

Modeling first impressions from highly variable facial images

Richard J. W. Vernon, Clare A. M. Sutherland, Andrew W. Young, and Tom Hartley¹

Department of Psychology, University of York, Heslington, York YO10 5DD, United Kingdom

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved July 7, 2014 (received for review May 27, 2014)

First impressions of social traits, such as trustworthiness or dominance, are reliably perceived in faces, and despite their questionable validity they can have considerable real-world consequences. We sought to uncover the information driving such judgments, using an attribute-based approach. Attributes (physical facial features) were objectively measured from feature positions and colors in a database of highly variable “ambient” face photographs, and then used as input for a neural network to model factor dimensions (approachability, youthful-attractiveness, and dominance) thought to underlie social attributions. A linear model based on this approach was able to account for 58% of the variance in raters’ impressions of previously unseen faces, and factor-attribute correlations could be used to rank attributes by their importance to each factor. Reversing this process, neural networks were then used to predict facial attributes and corresponding image properties from specific combinations of factor scores. In this way, the factors driving social trait impressions could be visualized as a series of computer-generated cartoon face-like images, depicting how attributes change along each dimension. This study shows that despite enormous variation in ambient images of faces, a substantial proportion of the variance in first impressions can be accounted for through linear changes in objectively defined features.

face perception | social cognition | person perception | impression formation

A variety of relatively objective assessments can be made upon perceiving an individual’s face. Their age, sex, and often their emotional state can be accurately judged (1). However, inferences are also made about social traits; for example, certain people may appear more trustworthy or dominant than others. These traits can be “read” from a glimpse as brief as 100 ms (2) or less (3), and brain activity appears to track social traits, such as trustworthiness, even when no explicit evaluation is required (4). This finding suggests that trait judgments are first impressions that are made automatically, likely outside of conscious control. Such phenomena link facial first impressions to a wider body of research and theory concerned with interpersonal perception (5).

For many reasons, including the increasingly pervasive use of images of faces in social media, it is important to understand how first impressions arise (6). This is particularly necessary because although first impressions are formed rapidly to faces, they are by no means fleeting in their consequences. Instead, many studies show how facial appearance can affect our behavior, changing the way we interpret social encounters and influencing their outcomes. For example, the same behavior can be interpreted as assertive or unconfident depending on the perceived dominance of an accompanying face (7), and inferences of competence based on facial cues have even been shown to predict election results (8). Nonetheless, support for the validity of these first impressions of faces is inconclusive (9–15), raising the question of why we form them so readily.

One prominent theory, the overgeneralization hypothesis, suggests that trait inferences are overgeneralized responses to underlying cues. For example, a person may be considered to

have other immature characteristics based on a “babyfaced” appearance (16). Exploring this general approach of seeking the factors that might underlie facial first impressions, Oosterhof and Todorov (17) found that a range of trait ratings actually seem to reflect judgments along two near-orthogonal dimensions: trustworthiness (valence) and dominance. The trustworthiness dimension appeared to rely heavily on anger-to-happiness cues, whereas dominance appeared to reflect facial maturity or masculinity. The use of such cues by human perceivers can be very subtle; even supposedly neutral faces can have a structural resemblance to emotional expressions that can guide trait judgments (18).

Oosterhof and Todorov’s (17) findings imply that trait judgments are likely to be based upon both stable (e.g., masculinity) and more transient (e.g., smiling) physical properties (“attributes”) of an individual’s face. However, the trait ratings they analyzed were derived from constrained sets of photographs or computer-generated neutral facial images. Although this approach allows careful control over experimental stimuli, such image sets do not fully reflect the wide range of variability between real faces and images of faces encountered in everyday life.

Jenkins et al. (19) introduced the concept of ambient images to encompass this considerable natural variability. The term “ambient images” refers to images typical of those we see every day. The variability between ambient images includes face-irrelevant differences in angle of view and lighting, as well as the range of expressions, ages, hairstyles, and so forth. Such variability is important: Jenkins et al. (19) and Todorov and Porter

Significance

Understanding how first impressions are formed to faces is a topic of major theoretical and practical interest that has been given added importance through the widespread use of images of faces in social media. We create a quantitative model that can predict first impressions of previously unseen ambient images of faces (photographs reflecting the variability encountered in everyday life) from a linear combination of facial attributes, explaining 58% of the variance in raters’ impressions despite the considerable variability of the photographs. Reversing this process, we then demonstrate that face-like images can be generated that yield predictable social trait impressions in naive raters because they capture key aspects of the systematic variation in the relevant physical features of real faces.

Author contributions: A.W.Y. and T.H. designed research; R.J.W.V. and C.A.M.S. performed research; C.A.M.S. contributed new reagents/analytic tools; R.J.W.V. and C.A.M.S. analyzed data; R.J.W.V., C.A.M.S., A.W.Y., and T.H. wrote the paper; R.J.W.V. contributed to the development of the model and face coding scheme; C.A.M.S. contributed to the design of the validation experiment; A.W.Y. conceived the study; and T.H. conceived the study and contributed to modeling methods.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¹To whom correspondence should be addressed. Email: tom.hartley@york.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1409860111/-DCSupplemental.

