Four dimensions characterizing comprehensive trait judgments of faces

Chujun Lin

California Institute of Technology

Umit Keles

California Institute of Technology

Ralph Adolphs (✉ Radolphs@caltech.edu)

California Institute of Technology

Article

Keywords: face judgments, traits, psychological dimensions, neural networks

Posted Date: July 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-41215/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on August 27th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-25500-y>.

Abstract

People readily attribute many traits to faces: some look beautiful, some competent, some aggressive. 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 across social and developmental psychology, social neuroscience, and 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.

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*)^{13,14}. These trait judgments are made ubiquitously and rapidly¹⁵, and are known to influence most subsequent processing, such as conscious perception¹⁶ and memory¹⁷ of the face. Although trait judgments of faces are in many cases inaccurate and reveal more about our own stereotypes than ground truth^{15,18}, 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^{19–23}.

Despite the considerable amount of work on the topic^{15, 24–33}, it remains unclear how people make these rapid judgements: 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^{1–10, 15,24,26, 34–36}, but those approaches used only a small number of trait words (12 to 18) that were common across studies^{14,34,37} or in use by lay people^{1,13}. 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.

Here we argue that to understand the true dimensionality of trait judgments from faces, it is essential to investigate a more comprehensively sampled set of trait words. To meet this challenge, we assembled an extensive list of trait words that people use to describe faces from multiple sources^{12–15, 19, 21, 25–31, 33, 38, 39} and applied a pre-trained neural network to derive a representative subset of 100 traits (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 the 100 selected traits were representative of the trait words people spontaneously generated for the selected 100 face images (Fig. 2a–b), and that the 100 selected face images were representative of the structural physiognomy of natural faces (Fig. 2c–d), although we note that only Caucasian faces with no emotional expressions were included [see **Methods**]. We collected ratings of the 100 faces on the 100 traits 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**).

Results

Four dimensions underlie comprehensive trait judgments of faces

Study 1 examined the underlying dimensions of the ratings that participants had given to the faces (ratings aggregated across participants) by first applying an exploratory method (exploratory factor analysis [EFA]; preregistered) and subsequently a confirmatory method with cross-validation (an autoencoder artificial neural network [ANN]). We confirmed that these ratings showed sufficient variance (**Supplementary Fig. 2a**), within-subject consistency (assessed with Pearson's correlations, $M = 0.47$, $Range = [0.28, 0.84]$, as well as linear mixed-effect modeling [preregistered]; Fig. 3), and between-subject consensus (preregistered; all $ICCs > 0.60$) [Fig. 3 and **Methods**]. Eight traits with low factorizability were excluded from further analyses (**Supplementary Fig. 2b**; including them did not change the dimensions we eventually found).

We determined the optimal number of factors to retain in EFA using five widely recommended methods^{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 (**Supplementary 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 ($r_{13} = -0.33$, $r_{14} = -0.23$, $r_{23} = 0.21$, $r_{24} = 0.33$ [$ps < 0.05$]; $r_{12} = -0.15$, $r_{34} = 0.12$ [$ps > 0.05$]). 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. 3). We interpreted these four factors as describing judgments of warmth, competence, femininity, and youth (Fig. 3; see **Supplementary Fig. 4a** for factor loadings) [see **Methods**].

To corroborate the four dimensions discovered from EFA, we applied an approach with minimal assumptions—artificial neural networks (ANN) with cross-validation to compare different factor structures (see **Methods**). Autoencoder ANNs with one hidden layer that differed in the number of neurons (range from 1 to 10) were constructed (Fig. 4a). These ANNs were trained on half of the data (i.e., aggregated ratings across half of the participants) and tested on the other held-out half (Adam optimization algorithm⁵² and mean squared error loss function with a batch size of 32 and 1500 epochs were used to train the ANNs, repeated for 50 iterations). Both linear and nonlinear activation functions were examined (Fig. 4b). Model performance of the best configuration (i.e., linear activation functions in both the encoder and decoder layers) increased substantially as the number of neurons in the hidden layer increased from 1 to 4 (explained variance on the test data increased by 18%, 5%, and 5%, respectively); the improvement was trivial beyond 4 neurons (increased by less than 1%) [Fig. 4c]. Critically, the four-dimensional representation learned by the ANN reproduced the four dimensions discovered from EFA (mean r s = 0.98, 0.92, 0.91, 0.94 [SDs = 0.01, 0.05, 0.02, 0.05] between factor loadings from EFA and the ANN's decoder layer weights with varimax rotation) and confirmed good performance (explained variance obtained with linear activation functions was 75% [SD = 0.6%] on the test data, comparable to PCA).

