be manipulated using traits learned in the same vector space as the encodings.

Figure 2. Random example stimuli from our dataset of 1,000 curated synthetic face images generated using StyleGAN2 (Karras et al., 2018, 2020) for use in all of our experiments.

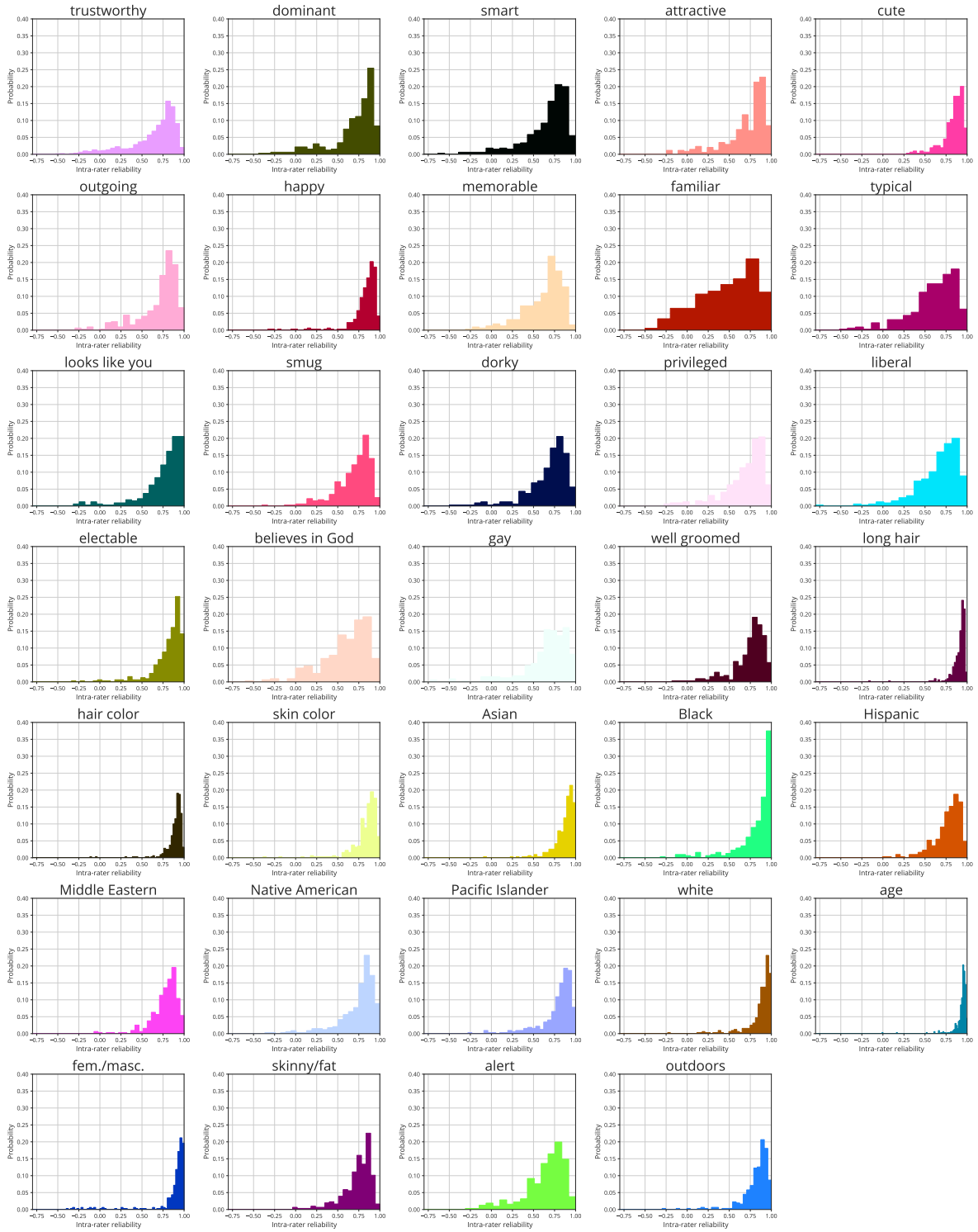

Figure 3. Intra-rater reliability distributions for each measured trait.

