Four dimensions characterize comprehensive trait judgments of faces

Chujun Lin\textsuperscript{1,}, Umit Keles\textsuperscript{1,}, Ralph Adolphs\textsuperscript{1,2}

\textsuperscript{1}Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA.

\textsuperscript{2}Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA.

*Correspondence to: clin7@caltech.edu
Abstract

People readily attribute many traits to faces: some look beautiful, some competent, some aggressive. These snap judgments have important consequences in real life, ranging from success in political elections to decisions in courtroom sentencing. Modern psychological theories argue that the hundreds of different words people use to describe others from their faces are well captured by only two or three dimensions, such as valence and dominance, a highly influential framework that has been the basis for numerous studies in social and developmental psychology, social neuroscience, and in engineering applications. However, all prior work has used only a small number of words (12 to 18) to derive underlying dimensions, limiting conclusions to date. Here we employed deep neural networks to select a comprehensive set of 100 words that are representative of the trait words people use to describe faces, and to select a set of 100 faces. In two large-scale, preregistered studies we asked participants to rate the 100 faces on the 100 words (obtaining 2,850,000 ratings from 1,710 participants), and discovered a novel set of four psychological dimensions that best explain trait judgments of faces: warmth, competence, femininity, and youth. We reproduced these four dimensions across different regions around the world, in both aggregated and individual-level data. These results provide a new and most comprehensive characterization of face judgments, and reconcile prior work on face perception with work in social cognition and personality psychology.
Main

People attribute a wide range of traits (temporally stable characteristics, see Methods) to other individuals upon viewing their faces, such as demographics (e.g., gender, age), physical appearance (e.g., baby-faced, beautiful), social evaluation (e.g., trustworthy, competent), and personality (e.g., aggressive, sociable). These trait judgments are made ubiquitously and rapidly, and are known to influence most subsequent processing, such as conscious perception and memory of the face. Although trait judgments of faces are in many cases inaccurate and reveal more about our own stereotypes than ground truth, they have major consequences for social decision-making in real life, ranging from success in job markets and social relationships to political elections and courtroom decisions.

Despite the considerable amount of work on the topic, it remains unclear how people make these rapid judgments: do they have distinct representations for each of the hundreds of possible words that describe somebody based on the face, or do they map their judgments of the face into a much lower-dimensional space? By analogy, we can perceive (and have words for) many different shades of colors, but they are all the result of a three-dimensional color space. In the case of color, the answer is easier because we know that there are only three kinds of cones in the retina; in the case of trait judgments of faces, we must infer the psychological space from behavioral data (human subjects’ ratings of faces on different trait words). Prior approaches have discovered dimensional frameworks that have largely shaped studies both within and outside the field, but those approaches used only a small number of trait words (12 to 18) that were common across studies or in use by lay people. Moreover, those words are partly redundant in meaning and may not encompass the full range of trait words that people can use to describe faces, contributing to disagreements in the literature.
Here we argue that to understand the true dimensionality of face judgments, it is essential to investigate a more comprehensively sampled set of judgments. To meet this challenge, we assembled an extensive list of trait words that people use to describe faces from multiple sources\textsuperscript{1,3,4,16,17,21,25–31,33,38,39} and applied a pre-trained neural network to derive a representative subset of 100 words (Fig. 1a-d). Similarly, we combined multiple extant face databases and applied a pre-trained neural network to derive a representative subset of 100 face images (Fig. 1e-h) [see Methods]. We verified that our 100 words were indeed representative of the trait words people spontaneously generate for the selected 100 faces (Extended Data Fig. 1a,b; Fig. 1d), and that our 100 faces were representative of the structural physiognomy of natural faces (Extended Data Fig. 1c,d,e; Fig. 1h) although we note that we only used Caucasian faces with no emotional expressions [see Methods]. We collected ratings of the 100 faces on the 100 words both sparsely online (Study 1) [750,000 ratings from 1,500 participants with repeated ratings for assessing within-subject consistency for every trait] and densely on-site (Study 2) [10,000 ratings from each of 210 participants across North America, Latvia, Peru, the Philippines, India, Kenya, and Gaza]. All experiments were preregistered on the Open Science Framework (see Methods).
Fig. 1: Sampling trait words (a-d) and face images (e-h) to generate a comprehensive set.

a. We began by assembling an extensive list of trait words spanning all important categories of trait judgments of faces. b. Each adjective was represented with a vector of 300 semantic features that describe word embeddings and text classification using a state-of-
the-art neural network that had been pre-trained to assign words to their contexts across 600 billion words. \(c\), Three filters were applied to remove words with similar meanings, unclear meaning, and infrequent usage (see Methods). \(d\), Comparing the final selected 100 traits (Extended Data Table S1) with spontaneous trait judgments of faces (Extended Data Fig. 1).

**Uniform Manifold Approximation and Projection (UMAP)**, a dimensionality reduction technique that generalizes to nonlinearities, showed that the 100 selected traits (blue dots; examples labeled in blue) were representative of the words people freely generated to describe spontaneous judgments of the 100 faces (gray dots, see Methods; non-overlapping examples labeled in gray, which were mostly momentary mental states rather than traits). \(e\), For faces we began by assembling a set of frontal, neutral, white faces from three popular face databases.

\(f\), Each face was represented with a vector of 128 facial features that are used to classify individual identities using neural networks pre-trained to identify individuals across millions of faces of all different aspects and races. \(g\), Maximum variation sampling was applied to select faces with maximum variability in facial structure in this 128-D space. \(h\), UMAP showed that the final selected 100 faces (stars) were representative of a larger set of frontal, neutral, white faces from various databases (dots) [Extended Data Fig. 1].

**Four dimensions underlie trait judgments of faces**

Study 1 applied exploratory factor analysis (EFA; preregistered) on aggregate-level ratings that participants had given to faces. We confirmed that these ratings showed sufficient variance (Extended Data Fig. 2a), within-subject consistency (assessed with Pearson’s correlations, \(M = 0.47, \text{Range} = [0.28, 0.84]\), as well as linear mixed-effect modeling [preregistered]; Fig. 2), and between-subject consensus (preregistered; all ICCs > 0.60) [Fig. 2 and Methods]. Eight traits
with low factorizability were excluded from EFA (Extended Data Fig. 2b; including them did not change the dimensions we eventually found).

We first determined the optimal number of factors to retain in EFA using five widely recommended methods\textsuperscript{50,51} (see Methods), as solutions are considered most reliable when multiple methods agree. Four methods—Horn’s parallel analysis, Cattell’s scree test, optimal coordinates, and empirical BIC—all indicated that the optimal number of factors to retain was four (Extended Data Fig. 3a).

EFA was thus applied to extract four factors using the minimal residual method, and the solutions were rotated with oblimin for interpretability. The four factors each explained 31%, 31%, 11%, and 12% of the common variance in the data (85% in total; 87% in total if five factors were extracted) and were weakly correlated ($r_{13} = -0.33$, $r_{14} = -0.23$, $r_{23} = 0.21$, $r_{24} = 0.33$ [$p < 0.05$]; $r_{12} = -0.15$, $r_{34} = 0.12$ [$p > 0.05$]). None of the factors were biased by words with particularly low or high within-subject consistency or between-subject consensus; and the trait words occupied the four-dimensional space fairly homogeneously (Fig. 2). We interpreted these four factors as describing judgments of warmth, competence, femininity, and youth (Fig. 2; see Extended Data Fig. 4a for factor loadings), labels that were validated both with an independent set of participant ratings and with word embedding metrics [see Methods].
Fig. 2: Reliability and dimensionality of comprehensive trait judgments of faces.

Upper right scatterplot: within-subject consistency as assessed with linear mixed-effect modeling (y-axis, regression coefficients) plotted against between-subject consensus as assessed with intraclass correlation coefficients (x-axis) of the 100 traits. The color scale indicates the product between the x- and y-values. Four histograms in diagonal: each plots the distribution of the factor loadings across all words in EFA, on each of the four dimensions (see Extended Data Fig. 4a for factor loadings; color code as in upper right scatterplot). Six scatterplots in lower left: each plots the factor loading of all words in EFA against two of the four dimensions (dots). Labels are shown for a small subset of datapoints (blue dots) due to limited space (see Extended Data Fig. 4b for full labels).
Comparison with existing dimensional frameworks

Prior work\textsuperscript{4,5,17,35,37} suggests that the various words people use to describe faces can be represented by two or three dimensions (e.g., valence and dominance\textsuperscript{4}). Our findings support the general idea of a low-dimensional space, but revealed four dimensions that differ from those previously proposed. This discrepancy was not explained by methodological differences: we reanalyzed our data using principal components analysis (PCA), a method used in prior work\textsuperscript{4,17,37}, and reproduced the same four dimensions as reported here (Extended Data Fig. 5a).

Instead, our four-dimensional space did not appear in previous studies because of limited sampling of traits in prior work: we interrogated two subsets of our data which each consisted of 13 traits that corresponded to those used in the discovery of the two most popular prior dimensional frameworks (2D and 3D frameworks)\textsuperscript{4,37}. Our four-dimensional space was not evident when analyses were restricted to these two small subsets of traits; instead, we reproduced the prior 2D and 3D frameworks, respectively (Extended Data Table 2a-b).

We next used our reproduction of the popular prior 2D-framework for more detailed comparisons with our four dimensions. Replicating prior findings\textsuperscript{4}, we found that judgments of faces on the traits sociable, trustworthy, responsible, and weird were represented by the 2D-framework’s valence dimension (absolute $rs = 0.94, 0.88, 0.86, 0.85$ between factor scores on the valence factor we derived, and ratings for these words across the 100 faces), but less well represented by our own warmth dimension (absolute $rs = 0.47, 0.67, 0.43, 0.23$; Extended Data Fig. 4a); the valence and warmth dimensions were moderately correlated (absolute $r = 0.41$ between factor scores). Similarly, as previously found\textsuperscript{4}, judgments on aggressive and submissive were represented by the 2D-framework’s dominance dimension (absolute $rs = 0.94, 0.95$), but...
not by our own competence dimension (absolute $rs = 0.15, 0.14, ps > 0.05$; Extended Data Fig. 4a); the two dimensions were not significantly correlated ($r = 0.01$) [see Methods].

We directly compared how well different frameworks characterized trait judgments of faces. Using linear combinations of traits with the highest loadings on each dimension as regressors (two for each dimension, due to only two traits loading on one of the dimensions in the 3D framework), we found that our four-dimensional framework better explained the variance for 82% of the trait judgments (that were not part of the linear combinations) than did any of the existing frameworks (Extended Data Fig. 5b; mean adjusted R-squared across all predictions was 0.81 for our framework, 0.72 for the 3D framework$^3$, and 0.72 for the 2D framework$^4$).