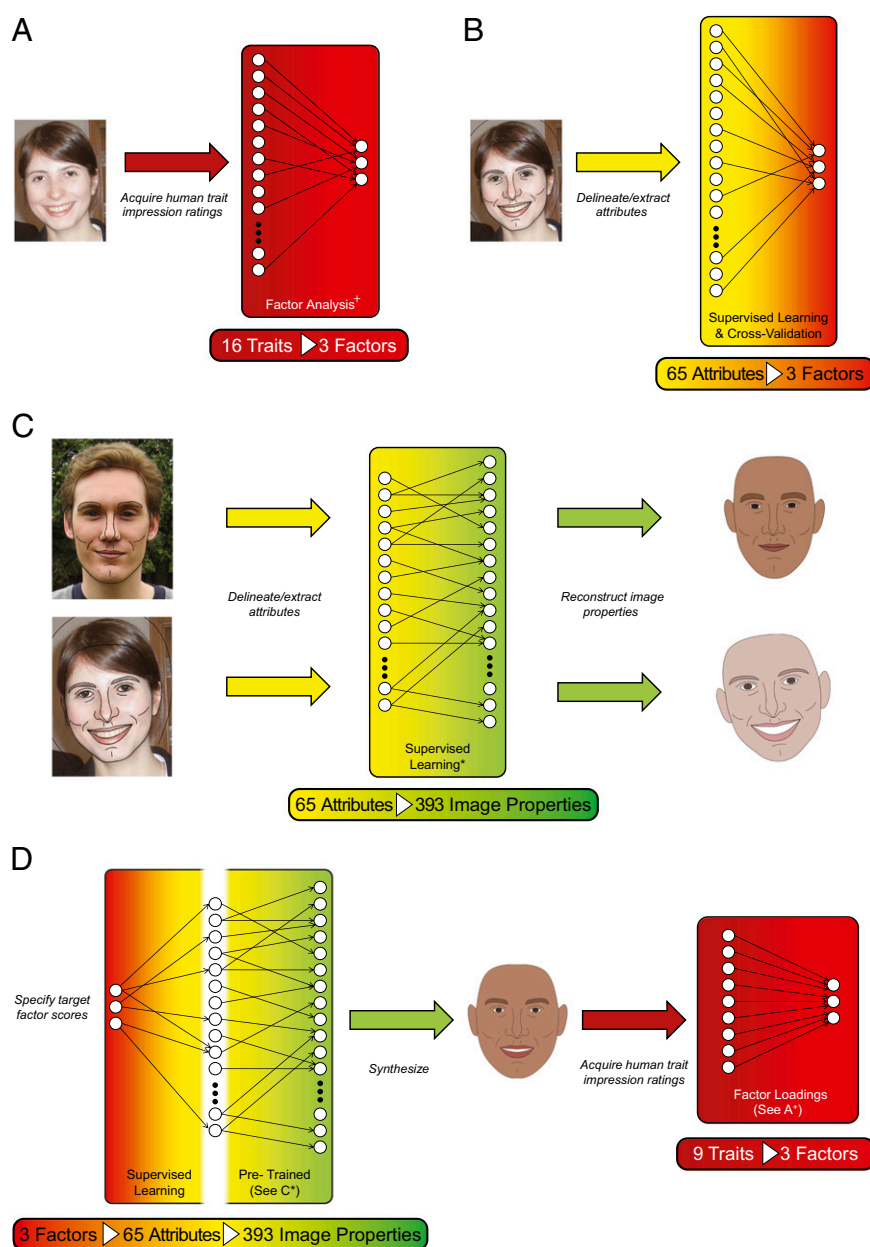


Fig. 1. Summary of methods. (A) For each of 1,000 highly variable face photographs, judgments for first impressions of 16 social traits (“traits”) were acquired from human raters. These 16 trait scores were reduced to scores on three underlying dimensions (“factors”) by means of factor analysis [see Sutherland et al. (21) and *Methods* for further details]. (B) Faces were delineated by the placement of 179 fiducial points outlining the locations of key features. The fiducial points were used to calculate 65 numerical attributes (“attributes”) summarizing local or global shape and color information (e.g., “lower lip curvature, nose area”). Neural networks were trained to predict factor scores based on these 65 attributes (Table S1 describes the fiducial points and attributes in detail). The performance of the trained networks (i.e., the proportion of variance in factor scores that could then be explained by applying the network to untrained images) was evaluated using a 10-fold cross-validation procedure. (C) Using the full set of images a separate network was trained to reproduce 393 geometric properties (e.g., fiducial point coordinates) from the 65 image attributes (pixel colors were recovered directly from the attribute code). This process permits novel faces to be reconstructed accurately on the basis of the attributes illustrating the information available to the model (see Fig. S2 for additional examples). (D) A cascade of networks was used to synthesize cartoon face-like images corresponding to specified factor scores. This process entailed training a new network to translate three factor scores into 65 attributes that could then be used as the input to the network shown in C, and generate the face-like image. The social trait impressions evoked by these synthesized face-like images were then assessed by naive raters using methods equivalent to Sutherland et al. (21).

factor scores based on unseen test cases were only marginally less accurate than the fit obtained for training cases ($r_{\text{approachability}} = 0.92$; $r_{\text{youthful-attractiveness}} = 0.75$; $r_{\text{dominance}} = 0.73$), indicating that the linear model generalizes very well to novel cases.