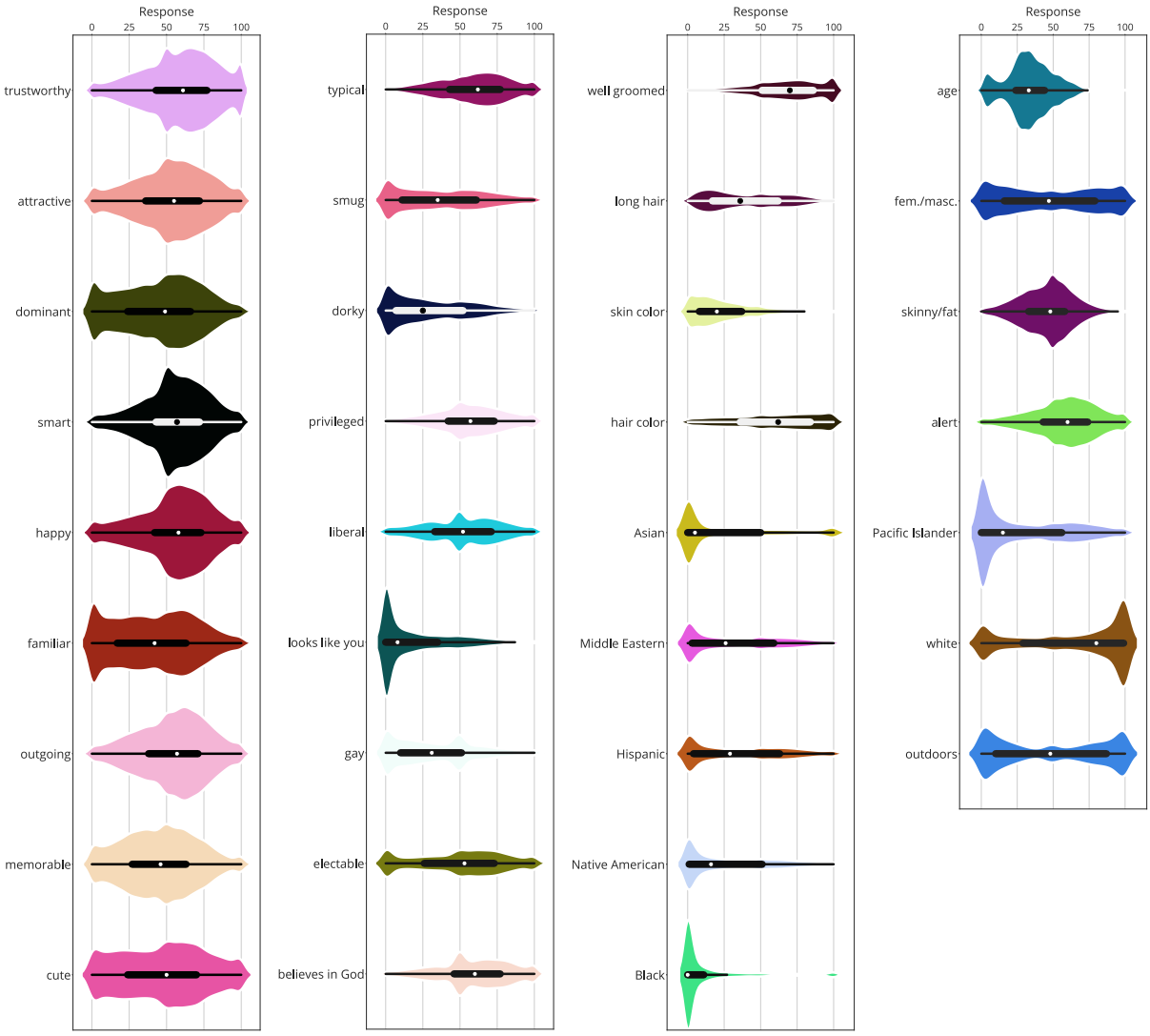

Figure 4. Distribution of raw responses given by participants for images rated along each trait. Boxplots at the center of each distribution represent the median as a white/black dot (depending on the contrast), the interquartile range as the thick opposite-colored line, and the remainder of the distribution (sans outliers) as the thinner lines (i.e., the "whiskers").