A final question of interest was how our dimensions that characterize face judgments might relate to dimensions that characterize personality, such as the Big Five$^1$. We asked an independent sample of 343 participants to rate themselves on a subset of 68 of our trait words that correspond to personality traits (no faces were shown; the task was self-report on the words; see Methods). As expected, the five Big Five personality dimensions emerged in this dataset (Extended Data Fig. 5c; using the same EFA method as for the face-trait ratings). Although there was some overlap in the way that our four face-judgment dimensions and the personality dimensions captured the variance in ratings evoked by these 68 words (Extended Data Fig. 5d), further analysis showed distinct hierarchical structures (Extended Data Fig. 5e,f). We conclude that trait judgments of unfamiliar faces (the focus of the present study), and self-reports of personality (as in the Big Five), are best characterized by two distinct dimensional spaces.

**Robustness and validity of the four dimensions**
We quantified the robustness of our results across different numbers of trait words or participants. We removed trait words one by one (based on their rank-ordered meaning similarity and unclarity) and reperformed EFA as before. Our four dimensions were highly robust (with 75 or more of our 100 words, rs between factors were > 0.95; with 65 or more words, rs were > 0.7; Extended Data Table S2.c). Similarly, we randomly removed participants one by one (50 randomizations each) and used the new aggregated ratings for EFA to show that the four dimensions were robust to participant sample size (Tucker indices of factor congruence > 0.95 for all four factors between the full dataset and all sub-datasets with no fewer than 19 participants per trait). Finally, we extracted the smallest subset of specific trait words that still yield the original four-dimensional space, a set of 18 words that could be used most efficiently in future studies when collecting ratings for a larger set of traits is not feasible (Extended Data Table S2.d).

To confirm our four dimensions and rule out the possibility of more complex hierarchical structure, we adopted an approach with minimal assumptions, using artificial neural networks (ANN) and cross-validation to compare different factor structures (see Methods). Autoencoder ANNs that differed in the number of neurons and hidden layers were constructed so as to model the factor structures that we wished to confirm (the existing 2D and 3D, our 4D, and hierarchical versions thereof). ANNs trained on half of the data and tested on the held-out other half confirmed a four-dimensional representation (explained variance obtained with linear activation functions = 75% on the test data [SD = 0.6%], Extended Data Fig. 6). This performance is comparable to PCA, and the improvement in model performance became trivial beyond four dimensions (explained variance on the test data from 1 to 4 neurons in the hidden layer increased by 18%, 5%, and 5%; but by less than 1% beyond 4 neurons). The four-
dimensional representation learned by the ANN reproduced our four dimensions (mean $r_s =$ 0.98, 0.92, 0.91, 0.94 [SDs = 0.01, 0.05, 0.02, 0.05] between our original factor loadings from EFA and the ANN’s decoder layer weights using varimax rotation). Adding hierarchical structure (additional layers to the ANN) did not explain more variance (Extended Data Fig.6e,f).

Generalizability across different countries and regions

Prior work has reported both common and discrepant dimensions in different cultures4,17,24,35,37. To test the generalizability of our findings, we conducted a second preregistered study to collect data across seven different regions of the world. We first analyzed the aggregate-level ratings for each sample (preregistered; we confirmed these ratings had satisfactory consistency and consensus, see Methods).

We began by asking whether the seven samples shared a similar correlation structure (the Pearson correlation matrix across trait ratings) with the sample in Study 1, using representational similarity analysis33 [RSA; Fisher z-transformation was applied before computing the correlation between correlation matrices]. Highly similar correlation structures were found across samples (RSAs with Study 1 = 0.96, 0.92, 0.85, 0.85, 0.75, 0.83, 0.86 for North America, Latvia, Peru, Philippines, India, Kenya, and Gaza, respectively). These high RSAs strongly suggest that a similar psychological space underlies face judgments across different samples.

Parallel analysis, optimal coordinates, and empirical BIC all showed that a four-dimensional space was most common across samples (in 5 of 7 samples: North America, Latvia, Peru, the Philippines, India) [Fig. 3a and Extended Data Fig. 3b-h]. We therefore applied EFA to extract four factors from each sample. Results showed that the warmth, competence, femininity, and
youth dimensions emerged in multiple samples (interpreted based on factor loadings, see
Extended Data Fig. 7). We further computed Tucker indices of factor congruence (the cosine
distance between pairs of factor loadings), which confirmed that the four-dimensional space was
largely reproduced across samples (Fig. 3b), but, as expected, reproducibility was attenuated by
the data quality available (as assessed by within-subject consistency, Fig. 3c)

Fig. 3: Dimensionality of comprehensive trait judgments of faces across different samples.
a, Eigenvalue decomposition. Dots plot the eigenvalues of the first 10 factors across seven samples, indicated by different colors. b, Tucker indices of factor congruence. Columns indicate the four dimensions found in Study 1: warmth (W), competence (C), femininity (F), and youth (Y). Rows indicate the dimensions found in the samples from North America (NA), Latvia (LV), Peru (PE), the Philippines (PH), Kenya (KE), India (IN), and Gaza (GZ). Numbers report the Tucker indices (with orthogonal Procrustes rotation). The color scale shows the sign and strength of the indices. c, Individual within-subject consistency by sample (assessed with Pearson’s correlations). Every participant in Study 2 had rated a subset of 20 traits twice for all faces to provide an assessment of individual data quality in terms of within-subject consistency.

Reproducibility across individual participants

So far, we have reproduced the four-dimensional space across samples, but we have not ruled out the possibility that this space might be an artifact of aggregating data across participants. Could the same four-dimensional space be reproduced in a single participant? This important question has been difficult to address since one needs to have complete data per participant. We met this challenge by collecting ratings on all traits for all faces from every participant in Study 2 (requiring approximately 10 hours of testing per participant; see Methods).

We first performed RSA to investigate whether single participants (n = 86 who had complete datasets for all traits after data exclusion; see Methods) shared the correlation structure of our Study 1 sample. RSAs varied considerably across participants (range = [0.14, 0.85], M = 0.56, SD = 0.16) and, as expected, were attenuated by data quality as assessed by within-subject consistency (Fig. 4a,b).
We next analyzed the dimensionality of each individual dataset. Parallel analysis (preregistered) showed that a four-dimensional space was most common (Fig. 4c) but, again, attenuated by data quality (four-dimensional spaces were found for data with higher within-subject consistency than data that produced other-dimensional spaces [unpaired t-test $t(34.57) = 3.29$, $p = 0.001$]). We therefore applied EFA to extract four factors from each participant’s dataset and computed their factor congruence with the data from Study 1. We found that our four dimensions were reproduced in some participants (see examples of factor loading matrices in Extended Data Fig. 8a, and Tucker indices for all participants in Extended Data Fig. 8b), but also found a considerable amount of individual differences, in line with prior research.

**Fig. 4: Dimensionality of comprehensive trait judgments of faces in individual data.**

* a, Representational similarity between aggregated data from Study 1 and individual-level data from Study 2 for individuals who had complete data after exclusion ($n = 86$, see Methods).
Colors indicate different samples (as in Fig. 3). Boxplots indicate the minima (bottommost line), first quartiles (box bottom), medians (line in box), third quartiles (box top), and maxima (topmost line) of RSAs. b, Correlation between within-subject consistency and RSA (R = 0.66, p < 0.001). Each point plots an individual’s within-subject consistency (x-axis) and that individual’s RSA with the aggregated data in Study 1 (y-axis). c, Distribution of dimensionality (from parallel analysis) across 86 individual-level datasets.

Discussion

Across two large-scale, pre-registered studies we found that comprehensive trait judgments of faces are best described by a four-dimensional space, with dimensions interpreted as warmth, competence, femininity and youth (Fig. 2, Extended Data Fig. 4). This finding was largely reproduced across samples from different regions, even using different languages (Spanish in Peru) [Fig. 3, Extended Data Fig. 7], as well as in individual participants (although this was more difficult to assess, due to data quality) [Fig. 4 and Extended Data Fig. 8]. We showed that our divergence from prior work was not due simply to methodological differences, but to the prior lack of comprehensively and representatively sampled trait words (Fig. 1, Extended Data Figs. 1, 5a,b, 6, and Table S2a,b).

These findings help to reconcile studies of face perception with the broader social cognition literature, which has long theorized that warmth and competence are two universal dimensions of social cognition. The other two dimensions we found, femininity and youth, are likely linked to overgeneralization and corroborate recent neuroimaging findings on social categorization from face perception. With an inclusion of a representative list of personality words, our results also provide new insights into the distinctions between the dimensions people use to describe
people from faces, and from self-reported personality, such as the Big Five (Extended Data Fig. 5d,e,f; see also Methods).

Despite the predominance of our four-dimensional space, we also found notable variation across samples and individuals (Figs. 3, 4). Since the sources of this variation are unknown and may largely reflect measurement error (Fig. 3c, 4b), we refrain from drawing any specific conclusions about cultural differences, for which larger-scale studies focusing on cultural effects will be needed. Similarly, conclusions about individual differences will require future studies that collect much denser, and likely longitudinal, data in individual participants. Face stimuli incorporating various races or emotional expressions will likely modify the dimensions of face judgments, as will viewing angle, background, and other context effects. Our findings provide the most comprehensive characterization of trait judgments from the physiognomy of faces alone, yielding candidate mental dimensions to investigate with respect to all these further variables, as well as in neuroimaging studies of face judgments.
Methods

Sampling of trait words

Here we follow the definition of a biological trait as being a temporally stable characteristic. Traits in our study include personality traits as well as other temporally stable characteristics that people spontaneously infer from faces, such as age, gender, race, socioeconomic status, and social evaluative qualities (Extended Data Fig. 1a, e.g., “young”, “female”, “white”, “educated”, “trustworthy”). By contrast, we excluded state attributions, such as “smiling” or “thinking” (words that can describe both trait and state variables were not excluded, e.g., we included “happy,” but disambiguated its usage as a trait in our instructions to participants, e.g., “A person who is usually cheerful;” Extended Data Table S1).

Our goal was to representatively sample a comprehensive list of trait words that are used to describe people from their faces. We derived a final set of 100 traits (Extended Data Table 1) through a series of combinations and filters (detailed below; also in our preregistration at https://osf.io/6p542). These 100 traits were further verified to be representative of words that people freely generate to describe trait judgments of our face stimuli (Fig. 1d and Extended Data Fig. 1a,b).