Because of their greater ability to capture nonlinear and multivariate relationships, we might intuitively have expected nonlinear architectures to outperform linear models. Perhaps

surprisingly, however, we found no additional benefits of a non-linear approach. For example, using our standard training and validation procedures, a network with five nonlinear hidden units generated correlations of $r_{\text{approachability}} = 0.88$, $r_{\text{youthful-attractiveness}} = 0.65$, $r_{\text{dominance}} = 0.62$ (all $P < 0.001$). Furthermore, there were significant negative correlations between the number of hidden units and performance for all three factors ($r_{\text{approachability}} = -0.96$,

$r_{\text{youthful-attractiveness}} = -0.98$, $r_{\text{dominance}} = -0.97$, all $P < 0.001$). It seems that any nonlinear or multivariate relationships that the more complex architectures are able to exploit in fitting the training data, do not generalize. Instead, nonlinear network performance for untrained test cases suffers from overfitting. Importantly then, the critical relationships in the data are largely captured by the simple linear model, which generalizes well to the new cases.

The fact that the linear model works so well allows us to quantify which physical features are correlated with each social trait dimension. Table 1 summarizes statistically significant associations (Pearson correlations surviving Bonferroni correction for multiple comparisons) between physical attributes and factor scores (a full description and numerical key to all 65 attributes is provided in Table S1).

The most striking thing in Table 1 is that almost all attributes we considered are significantly correlated with one or more of the dimensions of evaluation. It is clear that social traits can be signaled through multiple covarying cues, and this is consistent with findings that no particular region of the face is absolutely essential to making reliable social inferences (24).

That said, the substantial roles of certain types of attribute for each dimension also emerge clearly from closer inspection of Table 1. The five features that are most strongly positively correlated with the approachability dimension are all linked to the mouth and mouth shape (feature #25 mouth area, #26 mouth height, #29 mouth width, #30 mouth gap, #32 bottom lip curve), and this is consistent with Oosterhof and Todorov's (17) observation that a smiling expression is a key component of an impression of approachability. Four of the five features that are most strongly positively correlated with the youthful-attractiveness dimension relate to the eyes (feature #11 eye area, #12 iris area, #13 eye height, #14 eye width), in line with Zebrowitz et al.'s (16) views linking relatively large eyes to a youthful appearance. In Oosterhof and Todorov's (17) model the dominance dimension is linked to stereotypically masculine appearance, and here we find it to be most closely correlated with structural features linked to masculine face shape (feature #8 eyebrow height, #35 cheek gradient, #36 eye gradient) and to color and texture differences that may also relate to masculinity (28) or a healthy or tanned overall appearance (feature #49 skin saturation, #62 skin value variation).

Although this agreement between the features, which we found to be most closely linked to each dimension and theoretical approaches to face evaluation, is reassuring, it is nonetheless based on correlations, and correlational data are of course notoriously susceptible to alternative interpretations. We therefore sought to validate our approach with a strong test. The largely linear character of the mapping we have identified implies that it might be possible to reverse-engineer the process, using trait-factor scores as inputs (instead of outputs) to a neural network that will generate 65 predicted features from the input combination of factor scores. From these 65 attributes, the requisite image properties can then be recovered and used to reconstruct a face-like image (Fig. 1). The critical test is then whether the reconstructed image exemplifies the intended social traits. This process provides us with a way to visualize the patterns of physical change that are expected to drive the perception of specific social traits, and to test the validity of these predictions with naive human raters.

We carried out this process in three steps. We first created a linear model, allowing us to generate physical attribute scores (including pixel colors) characteristic of specific combinations of social trait judgments. We then created a linear model relating attribute scores to normalized image coordinates, allowing us to reconstruct face-like images from specified attribute scores (see *Methods* and Fig. 1C for details). We then combined these models to reconstruct faces expected to elicit a range of social

trait judgments along each dimension (see Fig. 3 for examples), and obtained a new set of ratings of these images, which we compared with the model's predictions (Fig. 1D).

In all cases the predicted scores on a given dimension correlated significantly with the new obtained ratings on that dimension (Table 2), showing that the intended manipulation of each dimension was effective. However, it is also evident from Table 2 that the dimensional manipulations were not fully independent from each other, as would be expected from the fact that many features are correlated with more than one dimension evident in Table 1. Nonetheless, for both approachability and dominance, the magnitude of the correlation of ratings with the corresponding predicted trait dimension was significantly greater than the correlation with either of the remaining dimensions, showing clear discriminant validity (all $P < 0.011$) (see *Methods* for details). For the youthful-attractiveness dimension, the magnitude of the correlation between predicted youthful-attractiveness and ratings of youthful-attractiveness was greater than that of its correlation with ratings of approachability ($P < 0.001$), and its correlation with dominance approached statistical significance ($P = 0.081$).

In sum, the generated faces were successful in capturing the key features associated with each dimension. Furthermore, the faces strikingly bear out the conclusions reached from the pattern of feature-to-dimension correlations reported in Table 1. The multiple cues for each dimension are clearly captured in the cartoon faces, and the importance of the mouth to approachability, the eyes to youthful-attractiveness, and the masculine face shape and change in skin tone for dominance are all evident. Increased youthful-attractiveness is also apparently linked to a more "feminized" face shape in Fig. 3. This result was in fact also apparent in Table 1, where the "jaw height" feature (no. 21 in Table 1) was the among the "top 5" positive correlations with youthful-attractiveness (together with the four eye-related features we listed previously), and the other jawline features (22–24) were all in the "top 11."

General Discussion

To our knowledge, we have shown for the first time that trait dimensions underlying facial first impressions can be recovered from hugely variable ambient images based only on objective measures of physical attributes. We were able to quantify the relationships between those physical attributes and each dimension, and we were able to generate new face-like representations that could depict the physical changes associated with each dimension.

