nature human behaviour

LETTERS

PUBLISHED: 14 NOVEMBER 2016 | VOLUME: 1 | ARTICLE NUMBER: 0001

Statistical learning shapes face evaluation

Ron Dotsch[1,2]*, Ran R. Hassin[3] and Alexander Todorov[4]

The belief in physiognomy—the art of reading character from faces—has been with us for centuries[1–3]. People everywhere infer traits (for example, trustworthiness) from faces, and these inferences predict economic, legal and even voting decisions[2,4]. Research has identified many configurations of facial features that predict specific trait inferences[2,5–14], and detailed computational models of such inferences have recently been developed[5–7,15–17]. However, these configurations do not fully account for trait inferences from faces. Here, we propose a new direction in the study of inferences from faces, inspired by a cognitive-ecological[18–20] and implicit-learning approach[21,22]. Any face can be positioned in a statistical distribution of faces extracted from the environment. We argue that understanding inferences from faces requires consideration of the statistical position of the faces in this learned distribution. Four experiments show that the mere statistical position of faces imbues them with social meaning: faces are evaluated more negatively the more they deviate from a learned central tendency. Our findings open new possibilities for the study of face evaluation, providing a potential model for explaining both individual and cross-cultural variation, as individuals are immersed in varying environments that contain different distributions of facial features.

To examine the hypothesis that the location of a face on a statistical distribution of facial features is used for evaluative inferences, we exposed participants to faces drawn from different distributions that vary in a range of statistical properties (such as the central tendency, dispersion, shape) and asked them to judge novel test faces. Crucially, because the test faces were identical in all conditions, any differences in the judgments between conditions can only be explained by the statistical properties of the faces. To generate the faces, we used a statistical face space model that captures the variance from a large sample of real faces with 130 orthogonal dimensions. Each dimension codes for different feature variance, and its magnitude corresponds to one standard deviation (s.d.) of the same feature variance in the sample of real faces. For instance, one dimension may code primarily for face width, and a face located at 1 s.d. on that dimension corresponds to a face width of 1 s.d. above the mean in the original sample. Thus, any face can be represented as a coordinate in the space, and any coordinate can be visualized as an image. The faces for our experiments always varied on a randomly generated dimension (the target dimension, see Methods and Supplementary Fig. 1) in the face space, with the constraint that it was orthogonal to known dimensions that are correlated with social judgments such as trustworthiness and dominance. Learning -phase stimuli additionally varied randomly in directions orthogonal to this target dimension (see Methods).

In study 1, participants were exposed to 500 faces drawn from a normal distribution (see Supplementary Fig. 2) with either a central tendency of –5 s.d. or a tendency of +5 s.d. on the target dimension in the statistical face space. They then judged a set of 29 new test faces varying from –14 s.d. to +14 s.d., in equal steps of 1 s.d., on the target dimension only. Participants judged each test face on social meaningfulness and trustworthiness, a primary dimension of face evaluation that accounts for more than 60% of the variance of face judgments[5]. Participants also judged the faces on typicality as a manipulation check. We modelled the judgments using several curve-fitting approaches and report here the approach with best fit (local polynomial regression; see Supplementary Information for the similar results yielded by the other approaches). As can be seen in Fig. 1, participants' judgments shifted as a function of learning. Typicality judgments peaked on average at –1.56 s.d. (s.d. = 5.40) in the –5 s.d. condition and at 2.19 s.d. (s.d. = 5.01) in the +5 s.d. condition ($t(61) = 2.86$, $P = 0.006$, Cohen's $d = 0.72$). Likewise, trustworthiness judgments shifted as a function of learning. The judgments peaked on average at 1.13 s.d. (s.d. = 7.44) in the –5 s.d. condition and at 5.03 s.d. (s.d. = 6.26) in the +5 s.d. condition ($t(61) = 2.25$, $P = 0.028$, Cohen's $d = 0.57$). Judgments of social meaningfulness had low internal consistency (Cronbach's $\alpha = 0.39$) and did not meet assumptions for statistical testing (see Methods). These results demonstrate that a face can evoke different evaluations depending on its statistical properties, in this case its location in a distribution, and that the central tendency of such distributions on a priori low social dimensions is not only extracted from exposure to a set of faces[23–25] but also affects face evaluation.