Comparison with existing dimensional frameworks

Prior work^{1,13,14,35,37} suggests that the various words people use to describe faces can be represented by two or three dimensions. 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^{13,14,37}, and reproduced the same four dimensions as reported above (**Supplementary Fig. 5a**).

Instead, the 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^{13,37}). The 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 (Table 1).

a

b

Traits from our set [traits in 2D framework ¹³]	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

Traits from our set [traits in 3D framework ³⁷]	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

Table 1 Factor loadings from EFA on subsets of 13 traits used in previous studies.

a, Factor loadings from EFA on the subset of data corresponding to 13 traits (first column) that are the same or most similar to those used in a prior study that discovered the popular 2D framework¹³ (first column, in brackets). Two factors—the optimal number of factors as indicated by both Cattell's Scree Test and empirical BIC—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 13 traits (first column) that are the same or most similar to those used in a prior study that discovered the popular 3D framework³⁷ (first column, in brackets). Three factors—the optimal number of factors as indicated by Cattell's Scree Test, the optimal coordinates index, Velicer's MAP test, and empirical BIC—were extracted and rotated with oblimin. The largest absolute loading across factors for each trait is highlighted in bold.

We next showed that the first two dimensions (warmth, competence) discovered from comprehensive trait judgments were different from the two dimensions of the popular prior 2D framework (valence, dominance). Analyses with the subset of 13 traits (Table 1a) showed that, replicating prior findings¹³, judgments of traits such as *sociable*, *trustworthy*, *responsible*, and *weird* were represented by the valence

dimension (absolute r s = 0.94, 0.88, 0.86, 0.85 between factor scores on the valence dimension and ratings for these traits across the 100 faces), but less well represented by the warmth dimension (absolute r s = 0.47, 0.67, 0.44, 0.23; see also **Supplementary Fig. 4a**); the valence and warmth dimensions were moderately correlated (absolute r = 0.41 between factor scores). Since using PCA we replicated the four dimensions from our full dataset and the 2D framework from the subset of 13 traits, we repeated the above analyses using PCA scores, which confirmed that the valence and warmth dimensions best represented different types of trait judgments (r = 0.09, p = 0.370). Similarly, as previously found¹³, judgments on *aggressive* and *submissive* were represented by the dominance dimension (absolute r s = 0.94, 0.95), but not by the competence dimension (absolute r s = 0.15, 0.14, p s > 0.05; see also **Supplementary Fig. 4a**); the two dimensions were not significantly correlated (r = 0.01, p = 0.894); these results were corroborated by analyses using PCA scores (r = -0.09, p = 0.383).

Finally, 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 loaded on one of the dimensions in the 3D framework, Table 1b), we found that the 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 (**Supplementary Fig. 5b**; mean adjusted R-squared across all predictions was 0.81 for the four-dimensional framework, 0.72 for the 3D framework, and 0.72 for the 2D framework).

Robustness of the four dimensions

We quantified the robustness of our results both across different numbers of trait words and across different numbers of participants. First, we removed trait words one by one and reperformed EFA to extract four factors as before (all pairs of trait words were ranked from the most to the least similar, the trait with lower clarity rating was removed from each pair). The four dimensions discovered from the full set versus the subsets of traits were highly correlated (Fig. 5a; see **Supplemental Table 2a** for the complete list of correlations). Second, we randomly removed participants one by one (50 randomizations each) and used the new aggregated ratings for EFA to show that the four dimensions discovered from the full dataset were robust to participant sample size (Fig. 5b; Tucker indices of factor congruence > 0.95 for all sub-datasets with no fewer than 19 participants per trait).