The fact that our results are consistent with a number of previous studies based on ratings rather than physical measures suggests that this approach has successfully captured true changes in each dimension. To validate this claim, we first trained a separate model to reconstruct face-like images capturing the relevant featural variation in our ambient image database. We then generated images whose features varied systematically along each dimension and demonstrated that these yield predictable social trait impressions in naive raters.

Critically, our approach has been based entirely on highly variable ambient images. The results did not depend in any way on imposing significant constraints on the images selected, or on any preconceived experimental manipulation of the data. Faces were delineated, and a wide range of attributes chosen/calculated, with no a priori knowledge of how individual faces were judged, or which attributes might be important. This approach minimizes any subjectivity introduced by the researcher, and is as close to a double-blind design as can be achieved. The fact that the findings are based on such a diverse set of images also lends considerable support to both the replicated and novel findings described above. Furthermore, the approach used allowed us explore attribute-factor relationships.

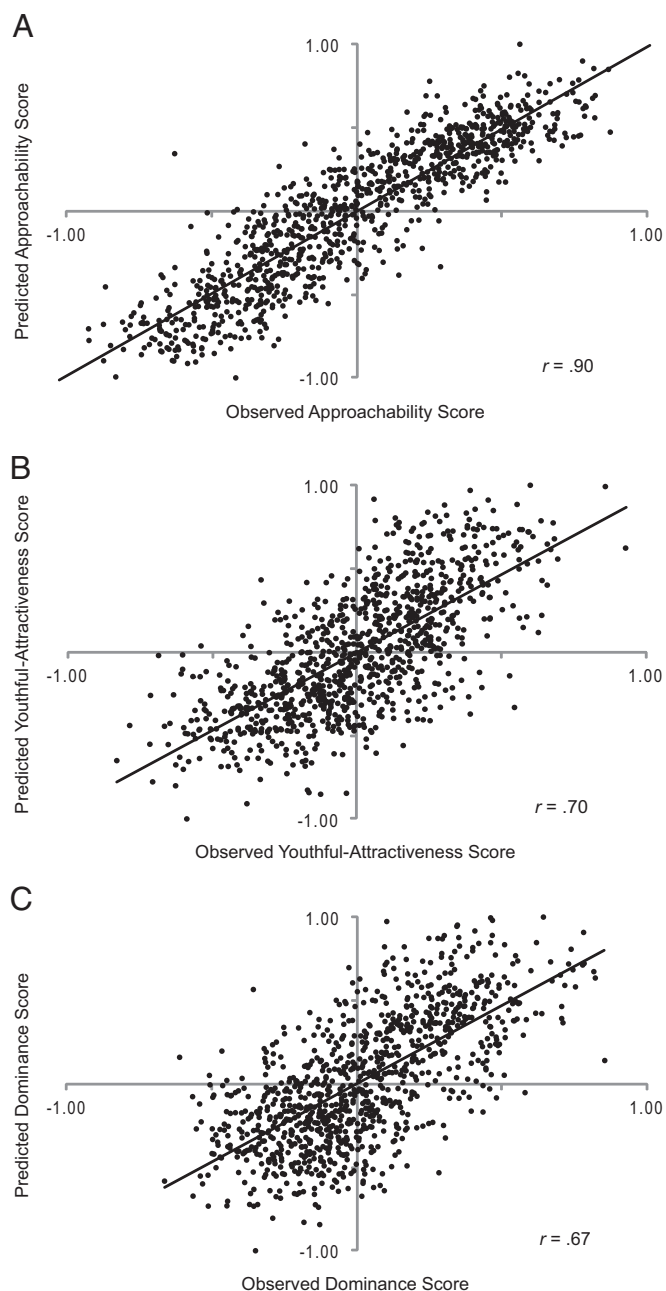


Fig. 2. Scatterplots indicating the correlations between experimentally derived factor scores (from human raters) with the corresponding predictions, for untrained images (see *Methods* for details), derived from a linear neural network (as illustrated in Fig. 1B). Each point ($n = 1,000$, for all axes) represents the observed and predicted ratings for a distinct face image in our database. Both experimental and predicted scores have been scaled into the range $(-1:1)$ (A) approachability, (B) youthful-attractiveness, or (C) dominance.

Oosterhof and Todorov's (17) demonstration that first impressions of many traits can be encompassed within an overarching dimensional model offered an important and elegant simplification of a previously disparate set of studies, and helped demystify how we as humans seem capable of reliably making such an extraordinary range of social inferences. Our findings take this demystification a significant step further by showing that these dimensions of evaluation can be based on simple linear combinations of features. Previous studies with computer-

generated stimuli had also shown that specific linear changes in a morphable computer model can capture such dimensions (17, 29). Achieving an equivalent demonstration here is noteworthy because the features present in the ambient images were not predetermined or manipulated on theoretical grounds and because their highly varied nature will clearly affect the reliability of individual feature-trait associations; unnatural lighting can compromise skin-tone calculations, angle of view affects the perceived size and the retinal shape of many features, and so on. It is therefore particularly impressive that our approach was able to capture the majority of the variance in ratings despite these potential limitations in individual attribute calculations. Part of the reason for this result surely lies in the point emphasized by Bruce (30) and by Burton (31) for the related field of face recognition: that naturally occurring image variations that are usually dismissed as "noise" can actually be a useful pointer that helps in extracting any underlying stability.