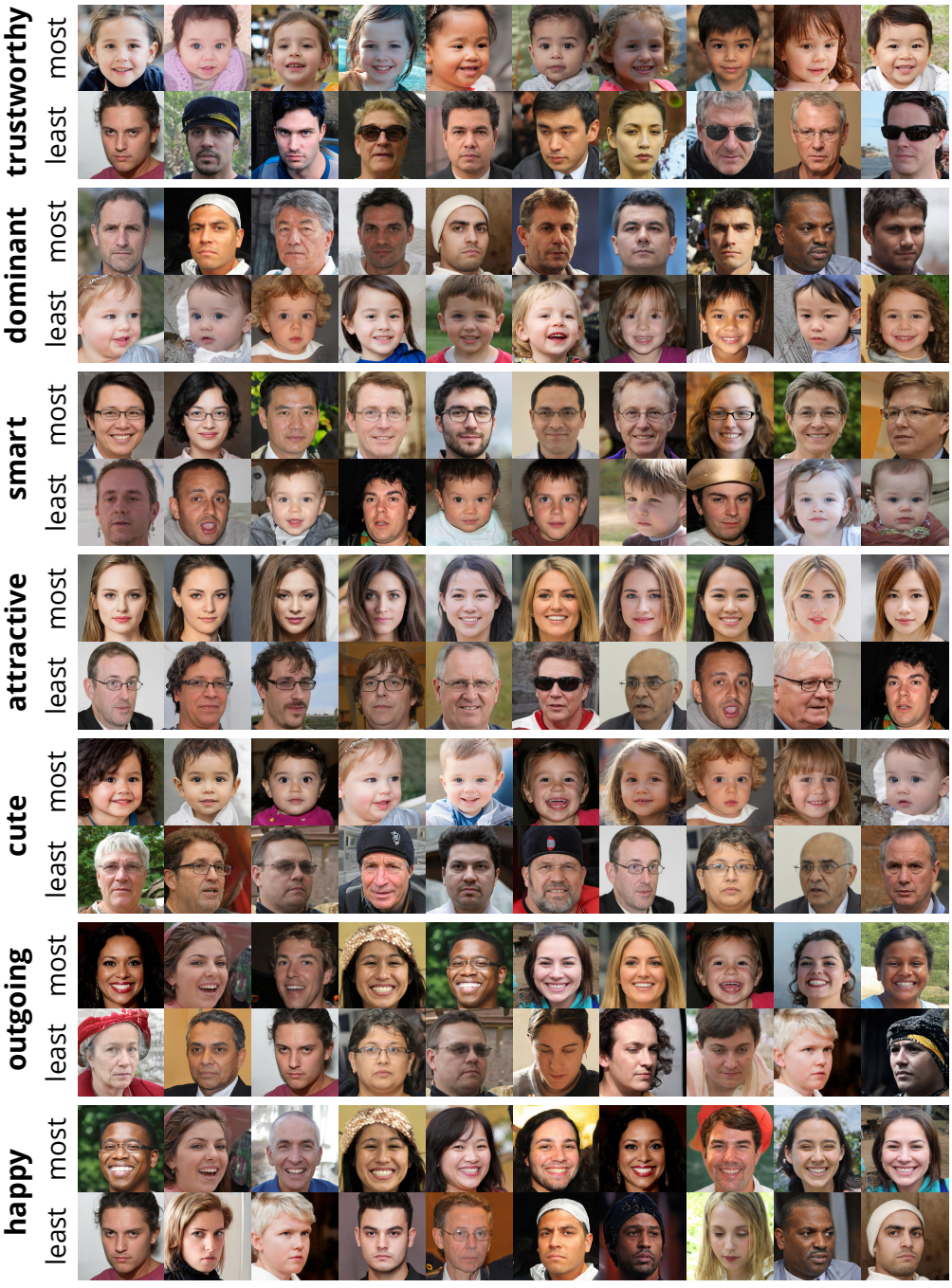

Figure 5. Ten most and ten least congruent faces (highest/lowest mean ratings) for traits *trustworthy*, *dominant*, *smart*, *attractive*, *cute*, *outgoing* and *happy*.