Finally, we extracted the smallest subset of specific trait words that still yielded the four-dimensional space discovered from the full dataset, a subset of 18 trait words that could be used most efficiently in future studies when collecting ratings for a larger set of traits is not feasible (**Supplementary Table 2b**).

Generalizability across different countries and regions

Prior work has reported both common³ and discrepant dimensions in different cultures^{13,14,24,35,37}. 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. 6a and **Supplementary 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 shown in **Supplementary Fig. 6**).

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. 6b); but, as expected, reproducibility was attenuated by the data quality available (as assessed by within-subject consistency, Fig. 6c)

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. 7a-b).

We next analyzed the dimensionality of each individual dataset. Parallel analysis (preregistered) showed that a four-dimensional space was most common (Fig. 7c) 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 the four-dimension space was successfully reproduced in individual participants (see examples of factor loading matrices in **Supplementary Fig. 7a**, and Tucker indices for all participants in **Supplementary Fig. 7b**), but also found a considerable amount of individual differences, in line with prior research⁵⁴.

Discussion

Across two large-scale, pre-registered studies we found that comprehensive trait judgments of faces are best described by a four-dimensional space (Figs. 3–4), with dimensions interpreted as warmth, competence, femininity and youth (**Supplementary Fig. 4**). We showed that our divergence from prior work was not due simply to methodological differences (**Supplementary Fig. 5a**), but to the prior lack of comprehensively and representatively sampled trait words (Figs. 1–2 and Table 1).

We showed that the warmth and competence dimensions reported here were different from the valence and dominance dimensions previously proposed. These findings help 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 cognition¹¹. The other two dimensions we found, femininity and youth, are likely linked to overgeneralization²⁷ and corroborate recent neuroimaging findings on social categorization from face perception^{32,55}.

This four-dimensional space was largely reproduced across samples from different regions, even using different languages (Spanish in Peru) [Fig. 6 and **Supplementary Fig. 6**], as well as in individual participants (although this was more difficult to assess, due to data quality) [Fig. 7 and **Supplementary Fig. 7**]. However, despite the predominance of the four-dimensional space, we also found notable variation across samples and individuals (Figs. 6b, 7c). Since the sources of this variation are unknown and may largely reflect measurement error (Figs. 6c, 7b), 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^{15,24,27}, 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 judgments⁵⁶.

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 (**Supplementary 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”).

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 (**Supplementary 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. 2a-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 sources^{12–15, 19, 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 library⁴⁰; 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 **Supplementary Fig. 1a** and **Methods** below), using the 300 computationally extracted semantic dimensions (Fig. 2a-b).

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 (**Supplementary Table 1**; 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 Database⁴³, the Chicago Face Database⁴², and the Face Research Lab London Set⁴¹. 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 library⁴⁴, 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 accuracy⁴⁴. 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 analysis⁴⁶ to compare the distribution of our selected faces with a) $N = 632$ neutral, frontal, white faces from a broader set of databases^{47–49} (Fig. 2c-d) and b) $N = 5376$ white faces “in the wild”^{57,58} that varied in angle, gaze, facial expression, lighting, and backgrounds (**Supplementary Fig. 1b**), using the 128

computationally extracted facial identity dimensions⁴⁴ as well as 30 traditional facial metric dimensions⁴² (Supplementary Fig. 1c).

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/>.

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 faces⁵⁹: 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 $\geq 95\%$).

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 (**Supplementary 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. 3].

Determination of the optimal number of factors

As recommended^{50,51,60,61}, 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)^{60,61}. 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"⁶². 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.

Labeling of Dimensions

Dimensionality reduction methods do not provide labels for the factors discovered, which must instead be interpreted by the investigators. We note that our third and fourth dimensions describe stereotypes related to gender (femininity-stereotypes) and age (youth-stereotypes) commonly reported in the literature¹¹. 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, autoencoders⁶³, 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^{13,37}, the 4D framework from EFA). 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 algorithm⁵² 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, Fig. 4b-c); a simple linear activation function ended up with the best results.

Existing frameworks^{13,37} 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). 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).

In addition, we tested potential hierarchical factor structure in our data by adding 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 (i.e., three hidden layers). Results showed that adding hidden layers did not increase model performance.