A question for future research concerns the extent to which the model's success in accounting for social trait impressions is dependent on the particular selection of attributes we used. Our 65 input features were intended to provide a reasonably complete description of the facial features depicted in each image. The success of these features is demonstrated by our capacity to reconstruct distinctive facsimiles of individual faces based only on the attributes (e.g., Fig. 1C; see Fig. S2 for further examples). In generating these attributes, our approach was to establish numerical scores that would, where possible, relate to a list of characteristics we could label verbally based on the Psychomorph template (Fig. S1 and Table S1). In line with the overarching ambient image approach, our strategy was to be guided by the data as to the role of different attributes, and thus we included the full set in our analyses. Given the linearity we found, though, we broadly expect that any similarly complete description of the relevant facial features should yield similar results. However, it is also likely that our attribute description is overcomplete in the sense that there is substantial redundancy between the attributes. Because of intercorrelations between features, it might well be possible to achieve comparable performance with fewer, orthogonal, components, but these would be much harder to interpret in terms of individual verbally labeled cues.

Although the current model already includes a fairly detailed representation of the geometry of shape cues, the textural information we incorporate is less fine-grained, and an obvious development of our approach will be to increase the resolution of this aspect of the model, which may yield some improvement in its performance and will potentially allow for much more realistic reconstructions. For the present purposes, though, we consider it a strength that a relatively simple representation can capture much of the underlying process of human trait attribution.

Another important issue to which our approach can be addressed in future concerns the extent to which social trait impressions may depend on specific image characteristics, such as lighting and camera position, changeable properties of an individual's face, such as pose and expression, or alternatively, more stable characteristics determined by the face's underlying structure. It is important to note that the former, variable features are potentially specific to each image and therefore even to different images of the same individual. This critical distinction between individual identities and specific photographs has a long history in work on face perception (32, 33), and indeed recent work (19, 20) has demonstrated clearly that intraindividual and image cues play a role in determining social trait judgments alongside any interindividual cues. Our results are consistent with this in demonstrating that changeable features of the face (such as head tilt and bottom lip curvature) covary reliably with social trait impressions, but our approach could also be extended to allow estimation of the relative contributions of these different contributory factors.

Table 1. Significant associations between objective attributes and social trait impressions in 1,000 ambient face photographs

Attribute type		Attribute	App	Yo-At	Dom
Head size and posture	01.	Head area		0.14	
	03.	Head width	0.14	0.18	−0.20
	04.	Orientation (front-profile)		0.12	
	05.	Orientation (up-down)	0.17	0.28	
	06.	Head tilt	0.19	0.20	
Eyebrows	07.	Eyebrow area	−0.16	−0.21	0.23
	08.	Eyebrow height	−0.15	−0.33	0.27
	09.	Eyebrow width		0.22	−0.12
	10.	Eyebrow gradient		0.31	−0.15
Eyes	11.	Eye area	−0.26	0.40	−0.22
	12.	Iris area	−0.20	0.41	−0.31
	13.	Eye height	−0.30	0.39	−0.23
	14.	Eye width	−0.13	0.34	−0.19
Nose	15.	% Iris	−0.31	0.24	
	16.	Nose area	0.26	0.14	
	17.	Nose height		0.24	
	18.	Nose width	0.45		0.16
	19.	Nose curve	0.37		
Jawline	20.	Nose flare	−0.37		
	21.	Jaw height	0.17	0.35	
	22.	Jaw gradient	0.18	0.33	
	23.	Jaw deviation		0.25	0.14
Mouth	24.	Chin curve	0.18	0.31	
	25.	Mouth area	0.69	0.14	−0.15
	26.	Mouth height	0.51	0.15	−0.22
	27.	Top Lip height	−0.24	0.24	−0.25
	28.	Bottom lip height	−0.35	0.34	−0.15
	29.	Mouth width	0.73		
	30.	Mouth gap	0.71		
Other structural features	31.	Top lip curve	0.36	0.12	
	32.	Bottom lip curve	0.75		
	33.	Noseline separation	0.22		
	34.	Cheekbone position	0.16		
	35.	Cheek gradient		−0.17	0.37
Feature positions	36.	Eye gradient	−0.23	−0.21	0.32
	38.	Eyebrows position			−0.27
	39.	Mouth position	0.38	−0.28	
Feature spacing	40.	Nose position		−0.22	
	41.	Eye separation		0.23	−0.21
	42.	Eyes-to-mouth distance	−0.39	0.19	
	43.	Eyes-to-eyebrows distance			−0.44
Color and texture	46.	Mouth-to-chin distance		−0.38	0.13
	47.	Mouth-to-nose distance	−0.60	−0.12	
	49.	Skin saturation			0.28
	50.	Skin value (brightness)	−0.13		−0.23
	51.	Eyebrow hue			
	52.	Eyebrow saturation	0.13		0.15
	53.	Eyebrow value (brightness)		−0.13	−0.22
Other features	55.	Lip saturation	0.12	0.19	
	59.	Iris value (brightness)	−0.24		
	60.	Skin hue variation		−0.21	
	61.	Skin saturation variation		−0.22	0.21
	62.	Skin value variation		−0.24	0.25
	63.	Glasses		−0.26	
	64.	Beard or moustache		−0.20	0.24
	65.	Stubble	−0.15		0.24

App, Approachability; Dom, Dominance; Yo-At, Youthful-attractiveness. Significant attribute-factor correlations, after Bonferroni correction ($P < 0.050/195$). Highly significant results ($P < 0.001/195$) are highlighted in bold. See Table S1 for attribute descriptions.