Figure 6. Ten most and ten least congruent faces (highest/lowest mean ratings) for traits *memorable*, *familiar*, *typical*, *looks like you*, *smug*, *dorky*, and *privileged*.

Figure 7. Ten most and ten least congruent faces (highest/lowest mean ratings) for traits *liberal*, *electable*, *believes in God*, *gay*, *well groomed*, *long haired*, and *hair color*.

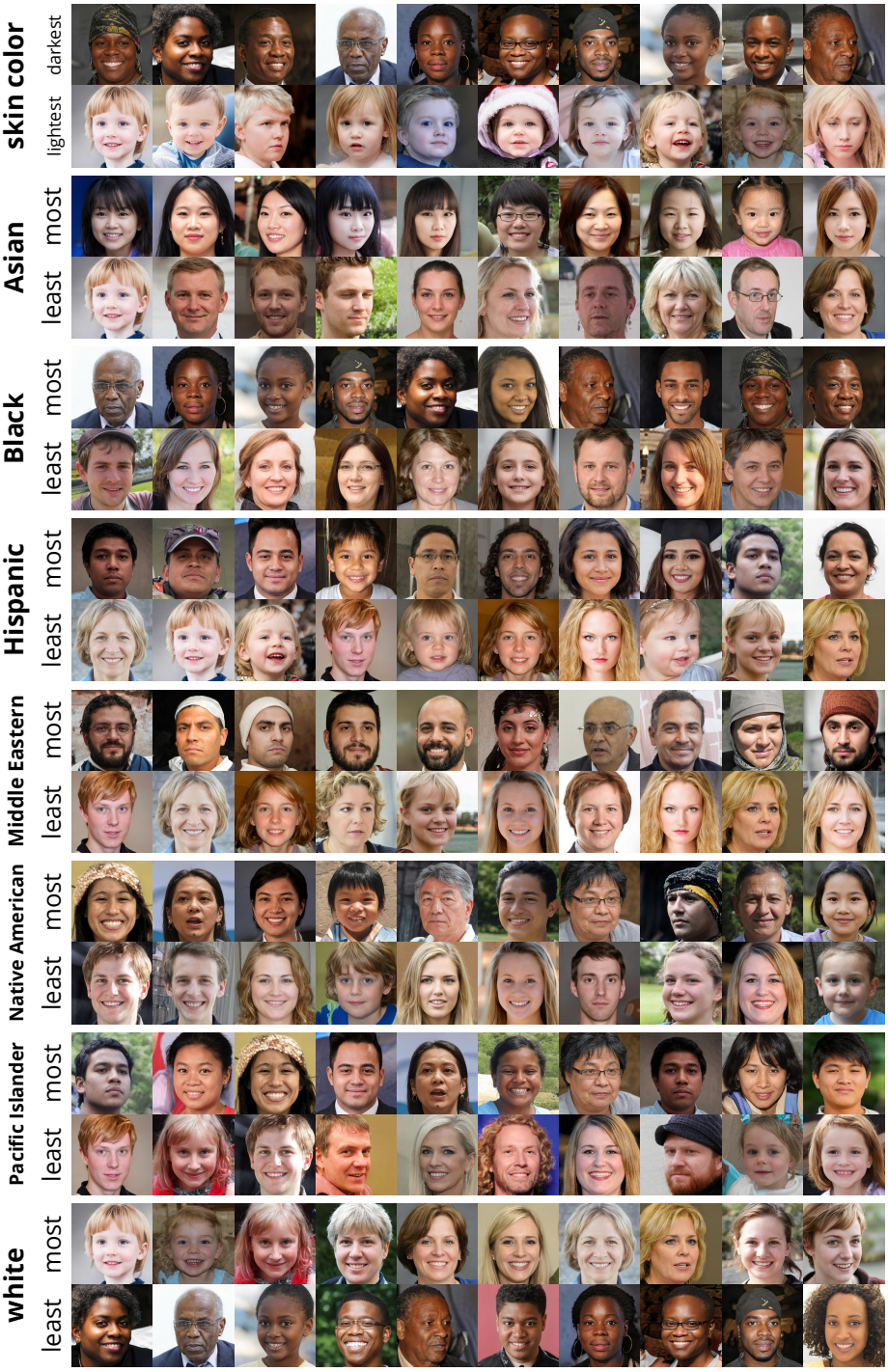

Figure 8. Ten most and ten least congruent faces (highest/lowest mean ratings) for traits *skin color*, *Asian*, *Black*, *Hispanic*, *Middle Eastern*, *Native American*, *Pacific Islander*, and *white*.