To derive the final set of trait words, we first gathered an inclusive list of 482 adjectives and 6 nouns that included all major categories of trait judgments of faces: demographic characteristics, physical appearance, social evaluative qualities, personality, and emotional traits, from multiple sources1,3,4,16,17,21,25–31,33,38,39. Many of the 482 adjectives were synonyms or antonyms. To avoid redundancy while conserving semantic variability, we sampled these adjectives according to three criteria: their semantic similarity (detailed below), clarity in meaning (from an independent
set of 29 MTurk participants), and frequency in usage (detailed below). For those words with similar meanings, clarity was the second selection criterion (the one with the highest clarity was retained). For those with similar meanings and clarity, usage frequency was the third selection criterion (the one with the highest usage frequency was retained).

To quantify the semantic similarity between these 482 adjectives, we represented each of them as a vector of 300 computationally extracted semantic features that describe word embeddings and text classification using a neural network provided within the FastText library \(^{40}\); this neural network had been trained on Common Crawl data of 600 billion words to predict the identity of a word given a context. We then applied hierarchical agglomerative clustering (HAC) on the word vectors based on their cosine distances to visualize their semantic similarities. To quantify clarity of meaning, we obtained ratings of clarity from an independent set of participants tested via MTurk (N = 31, 17 males, Age (M = 36, SD = 10)). To quantify usage frequency, we obtained the average monthly Google search frequency for the bigram of each adjective (i.e., the adjective together with the word “person” added after it) using the keyword research tool Keywords Everywhere (https://keywordseverywhere.com/).

The 94 adjectives representatively sampled using the above procedures and the additional 6 nouns consisted of our final set of 100 trait words. To verify the representativeness of these 100 trait words, we compared the distributions of our selected words and of 973 words human subjects freely generated to describe their spontaneous impressions of the same faces (see Extended Data Fig. 1a and Methods below), using the 300 computationally extracted semantic dimensions. We visualized these distributions using Uniform Manifold Approximation and Projection (UMAP\(^{41}\)) as shown in Fig. 1d.
To ensure that the dimensionality of the meanings of the words that we used was not limiting the dimensionality of the four factors we discovered in our study, we derived a similarity matrix among our 100 words using the FastText vector of their meanings in the specific one-sentence definitions we gave to participants in the experiments (Extended Data Table S1; basic stop-words such as “a”, “about”, “by”, “can”, “often”, “others” were removed from the one-sentence definitions for the computation of vector representations), and then conducted factor analysis on the similarity matrix. Parallel analysis, Optimal Coordinate Index, and Kaiser’s Rule all suggested 13 dimensions; Velicer’s MAP suggested 14 dimensions, and empirical BIC suggested 5 dimensions (empirical BIC penalizes model complexity). We used EFA to extract 5 and 13 factors using the same method as for the trait ratings (13 factors explained the same common variance as 14 factors, 70%; 5 factors explained 60%; factors were extracted with minimal residual method and rotated with oblimin to allow for potential factor correlations). None of the dimensions obtained bore resemblance to our four reported dimensions, arguing that the mere semantic similarity structure of our 100 trait words was not a constraint in deriving the four factors that we report.

### Sampling of face images

Our goal was to derive a representative set of neutral, frontal, white faces of high quality (clear, direct gaze, frontal, unoccluded, and high resolution) that are diverse in facial structure. We aimed to maximize variability in facial structure while controlling for factors such as race, expression, viewing angle, gaze, and background, which our present project did not intend to investigate. We first combined 909 high-resolution photographs of male and female faces from three publicly available face databases: the Oslo Face Database44, the Chicago Face Database43, and the Face Research Lab London Set42. We then excluded faces that were not front-facing, not
with direct-gaze, with glasses or other adornments obscuring the face. We further restricted
ourselves to images of Caucasian adults and neutral expression. This yielded a set of 426 faces
from the three databases.

To reduce the size of the stimulus set while conserving variability in facial structure, we sampled
from the 426 faces using maximum variation sampling. For each image, the face region was first
detected and cropped using the dlib library, and then represented with a vector of 128
computationally extracted facial features for face recognition, using a neural network provided
within the dlib library that had been trained to identify individuals across millions of faces of all
different aspects and races with very high accuracy. Next, we sampled 50 female faces and 50
male faces that respectively maximized the sum of the Euclidean distances between their face
vectors. Specifically, a face image was first randomly selected from the female or male sampling
set, and then other images of the same gender were selected so that each new selected image had
the farthest Euclidean distance from the previously selected images. We repeated this procedure
with 10,000 different initializations and selected the sample with the maximum sum of Euclidean
distances. We repeated the whole sampling procedure 50 times to ensure convergence of the
final sample. All 100 images in the final sample were high-resolution color images, with the eyes
at the same height across images, had a uniform grey background, and were cropped to a
standard size. See preregistration at https://osf.io/6p542.

To verify the representativeness of our selected 100 face images, we again performed UMAP
analysis to compare the distribution of our selected faces with a) $N = 632$ neutral, frontal, white
faces from a broader set of databases (Fig. 1h) and b) $N = 5376$ white faces “in the wild” that varied in angle, gaze, facial expression, lighting, and backgrounds (Extended Data Fig. 1c),
using the 128 computationally extracted facial identity dimensions as well as 30 traditional facial metric dimensions (Extended Data Fig. 1d-e).

3 Freely generated trait words

To verify that our selected 100 trait words were indeed representative of the trait judgments people spontaneously make from faces, we collected an independent dataset from participants who freely generated words about the person that came to mind upon viewing the face. As preregistered, 30 participants were recruited via MTurk (see preregistration at http://bit.ly/osfpre4); different from the preregistration, we decided to not only include Caucasian participants but included participants of any race (27 participants were white, 3 participants were black).

Participants viewed the 100 face images one by one, each for 1 second, and typed in the words (preferably single-word adjectives) that came to mind about the person whose face they just saw. Participants could type in as many as ten words and were encouraged to type in at least four words (the number of words entered per trial—words entered by a participant for a face—ranged from 0 words [for 8 trials] to 10 words [for 190 trials] with mean = 5 words). There was no time limit; participants clicked “confirm” to move on to the next trial when they finished entering all the words they wanted to enter for the current trial. All data can be accessed at https://osf.io/4mvyt/.

19 Study 1 Participants

All studies in this report were approved by the Institutional Review Board of the California Institute of Technology and informed consent was obtained from all participants. We predetermined our sample size for Study 1 based on a recent study that investigated the point of
stability for trait judgments of faces across 24 traits, a stable average rating could be obtained in a sample of 18 to 42 participants (ratings were elicited using a 7-point rating scale, the acceptable corridor of stability was +/- 0.5, and the confidence level was 95%). Based on these findings, we preregistered our sample size for Study 1 to be 60 participants for each trait (at https://osf.io/6p542).

Participants were recruited via MTurk (N = 1,500 (800 males), Age (M = 38 years, SD = 11), the median of educational attainment was “some post-high-school, no bachelor's degree”). All participants were required to be native English speakers located in the U.S. of 18 years old or older, with normal or corrected-to-normal vision, with an educational attainment of high school or above, and with a good MTurk participation history (approval rating ≥ 95%).

We also collected data about whether our participants were currently being treated for psychiatric or neurological illness. The majority of our participants (79.7%) were not currently being treated for any psychiatric or neurological illness. All dimensional analyses that are reported in the main text on the full sample were repeated also on those 79.7% of participants and the results corroborated all findings from the full dataset: Tucker indices of factor congruence for the four dimensions = 1.00, 1.00, 0.99, 0.99.

**Study 1 Procedures**

All experiments in Study 1 were completed online via MTurk. Considering the large amount of time it would take for a participant to complete ratings for all 100 traits and 100 faces, we divided the experiment into 25 modules: the 100 traits were randomly shuffled once and divided into 25 modules, each consisting of 4 traits. Each participant completed one module.
To encourage participants to use the full range of the rating scale, we briefly showed all faces (in five sets of arrays of 20 each) at the beginning of a module, so that participants had a sense of the range of the faces they were going to rate. In each module, participants rated all faces on each of the four traits (in random order) in the first four blocks; in the last (fifth) block they rerated all faces on the trait they were assigned in the first block again, thus providing sparse within-subject consistency data.

At the beginning of each block, participants were instructed on the trait they were asked to evaluate and were provided with a one-sentence definition of the trait (Extended Data Table 1). Participants viewed the faces one by one in random order (each for 1 second) and rated each face on a trait using a 7-point rating scale (by pressing the number keys on the computer keyboard). Participants could enter their ratings as soon as the face appeared or within four seconds after the face disappeared. The orientation of the rating scale in each block was randomized across participants. At the end of the experiment, participants completed a brief questionnaire on demographic information. See preregistration at https://osf.io/6p542.

Measures of reliability in Study 1

Data were first processed following three preregistered exclusion criteria (see preregistration at https://osf.io/6p542): of the full sample with a registered size of $N = 1,500$ participants and $L = 750,000$ ratings, $n = 48$ participants and $l = 27,491$ ratings were excluded from further analysis. Each of the 100 traits was rated twice for all faces by nonoverlapping subsets of participants (ca. $n = 15$ per trait). As preregistered, we applied linear mixed-effect modeling to assess within-subject consistency, which adjusted for non-independence in repeated individual ratings by incorporating both fixed effects (that were constant across participants) and random effects (that
varied across participants). Ratings from every participant for every face collected at the second 
time were regressed on those collected at the first time (ca. \( l = 1,445 \) pairs of ratings per trait) 
while controlling for the random effect of participants.

As preregistered, we assessed the between-subject consensus for each trait with intraclass 
correlation coefficients (ICC(2,k)), using ratings of every face by every participant (ca. \( n = 58 \) 
participants and \( l = 5,780 \) ratings per trait). A high intraclass correlation coefficient indicates that 
the total variance in the ratings is mainly explained by the variance across faces instead of 
participants. We observed excellent between-subject consensus (ICCs greater than 0.75) for 93 
of the 100 traits, and good between-subject consensus for the remaining 7 traits (ICCs greater 
than 0.60) [see Fig. 2].