Our methods provide a means to estimate first-impression dimensions objectively from any face image and to generate face-like images varying on each dimension. These are significant

steps that offer substantial benefits for future research. Our results are also of practical significance (e.g., in the context of social media) because we have shown how images of a given

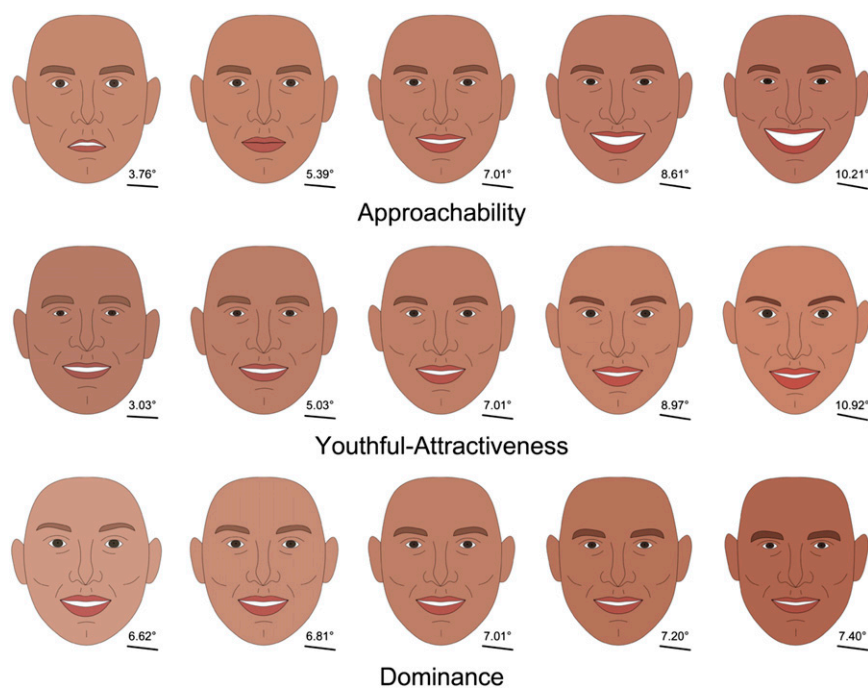


Fig. 3. Synthesized face-like images illustrating the changes in facial features that typify each of the three social trait dimensions. The images were generated using the methods described in the text and in Fig. 1 C and D, based on the same attributes as those used to derive social trait predictions in Fig. 2. The “low” end of each dimension is shown at the left end of each row and the “high” end is at the right. Faces are shown in upright orientation for easy comparison, but the model also suggests some systematic variation in the tilt of the face (indicated below each image). A sample of such synthesized images was used to validate the predicted trait impressions in a group of naive human raters (see Table 2 and the text for details). See [Movies S1–S3](#) for video animations of these changes.

individual can be selected on the basis of quantifiable physical features of the face to selectively convey desirable social trait impressions.

Methods

Social Trait Judgments. Each of the ambient faces had been rated for 16 social trait impressions, as reported in previous work (21–24). Briefly (Fig. 1A), each trait was rated for the full set of 1,000 faces by a minimum of six independent judges using a seven-point Likert scale (for example, attractiveness: 1 = very unattractive, 7 = very attractive). All traits had acceptable interrater reliability (Cronbach’s $\alpha > 0.70$). The means of the different raters’ scores for each face and each trait were subjected to principal axis factor analysis with orthogonal rotation, yielding factor scores (Anderson–Rubin method, to ensure orthogonality) for each of the previously identified dimensions (approachability, youthful-attractiveness, and dominance) (21).

Delineation and Quantification of Facial Features. The shape and internal features of each face were identified by manually placing 179 fiducial points onto each image using PsychoMorph (27). The initial delineation was checked visually by two experimenters, with further checking of the organization of fine-scale features using a custom Matlab script that identified errors that would not be found by visual inspection (e.g., incorrect sequencing of the fiducials). To facilitate the modeling of image shapes, the resulting 2D fiducials were then rotated to a vertical orientation such that the centroids of left- and right-sided points were level.

Sixty-five attributes were derived using these coordinates. These attributes corresponded to a range of structural, configurational, and featural measurements (see [Table S1](#) for full details). For example, “head width” was calculated as the mean horizontal separation between the three leftmost and three rightmost points; “bottom lip curvature” was calculated by fitting a quadratic curve through the points representing the edge of the bottom lip (curvature being indicated by the coefficient of the squared term); “mouth-to-chin distance” was the vertical separation between lowest point on the lower lip and the lowest point on the chin. Area measurements were calculated using polygons derived from subsets of the fiducial points. Overall colors within specified areas were calculated for specified regions (i.e., lips, iris, skin, eyebrows) by averaging RGB values of pixels within polygons defined by sets of fiducial points, converting these to hue, saturation, value

(HSV) measures, and (for skin pixels) additionally calculating a measure of dispersion, entropy, for each of the H, S, and V channels (15 texture attributes in total). A HSV description for color was chosen on the basis that hue and saturation would be relatively insensitive to the large overall luminance variations in the ambient images. Three Boolean attributes described the presence of glasses, facial hair (beards and moustaches), or stubble.

The raw attribute values were separately normalized to place them into a common range necessary for the neural network training. These normalization steps also helped to reduce the impact of image-level nonlinearities in the raw attributes of the highly varied images.

First, a square-root transform was applied to attributes measuring area. The overall scale of all geometric measures was normalized by dividing by the average distance between all possible pairs of points outlining the head (a robust measure of the size of the head in the 2D image). Finally the resulting values were scaled linearly into the range (–1:1).

The HSV color values were similarly scaled into the range (–1:1). As hue is organized circularly, it was necessary to apply a rotation to the raw hue values such that the median, reddish, hues associated with typical Caucasian skin tones were represented by middling values.