Figure 9. Ten most and ten least congruent faces (highest/lowest mean ratings) for traits *age*, *feminine/masculine*, *skinny/fat*, *alert*, and *outdoors*.

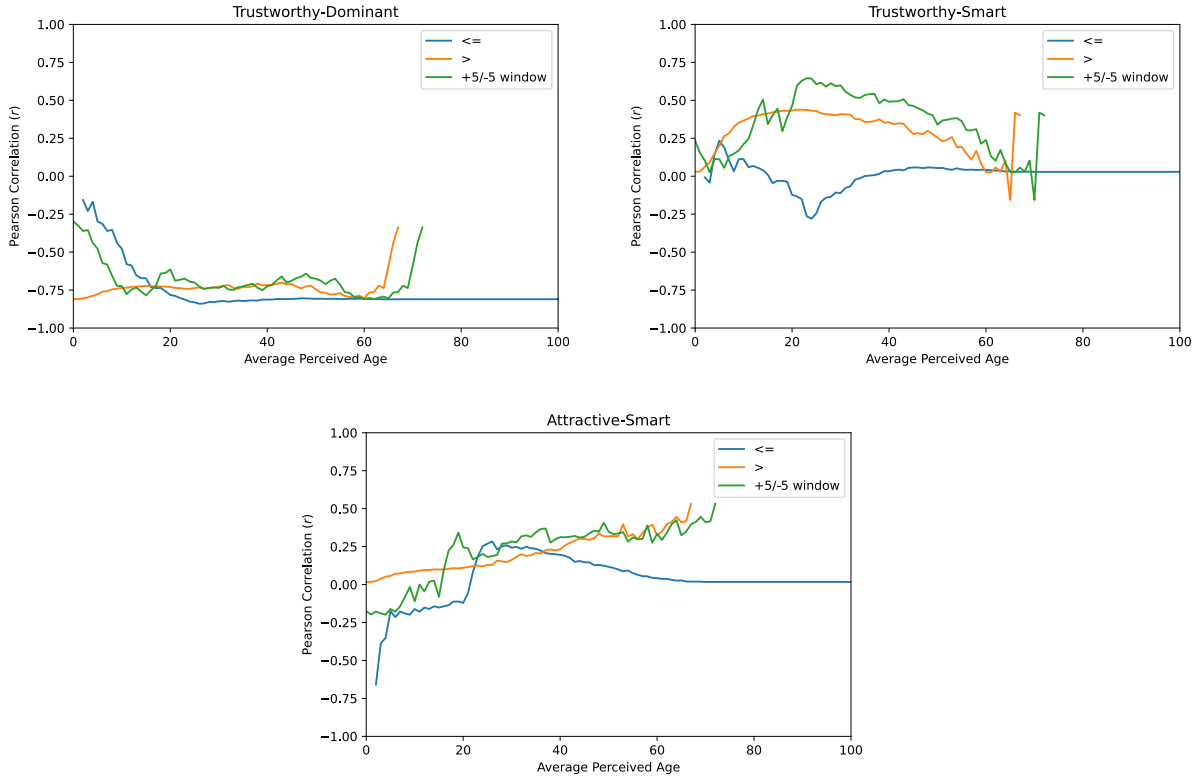

Figure 10. Correlations between average traits ratings as a function of age for three pairs of traits. The blue curves plot correlations for faces less than or equal the age threshold on the x-axis. The orange curves plot correlations for faces greater than the threshold on the x-axis. The green curves plot correlations for faces within a sliding 10-year window around the values on the x-axis. For the trait pair trustworthy-dominant (top left), the correlation is only affected at very young or very old ages. For trustworthy-smart (top right), the correlation increases up to ages around approximately 25, and then decreases again. For attractive-smart (bottom), the correlation becomes larger and more positive for older ages.