**Determination of the optimal number of factors**

As recommended\(^{50,51,58,59}\), five methods were included to determine the optimal number of 
factors to retain in EFA. No single method was regarded as the best method for determining the 
number of factors; solutions are considered most reliable when multiple methods agree. Parallel 
analysis retains factors that are not simply due to chance by comparing the eigenvalues of the 
observed data matrix with those of multiple randomly generated data matrices that match the 
sample size of the observed data matrix. Prior studies showed that parallel analysis produces 
accurate estimations of the number of factors consistently across different conditions (e.g., the 
distribution properties of the data) \(^{58,59}\). Cattell’s scree test retains factors to the left of the point 
from which the plotted ordered eigenvalues could be approximated with a straight line (i.e., 
retains factors “above the elbow”). The optimal coordinates index provides a non-graphical 
solution to Cattell’s scree test based on linear extrapolation. Empirical Bayesian information
criterion (eBIC) retains factors that minimize the overall discrepancy between the population’s and the model’s predicted covariance matrices while penalizing model complexity. Velicer’s minimum average partial (MAP) test is “most appropriate when component analysis is employed as an alternative to, or a first-stage solution for, factor analysis”\textsuperscript{60}. It is also included in our present study due to its popularity. MAP retains components by partialing out those that resulted in the lowest average squared partial correlation.

Comparing different dimensional frameworks

Using a subset of 13 traits in our set that are identical, or have very similar meanings, to those used in prior work\textsuperscript{4}, we successfully reproduced the previously reported two dimensions of valence and dominance (Extended Data Table S2.a). To compare the valence and dominance dimensions to the first two of our four dimensions reported here (“warmth” and “competence”) we assessed the factor scores of every face on each of the above dimensions (using the R function “factor.scores” with method “tenBerge”) and then correlated the faces’ scores between different dimensions or with their ratings on different traits.

Determination of the labeling of factors

Dimensionality reduction methods do not provide labels for the factors discovered, which must instead be interpreted by the investigators. Two influential theoretical frameworks from prior research\textsuperscript{4,15,37} have provided candidates for the labeling of our first two dimensions: the warmth-and-competence framework from the stereotype content model of the social cognition literature\textsuperscript{15} and the valence-and-dominance framework from the social attribution model of face impression literature\textsuperscript{4}. We empirically tested which labels best describe the first two dimensions found in our own study using two different approaches.
First, we collected data from an independent set of participants on the relevance of our traits to each of the possible labels (N = 30 per label). Specifically, participants were shown the 100 trait words one by one and were asked one of the following questions: “how warm or cold would the person described by this word be?”; “how pleasant or unpleasant would the person described by this word be?”; “how competent or incompetent would the person described by this word be?”; “how dominant or submissive would the person described by this word be?” These data can be accessed at https://osf.io/4myvt/. We averaged these ratings across participants per trait for each label and computed the correlation between (1) the ratings for all traits on a label from this task, and (2) the trait loadings on that dimension from our face-rating data (obtained from EFA as described in the main text and below). Results showed that “warmth” was a slightly more relevant label for our first dimension than “valence” (r$\text{S} = -0.773$ for “warmth” and -0.748 for “valence”), and “competence” was a more relevant label for our second dimension than “dominance” (r$\text{S} = 0.757$ for “competence” and 0.566 for “dominance”).

Second, we capitalized on recent advances in natural language processing, which provides a more objective assessment of semantic similarity. Using the pre-trained neural network FastText, which characterizes the similarity between words based on their co-occurrence in the Common Crawl data of 600 billion words, we computed the similarity between our trait words and the labels “valence”, “warmth”, “dominance”, and “competence”. We then again computed the correlation between (1) trait similarities to a label obtained from this text analysis, and (2) the traits’ loadings on that dimension (obtained from EFA as described in the main text and below). Corroborating what we found based on MTurk ratings, results again showed that “warmth” was a more relevant label for our first dimension than “valence” (r = -0.335 for “warmth” [t(90) = -3.37, p = 0.001] and r =0.002 for “valence” [p = 0.985]), and “competence” was a more relevant
label for our second dimension than “dominance” \(r = 0.329\) for “competence” \([t(90) = 3.31, p = 0.001]\) and \(r = -0.156\) for “dominance” \([p = 0.137]\)).

We note that our third and fourth dimensions describe stereotypes related to gender (femininity-stereotypes) and age (youth-stereotypes) commonly reported in the literature\(^{15}\). In fact, essentially all trait judgments based on faces, and therefore all of our dimensions, are a reflection of people’s stereotypes of some sort, since in our study nothing else is known about the people whose faces are used as stimuli, and therefore no ground truth is provided. We therefore omitted “-stereotypes” in our labeling of all dimensions, since it implicitly applies to all of them.

**Confirmatory analyses with artificial neural networks and cross-validation**

To compare different theoretical models and test potential nonlinearity in our data, we employed an artificial neural network approach, in particular, autoencoders\(^{61}\), with cross-validation. The aim of an autoencoder model is to learn a lower-dimensional representation of the data. We constructed different autoencoders based on the different models we wished to test (the existing 2D and 3D frameworks\(^4,37\), our 4D framework). We trained these autoencoders on half of the data (for each trait, 50% of the individuals were randomly selected and their ratings were used to compute new aggregated ratings per face per trait) and tested them on the held-out other half of the data. We used the Adam optimization algorithm\(^{62}\) and mean squared error loss function with a batch size of 32 and 1500 epochs to train the neural networks (the loss converged after 1000 epochs in all our models). We repeated this process for 50 iterations and compared the performance of different models. For completeness, both linear and nonlinear activation functions were explored for model fitting (linear, tanh, sigmoid, rectified linear activation unit,
L1-norm regularization; see details in Extended Data Fig. 6); a simple linear activation function ended up with the best results.

Existing frameworks suggest that all face-impression dimensions are of the same order (i.e., no dimension is a higher- or lower-order dimension of the others), but that the number of dimensions varies. Therefore, we first constructed different autoencoder models with only one hidden layer that varied in the number of neurons in this hidden layer, corresponding to the number of underlying dimensions (from 1 to 10; Extended Data Fig. 6a). The input layer and output layer were the same for all models, where each face was represented by a vector of ratings across the 92 traits and each trait corresponded to a neuron. All layers were densely connected.

We trained these different models and compared their performance (assessed with the explained variance on the held-out test data) (Extended Data Fig. 6c-d).

By contrast to models from the face impression literature, theories in personality psychology have suggested that there might be higher-order dimensions to the structure of personality. Although our study was not about personality in the usual sense, but about first impressions of faces, we wanted to test the possibility of hierarchical structure in the factors that explain our data. Therefore, we added one hidden encoder layer with various numbers of neurons (from 1 to 10) before the middle hidden layer (also with various numbers of neurons from 1 to 10); since autoencoder models are by definition symmetric, these hierarchical latent structures were mirrored in the decoder layers (Extended Data Fig. 6b). We trained these different models and compared their performance (Extended Data Fig. 6c-f).

Study 2 Participants
The study was approved by the Institutional Review Board of the California Institute of Technology and informed consent was obtained from all participants. We preregistered to recruit participants through Digital Divide Data, a social enterprise that delivers research services, in seven countries/regions of the world: North America (U.S. and Canada), Latvia, Peru, the Philippines, India, Kenya, and Gaza. All participants were required to be between 18-40 years old, proficient in English (except participants in Peru, where everything was translated to Spanish), have been educated at least through high school, have been trained in basic computer skills, and have never visited or lived in Western-culture countries (except participants in North America and Latvia). In addition, we aimed to have a roughly equal sex ratio of participants in all locations.

The sample size for each location was predetermined to be 30 participants. This sample size was determined based on two criteria: first, the sample size should be large enough to ensure stable average trait ratings (for a corridor of stability of +/- 1.00 and a level of confidence of 95%, the point of stability ranged from 5 to 11 participants across 24 traits); second, the sample size should be feasible to accrue at all seven locations given the requirements mentioned above and the availability of participants for paying multiple visits to complete all our experiment sessions over a 10-day period. See preregistration at [http://bit.ly/osfpre2](http://bit.ly/osfpre2). As planned, 30 individuals (15 females and 15 males) in each of the seven locations participated in our study (Age (M = 26, SD = 4) for North America; Age (M = 28, SD = 5) for Latvia; Age (M = 22, SD = 3) for Peru; Age (M = 25, SD = 4) for the Philippines; Age (M = 27, SD = 6) for India; Age (M = 24, SD = 2) for Kenya; and Age (M = 26, SD = 5) for Gaza).

**Study 2 Procedures**
All experiments were completed onsite in the Digital Divide Data local offices. Participants in North America, Latvia, the Philippines, India, Kenya, and Gaza completed all experiments in English. Participants in Peru completed all experiments in Spanish. An exact translation of the experiment instructions, trait words, and definitions of the traits from English to Spanish was provided by the Peru office of Digital Divide Data. Both the English and Spanish versions of those materials can be accessed at our preregistration (https://osf.io/qxgmw).

Eighty of the 100 trait words were used in Study 2—twenty words were excluded for their low correlations with other traits as found in Study 1 (sarcastic, white, thrifty, shallow, homosexual, nosy, conservative, and reserved), their ambiguity or similarity in meaning as found in feedback from Study 1 (trustful, natural, passive, reasonable, strict, enthusiastic, affectionate, and sincere), and their potential offensiveness in some cultures (idiot, loser, criminal, and abusive).

Participants in all seven countries/regions followed the same experimental procedures. Each participant provided ratings of all faces on all traits, of which 20 traits were rated twice for within-subject consistency (see our preregistration). The 80 traits were divided into 20 modules, each consisting of 4 distinct traits (the 20 retested traits were first assigned to distinct modules and then the other traits were randomly assigned across modules with the constraint that traits in the same module should be balanced in valence). All participants completed all 20 modules during multiple visits to the local offices in ten business days. Each module consisted of 5 blocks, with the retested trait always shown in the first and last blocks and the other traits shown in random order. The experimental procedure within each module was identical to Study 1.

Measure of reliability in Study 2
Data were first processed following our preregistered exclusion criteria A to C (see preregistration at https://osf.io/tbmsy): of the full sample with a preregistered size of $N = 30$ participants and $L = 300,000$ ratings at each of 7 locations ($N = 210$ total), we excluded from further analysis $n = 1$ participant in India and $l = 24,236$ ratings in North America, $l = 2,507$ ratings in Latvia, $l = 16,366$ ratings in Peru, $l = 3,178$ ratings in the Philippines, $l = 14,389$ ratings in India, $l = 9,117$ ratings in Kenya, and $l = 4,096$ ratings in Gaza. Registration criterion D was not applied for the analyses of within-subject consistency and between-subject consensus because it imposed a strict lower bound on the within-subject consistency to ensure data quality, which might lead to an overestimation of the reliability of the data.