Neural Network Training, Validation, and Cross-Validation. Neural networks were implemented using the MatLab Neural Network toolbox (MathWorks). For initial modeling of the determinants of social trait judgments, input units represented physical attributes as described above, with output units representing social trait factor scores. For the two-layer (linear) networks, both input and output units used a linear activation function, and these were connected in a fully feed-forward manner with each input unit being connected to each output unit (Fig. 1B). For the three-layer (potentially non-linear) networks an additional intervening layer of hidden units (with sigmoid activation function) was placed between input and output layers, such that each input unit was connected to each hidden unit, which was in turn connected to each output unit.

During training weights were adjusted using the MatLab toolbox’s default Levenberg–Marquardt algorithm. In essence, the weighted connections between input and output units were gradually and iteratively varied so as to minimize (on a least-squares measure) the discrepancy between the model’s output and the social trait judgments obtained from human raters.

Training, validation, and 10-fold cross-validation was carried out as described above. The 1,000 images were randomly partitioned into 10 discrete

Table 2. Spearman's correlations between expected and rated factor scores for synthesized face-like images

Rated factor scores	Expected factor scores		
	Approachability	Youthful-attractiveness	Dominance
Approachability	0.93**	0.03	−0.09
Youthful-attractiveness	0.78**	0.56*	0.10
Dominance	0.19	−0.53*	0.74**

** $P < 0.001$, * $P < 0.050$.

sets of 100, with 8 sets used to train the network, a further set used to determine when training had converged, and the remaining set used to evaluate the performance of the trained network. The predicted social trait judgment outputs were noted, and the whole process repeated until each of the 10 image sets had served as the test. This entire cross-validation procedure was then repeated 100 times using a new random partitioning of the data to ensure that the results did not depend on a specific partitioning of the data.

Statistical Analysis. The Pearson correlation between the outputs of the model to unseen (i.e., untrained) test case images and the corresponding factor scores provides a measure of how well the model predicts variation in the human ratings. We calculated these correlations for each of the three previously identified social trait dimensions (Fig. 2). We also report the correlations (Bonferroni-corrected) between individual attribute scores and each social trait dimension (Table 1). Note that our analysis suggests that social trait judgments are determined by multiple small contributions from different physical attributes, which means that it is impossible to unambiguously determine the contribution of each attribute (multicollinearity), although the correlations may serve to indicate relationships worthy of further investigation.

Generating Face-Like Images. To address interpretational difficulties arising from multicollinearity, we reversed the modeling process to generate face-like images with attributes expected to convey certain social trait impressions. To solve this engineering problem, we used a cascade of linear networks trained on the full set of 1,000 images. First (Fig. 1C), a network was trained to convert physical attributes (as described above) into the coordinates of fiducial points, which could be rendered using custom MatLab scripts (for this purpose we subdivided the output units, representing the full set of fiducial points (Fig. S1), feature centroids, and global rotation, into 19 sub-networks that were each trained separately for memory efficiency). Then (Fig. 1D), a new network was trained to generate the attributes corresponding to a specified social trait factor scores [each having been scaled into the range (−1:1)], which then acted as input to the face-rendering network. The resulting cascade generates synthetic face-like images that are expected to yield specific social percepts. For example, a face-like image generated using a trait factor score of (1, 0, 0) is expected to be perceived as highly approachable (first dimension) but neutral with respect to youthful-attractiveness (second dimension) and dominance (third dimension).

Validating Synthesized Faces. We tested the validity of the generated face-like images by having new participants assess them in an online questionnaire, closely following the procedure in Sutherland et al. (21). Trait ratings (as a proxy for scores on the three dimensions) were collected for 19 generated faces that covered the range of potential factor scores [each scaled into the range (−1:1)]. One face represented the neutral point (0, 0, 0), 6 faces represented the extremes (high and low) of each dimension [e.g., (0.8, 0, 0), approachable], and 12 faces represented all possible pairs of those extremes

[e.g., (−0.8, 0, 0.8) unapproachable, dominant]. We chose a value of 0.8 for to represent the extreme of each dimension because this was typical of the scaled social trait factor scores of the most extreme 30 of the 1,000 faces in our ambient image database.

To evaluate these images, we solicited social trait judgments from 30 naive participants (15 male, mean age 23.93 y) who took part after consenting to procedures approved by the ethics committee of the Psychology Department, University of York. All spoke fluent English and were from a culturally Western background.

To generate proxy factor scores for each trait dimension identified in our earlier factor analysis, we selected the most heavily loading traits and asked raters to evaluate those traits for each image in the set of synthesized face images. These ratings were then combined in an average, weighted by the trait's loadings on that dimension. For the approachability dimension raters assessed each face for "smiling," "pleasantness," and "approachability." For youthful-attractiveness they rated images for "attractiveness," "health," and "age." For dominance they rated "dominance," "sexual dimorphism," and "confidence."

The participants were randomly allocated into three sex-balanced groups, one per dimension. Each group only rated the three traits making up one dimension, to avoid the risk of judgments on one factor biasing another. Each trait was rated in a block of consecutive trials (with random block order), and within each block the 19 generated faces were randomly presented. Before rating the actual stimuli, the participants saw six practice images (created using random factor scores, indistinguishable from actual stimuli). All faces were rated on a 1–7 Likert scale with endpoints anchored as previous research (21–24).