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 traits⁵⁹); 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>. 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, nosey, 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.

Measures 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 (Figs. 3–4).

Assessment of between-subject consensus at each location used data from all participants within the same location ($I = 100$ 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. 3), 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. 7).

Data and code availability

All data, codes, and materials are available at Open Science Framework: <https://osf.io/4mvyt/> and <https://osf.io/xeb6w/>.

Declarations

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.

References

1. 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.
2. 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).
3. 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.
4. 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).
5. Mileva, M., Young, A. W., Kramer, R. S. S. & Burton, A. M. Understanding facial impressions between and within identities. *Cognition* **190**, 184–198 (2019).
6. 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.
7. Jessen, S. & Grossmann, T. Neural and Behavioral Evidence for Infants' Sensitivity to the Trustworthiness of Faces. *Journal of Cognitive Neuroscience* **28**, 1728–1736 (2016).
8. 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).
9. 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.
10. Song, A., Li, L., Atalla, C. & Cottrell, G. Learning to see people like people. *arXiv:1705.04282 [cs]* (2017).
11. 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).
12. 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).

13. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* **105**, 11087–11092 (2008).
14. 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).
15. Todorov, A. *Face value: The irresistible influence of first impressions*. (Princeton University Press, 2017).
16. 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).
17. Rule, N. O., Slepian, M. L. & Ambady, N. A memory advantage for untrustworthy faces. (2012).
18. 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).
19. 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).
20. Lin, C., Adolphs, R. & Alvarez, R. M. Inferring Whether Officials Are Corruptible From Looking At Their Faces. *Psychological Science* **29**, 1807–1823 (2018).
21. Todorov, A. Inferences of Competence from Faces Predict Election Outcomes. *Science* **308**, 1623–1626 (2005).
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., Stoller, 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).
25. 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).
26. 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).
27. Zebrowitz, L. A. & Montepare, J. M. Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass* **2**, 1497 (2008).
28. Said, C. P., Sebe, N. & Todorov, A. Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces. (2009).
29. 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).

30. Todorov, A., Mende-Siedlecki, P. & Dotsch, R. Social judgments from faces. *Current Opinion in Neurobiology* **23**, 373–380 (2013).
31. Secord, P. F., Dukes, W. F. & Bevan, W. Personalities in faces: I. An experiment in social perceiving. *Genetic Psychology Monographs* **49**, 231–270 (1954).
32. Freeman, J. B. & Johnson, K. L. More Than Meets the Eye: Split-Second Social Perception. *Trends Cogn Sci* **20**, 362–374 (2016).
33. Stoller, 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. DeBruine, L. & Jones, B. Face Research Lab London Set. (2017) doi:10.6084/m9.figshare.5047666.v3.
42. 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).
43. Chelnokova, O. *et al.* Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular Psychiatry* **19**, 746–747 (2014).
44. King, D. E. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 41755–1758 (2009).
45. Michael Quinn Patton. *Qualitative Research & Evaluation Methods*. (SAGE Publications, 2002).
46. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* (2018).
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. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2015).
53. Paunonen, S. V. On Chance and Factor Congruence Following Orthogonal Procrustes Rotation. *Educational and Psychological Measurement* **57**, 33–59 (1997).
54. Martinez, J. E., Funk, F. & Todorov, A. Quantifying idiosyncratic and shared contributions to judgment. *Behav Res* (2020) doi:10.3758/s13428-019-01323-0.
55. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nature Communications* **10**, 1258 (2019).
56. Brooks, J. A. & Freeman, J. B. Neuroimaging of person perception: A social-visual interface. *Neuroscience Letters* **693**, 40–43 (2019).
57. Bainbridge, W. A., Isola, P. & Oliva, A. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* **142**, 1323–1334 (2013).
58. 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.
59. 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.
60. Pearson, R., Mundfrom, D. & Piccone, A. A Comparison of Ten Methods for Determining the Number of Factors in Exploratory Factor Analysis. **39**, 15 (2013).
61. 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).
62. Velicer, W. F. Determining the number of components from the matrix of partial correlations. *Psychometrika* **41**, 321–327 (1976).
63. Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313**, 504–507 (2006).