All participants at all locations rated a subset of twenty traits twice for all faces. Analyses of within-subject consistency identical to those in Study 1 were performed for each of the seven datasets ($l = 100$ pairs of ratings across faces per participant for ca. $n = 28$ participants per location). We found acceptable within-subject consistency at all locations ($r_s > 0.20$, except for the ratings of competent, religious, anxious, and critical in India [$r_s = 0.18, 0.18, 0.19, 0.19$] and the ratings of anxious in Peru [$r = 0.19$]). As hypothesized in our preregistration, across all locations, ratings of traits regarding physical appearance had higher within-subject consistency (e.g., feminine, youthful, healthy, with mean $r_s = 0.74, 0.57, 0.51$, respectively) than traits that were more abstract (e.g., critical, anxious, religious, with mean $r_s = 0.31, 0.32, 0.33$, respectively), corroborating findings from Study 1 (Fig. 2).

Assessment of between-subject consensus at each location used data from all participants within the same location ($l = 100$ ratings per participant for the 100 faces from ca. $n = 28$ participants per trait per location). As hypothesized in our preregistration, traits regarding physical appearance such as feminine, youthful, beautiful, and baby-faced showed high between-subject
consensus in all seven locations (all ICCs > 0.86). At the other extreme, some locations had trait
ratings with near-zero consensus within that location (the ratings of compulsive in Gaza, prudish
in India and Kenya, self-critical in Gaza and the Philippines). This stood in contrast to the
findings from Study 1 where ICCs > 0.61 for all the one hundred traits (Fig. 2), and to the
samples from North America (ICCs > 0.61 for all traits) and Latvia (ICCs > 0.50 for all traits).

Data processing for RSA and dimensionality analysis in Study 2

To ensure high quality and complete data from individuals, we registered four exclusion criteria
(A-D) while data collection was underway and data had not yet been analyzed (see registration at
https://osf.io/tbmsy), in addition to those planned in our original preregistration
(https://osf.io/qxgmw). Analyses of representational similarity and dimensionality for both
aggregated and individual data were performed using data that were processed with exclusion
criteria A-D. Following those criteria, thirty-one participants across seven locations were
excluded for further analysis (n = 3 for North America, n = 2 for Latvia, n = 7 for Peru, n = 3 for
the Philippines, n = 10 for India, n = 2 for Kenya, and n = 4 for Gaza). Among those remaining
participants, n = 86 participants had complete data for all 80 traits—data from these 86
participants were used in the individual-level analyses (Fig. 4).

Relations between Face-impression Dimensions and Personality Dimensions

Our initial list of traits compiled from multiple sources included 435 most familiar personality-
trait adjectives in American English. Sixty-eight of those traits remained in our trait set after
our final selection process shown in Fig. 1 (together with the other thirty-two words from other
sources) and were included in the EFA analysis. To enable comparisons with our present
findings and the dimensions typically reported in studies of personality (e.g., the Big Five), we
collected self-report evaluations on each of these 68 personality traits from an independent sample of \( N = 343 \) participants via MTurk (i.e., self-reported personality ratings, no faces were shown; data can be assessed at [https://osf.io/4mvyt/](https://osf.io/4mvyt/)).

As expected, the five Big Five personality dimensions emerged in this dataset ([Extended Data Fig. 5c](https://osf.io/4mvyt/); using the same EFA method as for the face-trait ratings). To compare these five personality dimensions and our four face-impression dimensions, we computed Pearson’s correlations between these 68 traits’ factor loadings on the five personality dimensions and their factor loadings on our four face-impression dimensions (i.e., the factor loadings for this subset of 68 traits from [Extended Data Fig. 4a](https://osf.io/4mvyt/)). Our “warmth” dimension was most strongly associated with the “Agreeableness” personality dimension \((r = 0.79)\), our “competence” dimension was most strongly associated with the “Conscientiousness” personality dimension \((r = 0.71)\), our “femininity” dimension most strongly associated with the “Neuroticism” personality dimension \((r = 0.59)\), and our “youth” dimension was most strongly associated with the “Extraversion” personality dimension \((r = 0.64)\). The correlations between the “Openness” personality dimension and all our four dimensions were weak \((rs < 0.35)\). We provide a visualization of the correspondence between our four face-impression dimensions and the five personality dimensions in [Extended Data Fig. 5d](https://osf.io/4mvyt/).

To further explore whether our face-impression dimensions and the Big Five personality might share the same dimensional space, we analyzed the structural representation of each of the two datasets (face ratings, self-ratings without faces). Following previous approaches\(^{64}\), we extracted from each dataset one to five factors (unrotated for the one-factor level; “oblimin” rotation for the subsequent levels) and computed Pearson correlations between factor scores from different levels. We found considerable differences in the hierarchical structural representation of the
personality space versus the face-impression space (Extended Data Fig. 5e,f), indicating that trait ratings from faces and trait ratings of personality (without faces) are best characterized in two distinct spaces.

Data and code availability

All data, codes, and materials are available at Open Science Framework: https://osf.io/4mvyt/ and https://osf.io/xe6w/.

References

1. Todorov, A. *Face value: The irresistible influence of first impressions*. (Princeton University Press, 2017).
2. Lin, C., Adolphs, R. & Alvarez, R. M. Inferring Whether Officials Are Corruptible From Looking At Their Faces. *Psychological Science* 29, 1807–1823 (2018).
3. Todorov, A. Inferences of Competence from Faces Predict Election Outcomes. *Science* 308, 1623–1626 (2005).
4. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* 105, 11087–11092 (2008).
5. Collova, J. R., Sutherland, C. A. M. & Rhodes, G. 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* (2019) doi:10.1037/pspa0000167.
6. Stolier, R. M., Hehman, E., Keller, M. D., Walker, M. & Freeman, J. B. The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences* 115, 9210–9215 (2018).
7. Jones, B. C. et al. To Which World Regions Does the Valence-Dominance Model of Social Perception Apply?. PsyArXiv (2018) doi:10.31234/osf.io/n26dy.

8. Oh, D., Dotsch, R., Porter, J. & Todorov, A. Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General* (2019).

9. Mileva, M., Young, A. W., Kramer, R. S. S. & Burton, A. M. Understanding facial impressions between and within identities. *Cognition* 190, 184–198 (2019).

10. Stewart, L. H., Ajina, S., Getov, S. & Bahrami, B. Unconscious evaluation of faces on social dimensions. *Journal of Experimental Psychology: General* /fulltext/2012-08697-001.html (2012) doi:10.1037/a0027950.

11. Jessen, S. & Grossmann, T. Neural and Behavioral Evidence for Infants’ Sensitivity to the Trustworthiness of Faces. *Journal of Cognitive Neuroscience* 28, 1728–1736 (2016).

12. Getov, S., Kanai, R., Bahrami, B. & Rees, G. Human brain structure predicts individual differences in preconscious evaluation of facial dominance and trustworthiness. *Soc Cogn Affect Neurosci* 10, 690–699 (2014).

13. McCurrie, M. et al. Predicting First Impressions with Deep Learning. in 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017) 518–525 (2017). doi:10.1109/FG.2017.147.

14. Song, A., Li, L., Atalla, C. & Cottrell, G. Learning to see people like people. *arXiv:1705.04282 [cs]* (2017).

15. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences* 11, 77–83 (2007).
16. Saucier, G. & Goldberg, L. R. Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality* 10, 61–77 (1996).

17. Sutherland, C. A. M. *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).

18. Abir, Y., Sklar, A. Y., Dotsch, R., Todorov, A. & Hassin, R. R. The determinants of consciousness of human faces. *Nat Hum Behav* 2, 194–199 (2018).

19. Rule, N. O., Slepiain, M. L. & Ambady, N. A memory advantage for untrustworthy faces. (2012).

20. Olivola, C. Y. & Todorov, A. Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology* 46, 315–324 (2010).

21. Lin, C., Adolphs, R. & Alvarez, R. M. Cultural effects on the association between election outcomes and face-based trait inferences. *PLOS ONE* 12, e0180837 (2017).

22. Antonakis, J. & Eubanks, D. L. Looking Leadership in the Face. *Current Directions in Psychological Science* 26, 270–275 (2017).

23. Hamermesh, D. S. *Beauty Pays: Why Attractive People Are More Successful.* (Princeton University Press, 2011).

24. Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K. & Xie, S. Y. Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Soc Personal Psychol Compass* 13, e12431 (2019).
1. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology* 66, 519–545 (2015).

2. Hehman, E., Sutherland, C. A. M., Flake, J. K. & Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology* 113, 513–529 (2017).

3. Zebrowitz, L. A. & Montepare, J. M. Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass* 2, 1497 (2008).

4. Said, C. P., Sebe, N. & Todorov, A. Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces. (2009).

5. Rule, N. O., Ambady, N. & Hallett, K. C. Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology* 45, 1245–1251 (2009).

6. Todorov, A., Mende-Siedlecki, P. & Dotsch, R. Social judgments from faces. *Current Opinion in Neurobiology* 23, 373–380 (2013).

7. Secord, P. F., Dukes, W. F. & Bevan, W. Personalities in faces: I. An experiment in social perceiving. *Genetic Psychology Monographs* 49, 231–270 (1954).

8. Freeman, J. B. & Johnson, K. L. More Than Meets the Eye: Split-Second Social Perception. *Trends Cogn Sci* 20, 362–374 (2016).

9. Stolier, R. M., Hehman, E. & Freeman, J. B. Conceptual structure shapes a common trait space across social cognition. *PsyArXiv* (2019) doi:10.31234/osf.io/5na8m.
34. South Palomares, J. K., Sutherland, C. A. M. & Young, A. W. Facial first impressions and partner preference models: Comparable or distinct underlying structures? *Br J Psychol* 109, 538–563 (2018).

35. Wang, H. *et al.* A data-driven study of Chinese participants’ social judgments of Chinese faces. *PLOS ONE* 14, e0210315 (2019).

36. Tamir, D. I. & Thornton, M. A. Modeling the Predictive Social Mind. *Trends in Cognitive Sciences* 22, 201–212 (2018).

37. Sutherland, C. A. M. *et al.* Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* 127, 105–118 (2013).

38. Olivola, C. Y. & Todorov, A. Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. *J Nonverbal Behav* 34, 83–110 (2010).

39. Allport, G. W. & Odbert, H. S. Trait-names: A psycho-lexical study. *Psychological monographs* 47, i (1936).

40. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 135–146 (2017).

41. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* (2018).

42. DeBruine, L. & Jones, B. Face Research Lab London Set. (2017)

doi:10.6084/m9.figshare.5047666.v3.

43. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 1122–1135 (2015).

44. Chelnokova, O. *et al.* Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular Psychiatry* 19, 746–747 (2014).
45. King, D. E. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 41755–1758 (2009).

46. Michael Quinn Patton. *Qualitative Research & Evaluation Methods*. (SAGE Publications, 2002).

47. Walker, M., Schönborn, S., Greifeneder, R. & Vetter, T. The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLOS ONE* 13, e0193190 (2018).