To assess the correspondence between raters' judgments and the predicted factor scores (i.e., those used to synthesize the faces), we determined the correlation for each pairing of judged and predicted dimensions (Table 2). Independent *t* tests were used to compare these correlations at the level of individual raters. First, we calculated a Spearman's correlation for each pairing of rated and synthesized trait dimensions, then we used independent *t* tests to test the hypothesis that the absolute value of the Fisher transformed correlations was significantly greater for the predicted trait dimension than for each of the other dimensions.

Animations. By generating a series of images with incremental changes along each dimension, we were also able to create short movies to encapsulate each dimension in the model (Movies S1–S3). As well as the changes already noted, these movies show that the neural network has also captured superordinate variation resulting from synchronized changes across combinations of features; for example, increased dominance involves raising the head (as if to "look down" upon the viewer).

ACKNOWLEDGMENTS. The research was funded in part by an Economic and Social Research Council Studentship ES/I900748/1 (to C.A.M.S.).

1. Bruce V, Young AW (2012) *Face Perception* (Psychology Press, London).
2. Willis J, Todorov A (2006) First impressions: Making up your mind after a 100-ms exposure to a face. *Psychol Sci* 17(7):592–598.
3. Todorov A, Pakrashi M, Oosterhof NN (2009) Evaluating faces on trustworthiness after minimal time exposure. *Soc Cogn* 27(6):813–833.
4. Engell AD, Haxby JV, Todorov A (2007) Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *J Cogn Neurosci* 19(9):1508–1519.
5. Fiske ST, Cuddy AJC, Glick P (2007) Universal dimensions of social cognition: Warmth and competence. *Trends Cogn Sci* 11(2):77–83.
6. Perrett DI (2010) *In Your Face: The New Science of Human Attraction* (Palgrave Macmillan, London).
7. Hassin R, Trope Y (2000) Facing faces: Studies on the cognitive aspects of physiognomy. *J Pers Soc Psychol* 78(5):837–852.

8. Todorov A, Mandisodza AN, Goren A, Hall CC (2005) Inferences of competence from faces predict election outcomes. *Science* 308(5728):1623–1626.
9. Snyder M, Tanke ED, Berscheid E (1977) Social perception and interpersonal behaviour: On the self-fulfilling nature of social stereotypes. *J Pers Soc Psychol* 35(9):656–666.
10. Bond JCF, Berry DS, Omar A (1994) The kernel of truth in judgments of deceptiveness. *Basic Appl Soc Psych* 15(4):523–534.
11. Zebrowitz LA, Andreoletti C, Collins MA, Lee SY, Blumenthal J (1998) Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *J Pers Soc Psychol* 75(5):1300–1320.
12. Efferson C, Vogt S (2013) Viewing men's faces does not lead to accurate predictions of trustworthiness. *Sci Rep* 3:1047.
13. Carré JM, McCormick CM (2008) In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proc Biol Sci* 275(1651):2651–2656.

14. Stirrat M, Perrett DI (2010) Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychol Sci* 21(3):349–354.
15. Gómez-Valdés J, et al. (2013) Lack of support for the association between facial shape and aggression: A reappraisal based on a worldwide population genetics perspective. *PLoS ONE* 8(1):e52317.
16. Zebrowitz LA, Fellous JM, Mignault A, Andreoletti C (2003) Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Pers Soc Psychol Rev* 7(3):194–215.
17. Oosterhof NN, Todorov A (2008) The functional basis of face evaluation. *Proc Natl Acad Sci USA* 105(32):11087–11092.
18. Said CP, Sebe N, Todorov A (2009) Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* 9(2):260–264.
19. Jenkins R, White D, Van Montfort X, Burton AM (2011) Variability in photos of the same face. *Cognition* 121(3):313–323.
20. Todorov A, Porter JM (2014) Misleading first impressions: Different for different facial images of the same person. *Psychol Sci* 25(7):1404–1417.
21. Sutherland CAM, et al. (2013) Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* 127(1):105–118.
22. Santos IM, Young AW (2005) Exploring the perception of social characteristics in faces using the isolation effect. *Vis Cogn* 12:213–247.
23. Santos IM, Young AW (2008) Effects of inversion and negation on social inferences from faces. *Perception* 37(7):1061–1078.
24. Santos IM, Young AW (2011) Inferring social attributes from different face regions: Evidence for holistic processing. *Q J Exp Psychol (Hove)* 64(4):751–766.
25. Malpass RS, Kravitz J (1969) Recognition for faces of own and other race. *J Pers Soc Psychol* 13(4):330–334.
26. Meissner CA, Brigham JC (2001) Thirty years of investigating the own-race bias in memory for faces—A meta-analytic review. *Psychol Public Policy Law* 7(1): 3–35.
27. Tiddeman B, Burt M, Perrett D (2001) Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE* 21(5):42–50.
28. Russell R (2009) A sex difference in facial contrast and its exaggeration by cosmetics. *Perception* 38(8):1211–1219.
29. Walker M, Vetter T (2009) Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *J Vis* 9(11):1–13.
30. Bruce V (1994) Stability from variation: The case of face recognition. *Q J Exp Psychol A* 47(1):5–28.
31. Burton AM (2013) Why has research in face recognition progressed so slowly? The importance of variability. *Q J Exp Psychol (Hove)* 66(8):1467–1485.
32. Hay DC, Young AW (1982) The human face. *Normality and Pathology in Cognitive Functions*, ed Ellis AW (Academic, London), pp 173–202.
33. Bruce V, Young A (1986) Understanding face recognition. *Br J Psychol* 77(Pt 3): 305–327.