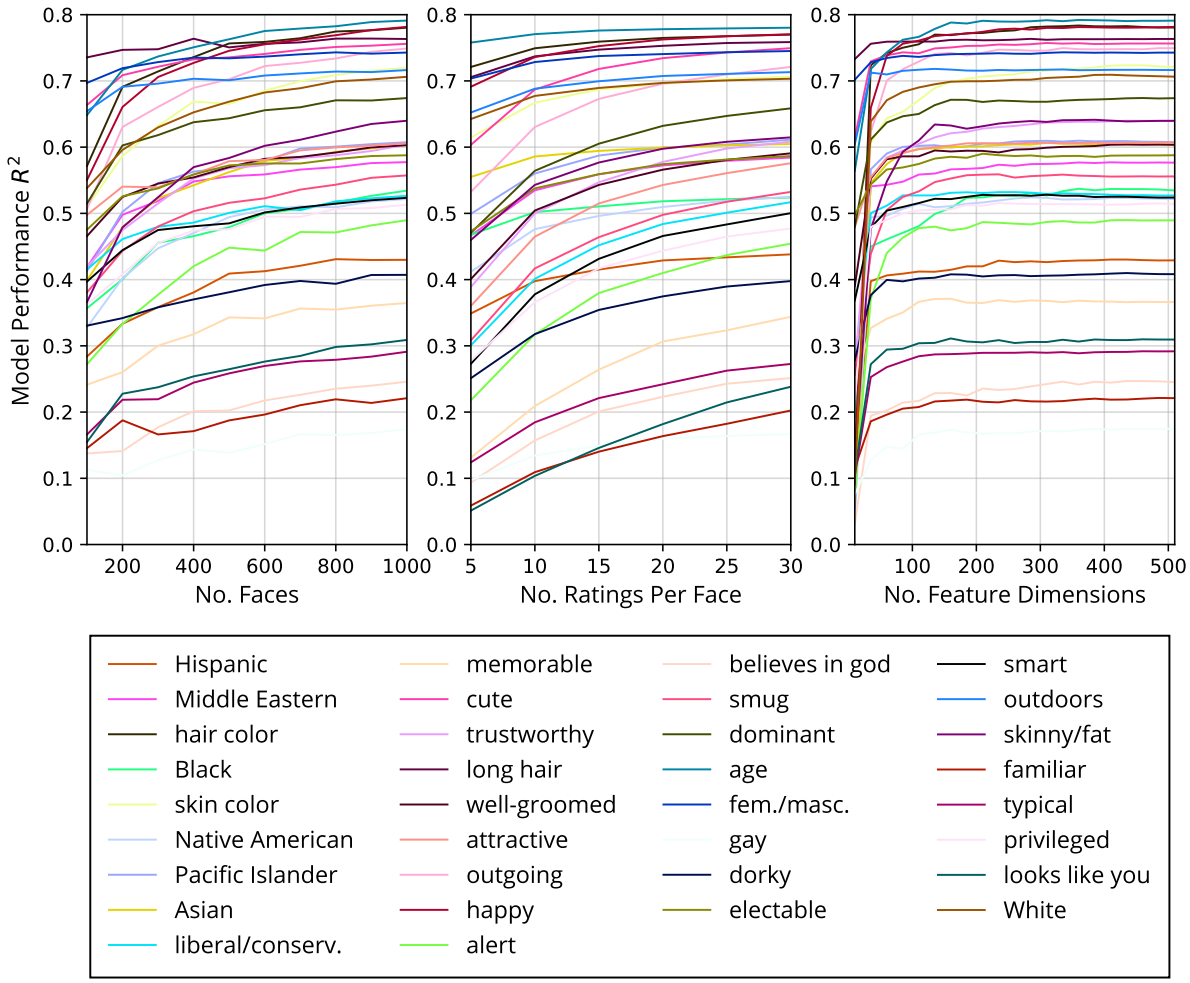

Figure 11. Model performance (R^2) for each trait as a function of the number of face examples (left), the number of participant ratings for each face example (middle), and the number of image feature dimensions (right).