48. Lundqvist, D., Flykt, A. & Öhman, A. The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet* 91, 2–2 (1998).

49. Morrison, D., Wang, H., Hahn, A. C., Jones, B. C. & DeBruine, L. M. Predicting the reward value of faces and bodies from social perception. *PLoS ONE* 12, e0185093 (2017).

50. Conway, J. M. & Huffcutt, A. I. A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. in (2003). doi:10.1177/1094428103251541.

51. Hayton, J. C., Allen, D. G. & Scarpello, V. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational research methods* 7, 191–205 (2004).

52. Martinez, J. E., Funk, F. & Todorov, A. Quantifying idiosyncratic and shared contributions to judgment. *Behav Res* (2020) doi:10.3758/s13428-019-01323-0.

53. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nature Communications* 10, 1258 (2019).

54. Brooks, J. A. & Freeman, J. B. Neuroimaging of person perception: A social-visual interface. *Neuroscience Letters* 693, 40–43 (2019).
55. Bainbridge, W. A., Isola, P. & Oliva, A. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 1323–1334 (2013).

56. Zhang, Z., Song, Y. & Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4352–4360 (IEEE, 2017). doi:10.1109/CVPR.2017.463.

57. Hehman, E., Xie, S. Y., Ofosu, E. K. & Nespoli, G. Assessing the point at which averages are stable: A tool illustrated in the context of person perception. (2018) doi:10.31234/osf.io/2n6jq.

58. Pearson, R., Mundfrom, D. & Piccone, A. A Comparison of Ten Methods for Determining the Number of Factors in Exploratory Factor Analysis. 39, 15 (2013).

59. Çokluk, Ö. & Koçak, D. Using Horn’s Parallel Analysis Method in Exploratory Factor Analysis for Determining the Number of Factors. *Educational Sciences: Theory & Practice* 16, 537–551 (2016).

60. Velicer, W. F. Determining the number of components from the matrix of partial correlations. *Psychometrika* 41, 321–327 (1976).

61. Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 504–507 (2006).

62. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2015).

63. DeYoung, C. G. Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology* 91, 1138–1151 (2006).

64. Goldberg, L. R. Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality* 40, 347–358 (2006).
Acknowledgements

We thank Dean Mobbs, R. Michael Alvarez, Mark Bowren, Antonio Rangel, Clare Sutherland, Uri Maoz, and William Revelle for their input, Remya Nair and Christopher J. Birtja for technology support, and Becky Santora for helping with testing participants in foreign locations through Digital Divide Data. Funded in part by NSF grants BCS-1840756 and BCS-1845958, and the Carver Mead New Adventures Fund.

Author contributions

C.L. and R.A. developed the study concept and designed the study; C.L. and U.K. prepared experimental materials; R.A. supervised the experiments and analyses; C.L. performed and supervised data collection; C.L. and U.K. performed data analyses; C.L. and R.A. drafted the manuscript; all authors revised and reviewed the manuscript and approved the final manuscript for submission.

Competing interest declaration

The authors declare no competing interests.

Additional information

Correspondence to clin7@caltech.edu.
Extended Data Fig. 1: Verifying the representativeness of the 100 selected traits and faces.

a, Word cloud of the 973 freely generated descriptions of our 100 faces (see Methods). All words that appeared at least twice are shown in the figure (words that appeared only once were excluded, as they were comprised mainly of misspelled words or words not included in the FastText vocabulary). The scale indicates frequency (ranging from 2 to 306 times).

b, Distribution of word similarities between the 100 selected traits and freely generated words (red dotted line), as well as between the 100 selected traits and freely generated words and the 100 selected traits (blue dotted line). The frequency values (y-axis) are shown for word similarities less than 0.8 (x-axis).

c, Dimensional representation of trait similarity in the UMAP space. The red circles indicate the 100 randomly selected traits, with distinct colors representing different nationalities. The circles (red and blue) indicate traintrait similarity, with the x-axis indicating feature similarity within the trait (100T10OF) and the y-axis indicating feature similarity between traits (100T10OF).

d, Dimensional representation of face similarity in the UMAP space. The color bar indicates the part of the face being represented, with colors ranging from purple (hair) to yellow (eyebrows). The face images are color-coded according to the part of the face they represent, with the x-axis indicating feature similarity within the trait (100T10OF) and the y-axis indicating feature similarity between traits (100T10OF).

e, UMAP dimensional representation of trait similarity in the UMAP space. The red circles indicate the 100 randomly selected traits, with distinct colors representing different nationalities. The circles (red and blue) indicate traintrait similarity, with the x-axis indicating feature similarity within the trait (100T10OF) and the y-axis indicating feature similarity between traits (100T10OF).
Distributions of word similarities. The similarity between two words was assessed with the cosine distance between the 300-feature vectors of the two words. The blue histogram plots the pairwise similarities among the 100 trait words we used in our study. The red histogram plots the similarities between each of the freely generated words and its closest counterpart among our 100 traits. Dashed lines indicate means. All freely generated words were found to be similar to at least one of our selected traits (all similarities between freely generated words [except for “moving” and “round”] and their closest counterparts in our trait set > 0.25). Eighty-five freely generated words were identical to those in our 100 selected traits. c, Uniform Manifold Approximation and Projection (UMAP, a dimensionality reduction technique similar to t-SNE but applicable also to nonlinear cases) of our 100 selected faces (stars), a larger set of frontal, neutral, white faces from multiple databases (dots with colors other than blue; $N = 632$), and ambient facial images in the wild (blue dots; with various angles, gazes, facial expressions, lighting, backgrounds, etc.; $N = 5376$). All faces were represented by the 128 computationally extracted features used by a state-of-the-art neural network for facial recognition. d, Automatic annotation of facial keypoints and facial parts, for automatically measuring 30 facial metrics (e.g., pupillary distance, eye width, eye length, eye size, eye shape, nose width, nose length, nose shape, lip thickness, lip fullness, cheekbone height, cheekbone prominence, various face width measures, various face height measures, face roundness, median luminance). e, UMAP of our 100 selected faces and the larger set of frontal, neutral, white faces from multiple databases (dots; $N = 632$). All faces were represented by the 30-dimensional vectors of facial metrics automatically measured in d, which are the ones largely used in the prior literature.
Extended Data Fig. 2: Variance and factorizability of ratings across 100 traits.

a, Each row plots the average ratings across participants for the 100 faces on a trait (grey dots), with the median (line in the box), the first quartile (left edge of the box), the third quartile (right edge of the box), and outliers that are more extreme than 3/2 times of the quartiles (open dots). b, Each row plots the mean (dot), median (triangle), and maximum (square) absolute correlations a trait has with all the other 99 traits (across faces, averaged over participants). The vertical dashed line indicates $r = 0.30$, which describes an inflection point in the curve of mean absolute correlations. The eight traits at the bottom (in bold) were excluded from EFA because of their low average correlations with all other traits (i.e. low factorizability).
Extended Data Fig. 3: Scree plots of data across eight samples.
a, Study 1 sample. b-h, Study 2 samples from b, North America, c, Latvia, d, Peru, e, the Philippines, f, India, g, Kenya, and h, Gaza. Circles plot the eigenvalues (the fraction of total common variance in the data as explained by each factor) of the original data across factors, ordered from the largest to the smallest. Triangles plot the 95th percentile of the eigenvalues of the simulated data from parallel analysis. The optimal number of factors to retain as recommended by each of the five methods is shown. Parallel analysis retains factors with eigenvalues (circles) greater than those from the simulated data (triangles) from 5,000 Monte Carlo simulations (see the close-up image for a clearer comparison). Cattell’s scree test retains factors to the left of the point from which the plotted ordered eigenvalues could be approximated with a straight line (i.e., “above the elbow”). The optimal coordinates index provides a non-graphical solution to Cattell’s scree test based on linear extrapolation. Empirical Bayesian information criterion (eBIC) retains factors that minimize the overall discrepancy between the population’s and the model’s predicted covariance matrices while penalizing model complexity (purple dots in inset graphs). Velicer’s minimum average partial (MAP) test is “most appropriate when component analysis is employed as an alternative to, or a first-stage solution for, factor analysis”\textsuperscript{60}. It is also included in our present study due to its popularity. MAP retains components by partialing those that resulted in the lowest average squared partial correlation. The MAP test gave variable numbers of components greater than 4; it is not plotted but the results are numerically provided in the legend inset.
Extended Data Fig. 4: Four dimensions from EFA in Study 1.

a, Factor loadings of trait ratings on the four dimensions from EFA. Each column plots the strength of the factor loadings (x-axis, absolute value) across traits (y-axis). Color indicates the sign of the loading (red for positive and blue for negative); more saturated colors for higher absolute values. Since oblique rotation allowed factors to be correlated with one another, the four dimensions turned out to be weakly correlated ($r_{13} = -0.33$, $r_{14} = -0.23$, $r_{23} = 0.21$, $r_{24} = 0.33$ [ps < 0.05]; $r_{12} = -0.15$, $r_{34} = 0.12$ [ps > 0.05]). b, Distributions of the 92 traits along each pair of dimensions based on their factor loadings on the four dimensions.
| Traits         | RC1     | RC2     | RC3     | RC4     |
|---------------|---------|---------|---------|---------|
| Patient       |         |         |         |         |
| Easygoing     |         |         |         |         |
| Agreeable     |         |         |         |         |
| Open-minded   |         |         |         |         |
| Empathetic    |         |         |         |         |
| Flexible      |         |         |         |         |
| Sensitive     |         |         |         |         |
| Submissive    |         |         |         |         |
| Trustful      |         |         |         |         |
| Helpful       |         |         |         |         |
| Ethical       |         |         |         |         |
| Happy         |         |         |         |         |
| Thoughtful    |         |         |         |         |
| Passive       |         |         |         |         |
| Optimistic    |         |         |         |         |
| Humble        |         |         |         |         |
| Natural       |         |         |         |         |
| Creative      |         |         |         |         |
| Sincere       |         |         |         |         |
| Curious       |         |         |         |         |
| Enthusiastic  |         |         |         |         |
| Affectionate  |         |         |         |         |
| Baby-faced    |         |         |         |         |
| Criminal      |         |         |         |         |
| Strict        |         |         |         |         |
| Religous      |         |         |         |         |
| Prejudiced    |         |         |         |         |
| Manipulative  |         |         |         |         |
| Compulsive    |         |         |         |         |
| Condescending |         |         |         |         |
| Skeptical     |         |         |         |         |
| Serious       |         |         |         |         |
| Abusive       |         |         |         |         |
| Jealous       |         |         |         |         |
| Aggressive    |         |         |         |         |
| Grumpy        |         |         |         |         |
| Discontented  |         |         |         |         |
| Angry         |         |         |         |         |
| Cruel         |         |         |         |         |
| Combative     |         |         |         |         |
| Critical      |         |         |         |         |
| Mean          |         |         |         |         |
| Interferer    |         |         |         |         |
| Leader-like   |         |         |         |         |
| Ambitious     |         |         |         |         |
| Competent     |         |         |         |         |
| Articulate    |         |         |         |         |
| Consistent    |         |         |         |         |
| Practical     |         |         |         |         |
| Determined    |         |         |         |         |
| Honest        |         |         |         |         |
| Mature        |         |         |         |         |
| Persistent    |         |         |         |         |
| Punctual      |         |         |         |         |
| Courageous    |         |         |         |         |
| Responsible   |         |         |         |         |
| Independent   |         |         |         |         |
| Conscientious |         |         |         |         |
| Confident     |         |         |         |         |
| Well-educated |         |         |         |         |
| Dignified     |         |         |         |         |
| Wise          |         |         |         |         |
| Clever        |         |         |         |         |
| Beautiful     |         |         |         |         |
| Socially      |         |         |         |         |
| Healthy       |         |         |         |         |
| Charismatic   |         |         |         |         |
| Charismatic   |         |         |         |         |
| Careful       |         |         |         |         |
| Trustworthy   |         |         |         |         |
| Reasonable    |         |         |         |         |
| Intellectual  |         |         |         |         |
| Outspoken     |         |         |         |         |
| Traditional   |         |         |         |         |
| Grouchy       |         |         |         |         |
| Atypical      |         |         |         |         |
| Self-pitying  |         |         |         |         |
| Unobservant   |         |         |         |         |
| Autistic      |         |         |         |         |
| Weed          |         |         |         |         |
| Ignorant      |         |         |         |         |
| Loser         |         |         |         |         |
| Idiot         |         |         |         |         |
| Prudish       |         |         |         |         |
| Feminine      |         |         |         |         |
| Self-critical |         |         |         |         |
| Religious     |         |         |         |         |
| Emotional     |         |         |         |         |
| Anxious       |         |         |         |         |
| Strong        |         |         |         |         |
| Youthful      |         |         |         |         |