Figures

Selection of trait stimuli

Selection of face images

Figure 1

Sampling traits (a-d) and face images (e-h) to generate a comprehensive set. a, Sampling of traits began by assembling an extensive list of trait words^{12–15,19,21,25–31,33,38,39} 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 words⁴⁰. c, Three filters

were applied to remove words with similar meanings, unclear meaning, and infrequent usage (see Methods). d, The final set of 100 traits consisted of the sampled adjectives and nouns (see Supplementary Table 1). e, Sampling of face images began by assembling a set of frontal, neutral, white faces from three popular face databases^{41–43}. f, Each face was represented with a vector of 128 facial features that are used to classify individual identities using neural network⁴⁴ pre-trained to identify individuals across millions of faces of all different aspects and races. g, Maximum variation sampling⁴⁵ was applied to select faces with maximum variability in facial structure in this 128-D space. h, Multidimensional scaling visualization of the sampled 100 face images (green and orange dots).

Figure 2

Representativeness of the sampled traits (a-b) and face images (c-d). a, Distributions of word similarities. The similarity between two words was assessed with the cosine distance between the 300-feature vectors⁴⁰ of the two words. The blue histogram plots the pairwise similarities among the 100 sampled traits. The red histogram plots the similarities between each of the freely generated words during spontaneous face judgments ($n = 973$, see Supplementary Fig. 1a) and its closest counterpart in the sampled 100 traits. Dashed lines indicate means. All freely generated words were found to be similar to at least one of the sampled traits (all similarities greater than the mean similarity among the sampled

traits [except for the words “moving” and “round”]). Eighty-five freely generated words were identical to those in the 100 sampled traits. b, Uniform Manifold Approximation and Projection of words (UMAP⁴⁶, a dimensionality reduction technique that generalizes to nonlinearities). Blue dots indicate the 100 sampled traits (examples labeled in blue) and gray dots indicate the freely generated words during spontaneous face judgments (see Methods; non-overlapping examples labeled in gray, which were mostly momentary mental states rather than temporally stable traits). c, UMAP of the final sampled 100 faces (stars) compared with a larger set of frontal, neutral, white faces from various databases^{47–49} (dots, N = 632; see also Supplementary Fig. 1b for comparison with faces in the wild). Each face was represented with 128 facial features as extracted by a state-of-the art deep neural network⁴⁴. d, UMAP of the final sampled 100 faces (stars) compared with the larger set of faces (dots) as in (c). Each face was represented here with 30 automatically measured simple facial metrics⁴² (e.g., pupillary distance, eye size, nose length, cheekbone prominence; see Supplementary Fig. 1c).

dimensions (color code as in upper right scatterplot; see also Supplementary Fig. 4a for factor loadings). Six scatterplots in lower left: each plots the factor loading of all traits 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 Supplementary Fig. 4b for full labels).

Figure 4

Confirmatory analysis with artificial neural network and cross-validation. a, An autoencoder model with one hidden layer to learn the underlying representation of the data. b, The means (points) and standard deviations (bars) of the explained variance on the training data from autoencoders with various number of neurons in the hidden layer (red dots in a). Colors indicate different configurations of activation functions in the encoder and decoder layers (linear, tanh, sigmoid, rectified linear activation unit, L1-norm regularization); for example, the blue line indicates configurations with linear functions in both the encoder and decoder layers (AE-linear-linear). c, Means (points) and standard deviations (bars) of the explained variance on the test data from autoencoders shown in (b).

Figure 5

Robustness of dimensions with respect to the number of traits and participants. a, Pearson’s correlations between factor scores from the full set versus subsets of traits. Colors indicate the four different

dimensions. b, Tucker indices of factor congruence (with orthogonal Procrustes rotation) between the full set versus subsets of participants. Colors indicate the four different dimensions. Error bars indicate standard deviations across the 50 iterations.

Figure 6

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 four factors derived from the samples in 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. Statistical significance was obtained using permutation test (with orthogonal Procrustes rotation, and permuting both the rows and columns of the compared factor loading matrix⁵³ over 1000 iterations): *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. 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.

Figure 7

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. 6). 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplemental.docx](#)