Loading Strength

2

3

4

5
Extended Data Fig. 5: Comparison with existing dimensional frameworks.

a, Four dimensions from PCA in Study 1. Columns plot the strength of the loadings (x-axis, absolute value) on the first four varimax rotated principal components across all 92 traits (y-axis). Colors indicate the sign of the loading (red for positive and blue for negative); more
saturated colors for higher absolute values. The first four principal components without rotation accounted for 52%, 21%, 7%, and 5% of the variance in our data, 86% in total; the fifth accounted for 2%. b, Predicting trait judgments using different dimensional frameworks. Regressors were linear combinations of the traits that showed highest loadings in each dimensional framework (two traits for each dimension because there were only two traits that loaded highest on one of the 3D37’s dimensions; for example, for our 4D, the model consisted of eight regressors). Each row indicates three different models that regressed the ratings of a targeted trait (row name; which was not one of the regressors) on the three different sets of regressors from the three frameworks, and plots the adjusted R². c, Factor loading matrices from EFA using self-reported personality ratings on a subset of 68 personality traits (see Methods). The Big Five personality dimensions emerged from this dataset. d, Correspondence between our four face-impression dimensions and the Big Five personality dimensions. Each of the 68 traits was color-coded to indicate the dimension which it had the highest absolute loading on among the four face-impression dimensions (inner ring; refer to Extended Data Fig. 4a for the factor loadings) and the five personality dimensions (outer ring; refer to c for the factor loadings). We also calculated the correlations between the factor loadings on our four face-judgment dimensions and these five personality dimensions: our warmth dimension was most strongly associated with the Agreeableness personality dimension ($r = 0.79$), our competence dimension was most strongly associated with the Conscientiousness personality dimension ($r = 0.71$), our femininity dimension most strongly associated with the Neuroticism personality dimension ($r = 0.59$), and our youth dimension was most strongly associated with the Extraversion personality dimension ($r = 0.64$). The correlations between the Openness personality dimension and all our four dimensions were weak ($r < 0.35$). e, Structural representation of the self-reported
personality ratings on the subset of 68 personality traits (see Methods). Boxes indicate factors; labels were given when the interpretation of the factor was clear. Arrows and numbers indicate Pearson’s correlations between factor scores of factors from different levels (paths with correlations of 0.4 or above are shown). f, Structural representation of the face-trait ratings from Study 1 on the subset of the same 68 personality traits as in e.
Extended Data Fig. 6: Confirmatory analysis with autoencoders and cross-validation.

a, We trained an autoencoder model with only one hidden layer to confirm our EFA results (see Methods). b, An autoencoder model with multiple hidden layers was used to explore possible hierarchical factor structures. c, The means (points) and standard deviations (bars) of the explained variance on the training data by autoencoders with one hidden layer (as in a) that varied in the number of neurons. Colors indicate different configurations of activation functions.
in the encoder and decoder layers (linear, tanh, sigmoid, rectified linear activation unit, L1-norm regularization). The configuration with linear functions in both the encoder and decoder layers performed the best (AE-linear-linear). d, Means (points) and standard deviations (bars) of the explained variance on the test data by autoencoders shown in e. Results showed that a four-dimensional representation (with the AE-linear-linear configuration) described the data well (explained variance = 75%, comparable to PCA [75%]) and the increases in performance beyond four dimensions were trivial (<1%). Critically, this four-dimensional representation reproduced our four dimensions ($r_s = 0.98, 0.92, 0.91, 0.94$ [SDs = 0.01, 0.05, 0.02, 0.05] for the four dimensions between factor loadings from EFA and decoder layer weights with varimax rotation).

e, The means (points) and standard deviations (bars) of the explained variance on the training data by autoencoders with multiple hidden layers (as in b) that varied in the number of neurons in the first and third (since autoencoders are by definition symmetric) and middle hidden layers. We plot only the results from linear activation functions in both the encoder and decoder layers (corresponding to label “AE-linear-linear” in c and d). f, Means (points) and standard deviations (bars) of the explained variance on the test data given by the autoencoders shown in e. Including hierarchical latent structure did not improve model performance.
Extended Data Fig. 7: Four factors extracted from aggregated data in Study 2.

The seven panels plot results for samples from a, North America, b, Latvia, c, Peru, d, the Philippines, e, India, f, Kenya, and g, Gaza. Each column plots the strength of the factor loadings across the 80 traits. The color of the bar indicates the sign of the loading (red: positive; blue: negative); the length and saturation of the bar indicates the magnitude of the loading.
Extended Data Fig. 8: Four factors extracted from individual data in Study 2.

a, Examples of the factor loading matrices of three participants (a1 for North America participant #2, a2 for Latvia participant #1, a3 for Latvia participant #15). b, Factor congruence between Study 1 sample and each individual participant who had a complete dataset in Study 2 (n = 86, who had data for all 80 traits after data exclusion according to preregistered criteria). Each table plots the Tucker indices of factor congruence (with orthogonal Procrustes rotation) between the four dimensions found in aggregate-level data in Study 1 (columns; W: warmth, C: competence, F: femininity, Y: youth) and those found in individual-level data in Study 2 (rows). The row label of each table indicates the location (NA for North America, LV for Latvia, PE for Peru, GZ for Gaza, PH for the Philippines, KE for Kenya, IN for India) and ID of the participant. The numbers report the Tucker indices. The color scale shows the sign and strength of the indices.
| TRAITS   | DEFINITIONS                                                                                                                                                                                                 |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| abusive  | A person who is extremely offensive and insulting                                                                                                                                                    |
| affectionate | A person who is comfortable showing his/her love, warmth, and kindness                                                                                                                             |
| aggressive | A person who pursues his/her aims and interests forcefully, sometimes with physical force                                                                                                           |
| agreeable | A person who is kind, cooperative, and sympathetic                                                                                                                                                    |
| ambitious | A person who has a strong desire and determination to succeed in their goals                                                                                                                              |
| angry    | A person who is usually angry                                                                                                                                                                |
| anxious  | A person who stresses and worries about things                                                                                                                                                       |
| articulate | A person who speaks fluently and clearly, and who can express their ideas well                                                                                                                      |
| atypical | The structure, texture, shape or other aspects of the appearance of the face is unusual or rare                                                                                                           |
| autistic | A person who has autism spectrum disorder—a developmental disorder characterized by troubles with social interaction and communication, and by restricted and repetitive behavior |
| baby-faced | A person who has facial features resembling a baby                                                                                                                                                |
| beautiful | A person who looks appealing and physically attractive                                                                                                                                                |
| bossy    | A person who likes giving people orders and wants things his/her own way                                                                                                                               |
| careful  | A person who works and thinks in a cautious, thorough, or thoughtful way to avoid potential danger                                                                                                   |
| charismatic | A person who is interesting and likeable because they have a charming personality                                                                                                                |
| clever   | A person who is quick to understand and learn, and who can figure things out quickly                                                                                                                  |
| combative | A person who likes to argue or pick a fight                                                                                                                                                        |
| compulsive | A person who is efficient and capable to do things in general                                                                                                                                     |
| condescending | A person who thinks he/she is better than others and puts other people down                                                                                                                              |
| confident | A person who is sure about his/her own abilities, correctness, and successfulness                                                                                                                     |
| conscientious | A person who does his/her work or duty thoroughly and responsibly                                                                                                                                   |
| conservative | A person who sticks to traditional values, especially in politics or religion, and who does not like new ideas or changes                                                                          |
| consistent | A person who behaves or responds in the same way over time; reliable                                                                                                                               |
| courageous | A person who is not afraid to do the right thing, even if it is dangerous to them                                                                                                                                 |
| creative  | A person who has good imagination or original ideas                                                                                                                                                  |
| criminal  | A person who looks like they could commit a crime                                                                                                                                                   |
| critical  | A person who judges others harshly, and often makes disapproving comments                                                                                                                              |
| cruel    | A person who willfully causes pain or suffering to other people or to animals, and feels no concern about it                                                                                           |
| curious  | A person who is eager to learn about or experience new things                                                                                                                                     |
| defensive | A person who is easily offended and always guards themselves against criticism                                                                                                                         |
| determined | A person who is able to make firm decisions and is resolved not to change them                                                                                                                          |
| dignified | A person who is polite and composed, and always shows good and respected manners                                                                                                                     |
| disordered | A person who is untidy and not organized                                                                                                                                                          |
| easygoing | A person who is relaxed, tolerant, and not prone to rigid rules or bouts of temper                                                                                                                  |
| emotional | A person who shows his/her feelings and laughs and cries easily                                                                                                                                       |
| empathetic | A person who is able to understand and share the feelings of others                                                                                                                                  |
| energetic | A person who is very active and full of energy                                                                                                                                                       |
| enthusiastic | A person who is filled with eager enjoyment and interest                                                                                                                                           |
| ethical  | A person who is careful to do things that are morally right to do                                                                                                                                 |
| feminine | A person whose facial appearance looks like a woman                                                                                                                                                   |
| flexible  | A person who is ready and able to change so as to adapt to different circumstances                                                                                                                   |
| grumpy   | A person who is bad-tempered and always complaining                                                                                                                                                  |
| happy    | A person who is usually cheerful                                                                                                                                                                     |
| healthy  | A person who is in good health                                                                                                                                                                      |
| helpful  | A person who gives help when others are in need                                                                                                                                                     |
| homosexual | A person who is sexually attracted to people of his/her own sex                                                                                                                                |
| humble   | A person who is modest and does not boast                                                                                                                                                           |
| idiot    | A person who is stupid                                                                                                                                                                              |
| ignorant | A person who doesn't know anything, and is also usually unaware of that                                                                                                                             |
| income   | A person’s income level                                                                                                                                                                             |
| independent | A person who is able to think and act without being influenced by others                                                                                                                            |
| intellectual | A person who thinks a lot about the deeper meaning of things and likes to analyze things                                                                                                              |
| intense  | A person who is very serious and expresses strong feelings                                                                                                                                          |
| jealous  | A person who feels resentment about what other people have                                                                                                                                        |
| leader-like | A person who can take charge and help a group accomplish a goal                                                                                                                                    |
| loser    | A person who fails frequently or is generally unsuccessful in life                                                                                                                                |
| **manipulative** | A person who likes to control people in order to meet his/her own needs |
|------------------|---------------------------------------------------------------------|
| **mature**       | A person who thinks and behaves like a responsible adult            |
| **mean**         | A person who is unkind, inconsiderate, and doesn’t share things    |
| **natural**      | A person who is relaxed and spontaneous                             |
| **nosey**        | A person who is overly curious about other people’s business        |
| **open-minded**  | A person who is willing to try new things or to hear and consider new ideas |
| **optimistic**   | A person who is hopeful and confident about the future              |
| **outspoken**    | A person who is frank in stating his/her opinions especially if they are critical or controversial |
| **passive**      | A person who allows things to happen or accepts what others do, without resistance or trying to change anything |
| **patient**      | A person who is able to accept or tolerate delays or problems and is very relaxed about getting things done |
| **persistent**   | A person who is able to continue in a course of action in spite of difficulty or opposition |
| **practical**    | A person who is sensible and realistic in dealing with a situation or problem |
| **prudish**      | A person who is overly proper and cannot stand hearing any sexual reference |
| **punctual**     | A person who is always on time                                      |
| **reasonable**   | A person who makes sense and whose opinions most people would agree with |
| **rebellious**   | A person who resists authority, control, or convention and wants to have their own way |
| **religious**    | A person who practices religion and believes in their faith         |
| **reserved**     | A person who tends not to show their emotions or opinions and is quiet |
| **responsible**  | A person who accepts the consequences of his or her own actions and decisions |
| **sarcastic**    | A person who likes using irony in order to mock others               |
| **self-critical**| A person who holds himself/herself responsible for any failures, always questioning if they did the right thing or not |
| **self-pitying** | A person who feels sorry for themselves                              |
| **sensitive**    | A person who is aware of or careful about others' attitudes, feelings, or circumstances |
| **serious**      | A person who shows deep thoughts and who doesn't smile or laugh easily |
| **shallow**      | A person who is concerned only about silly or inconsequential things; superficial |
| **sincere**      | A person who says what he/she genuinely feels or believes           |
| **skeptical**    | A person who questions things and is not easily convinced           |
| **sociable**     | A person who is friendly and enjoys talking and engaging in activities with other people |
| **strict**       | A person who follows rules exactly, and expects others to follow rules exactly |
| **strong**       | A person who is physically vigorous and is able to exert great bodily or muscular power |
| **submissive**   | A person who shows a willingness to be controlled by others or conforms to the authority or will of others |
| **thoughtful**   | A person who is considerate of others' needs                        |
| **thrifty**      | A person who uses money and other resources carefully and not wastefully |
| **traditional**  | A person who likes to do things the way they have always been done and accepted in the past |
| **trustful**     | A person who tends to trust other people easily (note: this is different from being trustworthy) |
| **trustworthy**  | A person who can be relied on as honest and truthful                |
| **unobservant**  | A person who does not notice things                                 |
| **weird**        | A person who does strange or bizarre things                        |
| **well-educated**| A person who has completed a high level of education, such as bachelor's, master’s and doctorate degrees |
| **white**        | A person whose face looks like they are Caucasian                   |
| **wise**         | A person who has mature experience, knowledge, and good judgments  |
| **youthful**     | A person who looks young                                           |

2 Extended Data Table S1: Definitions of 100 trait words.

The definition of each trait word was provided to participants in our study to eliminate possible heterogeneity in how each individual understands the meaning of a trait word. These definitions were obtained from Google dictionary, with necessary modifications to make the definition easy to understand and fit the context of describing a person.
### a

| Traits from our set [traits in 2D framework] | Valence | Dominance |
|---------------------------------------------|---------|------------|
| Sociable [Sociable]                         | 0.89    | 0.14       |
| Weird [Weird]                               | -0.88   | 0.13       |
| Beautiful [Attractive]                      | 0.86    | 0.03       |
| Confident [Confident]                        | 0.85    | -0.53      |
| Responsible [Responsible]                   | 0.82    | 0.12       |
| Trustworthy [Trustworthy]                   | 0.77    | 0.38       |
| Wise [Intelligent]                          | 0.70    | -0.06      |
| Thoughtful [Caring]                          | 0.64    | 0.55       |
| Happy [Unhappy]                             | 0.54    | 0.45       |
| Submissive [Dominant]                        | -0.18   | 1.00       |
| Aggressive [Aggressive]                      | -0.13   | -0.90      |
| Mean [Mean]                                 | -0.22   | -0.86      |
| Emotional [Emotionally stable]              | 0.48    | 0.54       |

### b

| Traits from our set [traits in 3D framework] | Approachability | Youthful/Attractiveness | Dominance |
|---------------------------------------------|-----------------|-------------------------|-----------|
| Wise [Intelligent]                          | 0.92            | -0.37                   | 0.02      |
| Trustworthy [Trustworthy]                   | 0.80            | 0.20                    | 0.24      |
| Agreeable [Approachable]                    | 0.68            | 0.20                    | 0.43      |
| Confident [Confident]                        | 0.63            | 0.13                    | -0.63     |
| Happy [No Smile-Big Smile]                  | 0.61            | 0.21                    | 0.26      |
| Beautiful [Attractive]                      | 0.60            | 0.54                    | -0.23     |
| Feminine [Feminine]                         | 0.31            | 0.28                    | 0.20      |
| Youthful [Youthful]                         | -0.11           | 0.98                    | 0.12      |
| Baby-faced [Baby-faced]                     | -0.09           | 0.82                    | 0.31      |
| Healthy [Healthy]                           | 0.52            | 0.67                    | -0.25     |
| White [Pallid-Tanned]                       | 0.16            | 0.27                    | 0.05      |
| Submissive [Dominant]                       | 0.05            | 0.21                    | 0.88      |
| Aggressive [Aggressive]                     | -0.38           | -0.12                   | -0.79     |

### c

| Number of traits remaining in the model | Factor 1 with Warmth | Factor 2 with Competence | Factor 3 with Femininity | Factor 4 with Youth |
|----------------------------------------|-----------------------|---------------------------|----------------------------|---------------------|
| 91                                     | 1.00**                | 1.00**                    | 1.00**                     | 0.99**              |
| 90                                     | 1.00**                | 1.00**                    | 0.98**                     | 0.99**              |
| 89                                     | 1.00**                | 1.00**                    | 0.98**                     | 0.99**              |
| 88                                     | 1.00**                | 1.00**                    | 0.98**                     | 0.99**              |
| 87                                     | 0.99**                | 1.00**                    | 0.97**                     | 0.98**              |
| 86                                     | 1.00**                | 1.00**                    | 0.97**                     | 0.98**              |
| 85                                     | 0.99**                | 0.99**                    | 0.97**                     | 0.98**              |
| 84                                     | 0.99**                | 0.99**                    | 0.97**                     | 0.97**              |
| 83                                     | 0.99**                | 0.99**                    | 0.96**                     | 0.97**              |
| 82                                     | 0.99**                | 0.99**                    | 0.96**                     | 0.97**              |
| 81                                     | 0.99**                | 0.99**                    | 0.96**                     | 0.96**              |
** Extended Data Table S2: Dimensional analyses on subsets of trait ratings.

a, Factor loadings from EFA on the subset of data corresponding to the 13 traits from our set (first column) that are the same or most similar to those used in a prior study that discovered the popular 2D framework (in brackets). Two factors—the optimal number of factors as indicated by both Kaiser’s Rule and Cattell’s Scree Test—were extracted and rotated with oblimin. The
largest absolute loading across factors for each trait is highlighted in bold. **b**, Factor loadings from EFA on the subset of data corresponding to the 13 traits from our set (first column) that are the same or most similar to those used in a prior study that discovered the popular 3D framework (in brackets). Three factors—the optimal number of factors as indicated by Kaiser’s Rule, Cattell’s Scree Test, and the optimal coordinates index—were extracted and rotated with oblimin. The largest absolute loading across factors for each trait is highlighted in bold. **c**, Pearson’s correlations between our four dimensions (Fig. 2) and the four factors extracted from subsets of traits as a function of their similarity and meaning clarity. All pairs of traits were ranked according to their semantic similarity (using FastText, see Methods); for each pair of traits, we removed the one with the lower clarity ratings given by an independent set of participants; we removed traits one by one starting from the pair with the highest similarity to the pairs with lowest similarity. A four-factor solution was extracted from each subset of traits (using the same EFA method as for the full set of traits). To obtain the relations between the factors found in these subsets and the warmth, competence, femininity, and youth factors, we assessed the value of every face on each factor (i.e., factor scores, computed using R function factor.scores with method “tenBerge”) and then correlated the faces’ scores between different factors. **d**, Factor loadings from EFA on the subset of data corresponding to the smallest subset of specific traits that still yields our four dimensions. The largest absolute loading across four factors for each trait is highlighted in bold. The four dimensions accounted for 88% of the common variance in this subset of data.