Study 2 extended these findings to judgments of attractiveness, competence, dominance and extroversion (Fig. 2). Because of our interest in general social perception, we analysed the estimated peaks of the five social judgments using multivariate analysis of variance (MANOVA). This analysis indicated that the central tendency manipulation generally affected social judgments in the same way as in study 1 (Wilks' $\Lambda = 0.75$, approximate $F(5, 58) = 3.89$, $P = 0.004$, partial $\eta^2 = 0.06$), although the differences in extroversion and dominance peaks were not significant in univariate tests (respectively, $P = 0.225$ and $P = 0.445$, see Supplementary Information). Participants also judged typicality as a manipulation check. Peak typicality indeed shifted in the direction of the central tendency manipulation (see Fig. 2; $t(62) = 2.56$, $P = 0.013$, Cohen's $d = 0.64$). These results were consistent with our hypothesis that learned statistical properties affect evaluation. The deviating pattern of results for dominance might be because judgments of dominance have a much weaker evaluative component than judgments of attractiveness, trustworthiness, extroversion and competence[5,7]. An alternative explanation is that, despite orthogonalization, the target dimension still contained residual dominance variance, causing the linear pattern of dominance judgments in Fig. 2, and potentially attenuating the predicted effects of statistical learning. Although extroversion followed the same pattern as the other social judgments, it showed the weakest effect. Note that extroversion was always the last social judgment to be made, which may have influenced the strength of the central tendency effect.

[1]Department of Psychology, Utrecht University, PO Box 80.149, Utrecht, 3508 TC, The Netherlands. [2]Behavioural Science Institute, Radboud University, PO Box 9104, 6500 HE, Nijmegen, The Netherlands. [3]Department of Psychology, Hebrew University, Mt. Scopus, Jerusalem 91905, Israel. [4]Department of Psychology, Princeton University, Peretsman-Scully Hall, Princeton, New Jersey 08540 USA. *e-mail: r.dotsch@uu.nl

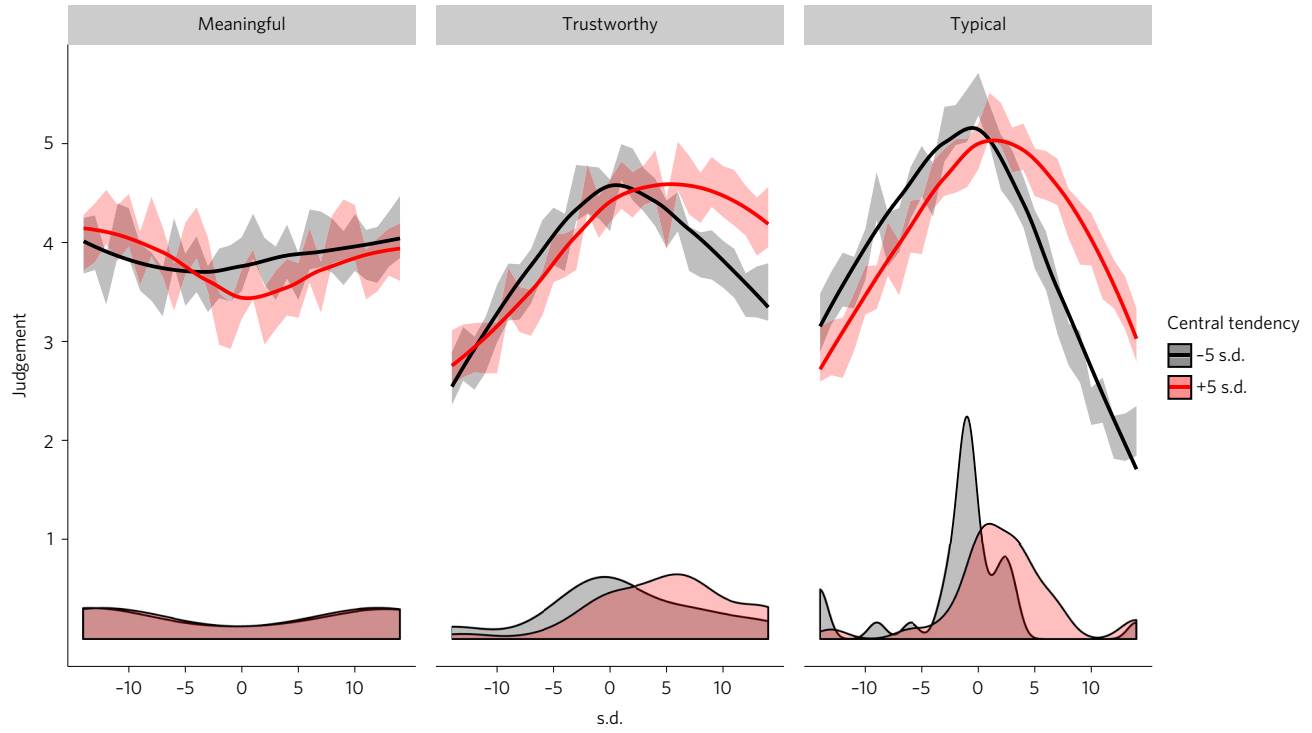

Figure 1 | Judgments study 1. Smoothed average social meaningfulness, trustworthiness and typicality judgments of faces in study 1 as a function of their location on the target dimension (on the *x* axis) and central tendency of the face distribution on that dimension in the learning phase (with unsmoothed between-subjects standard errors). The density plots below show the distribution of the estimated peaks of respective judgment across subjects.

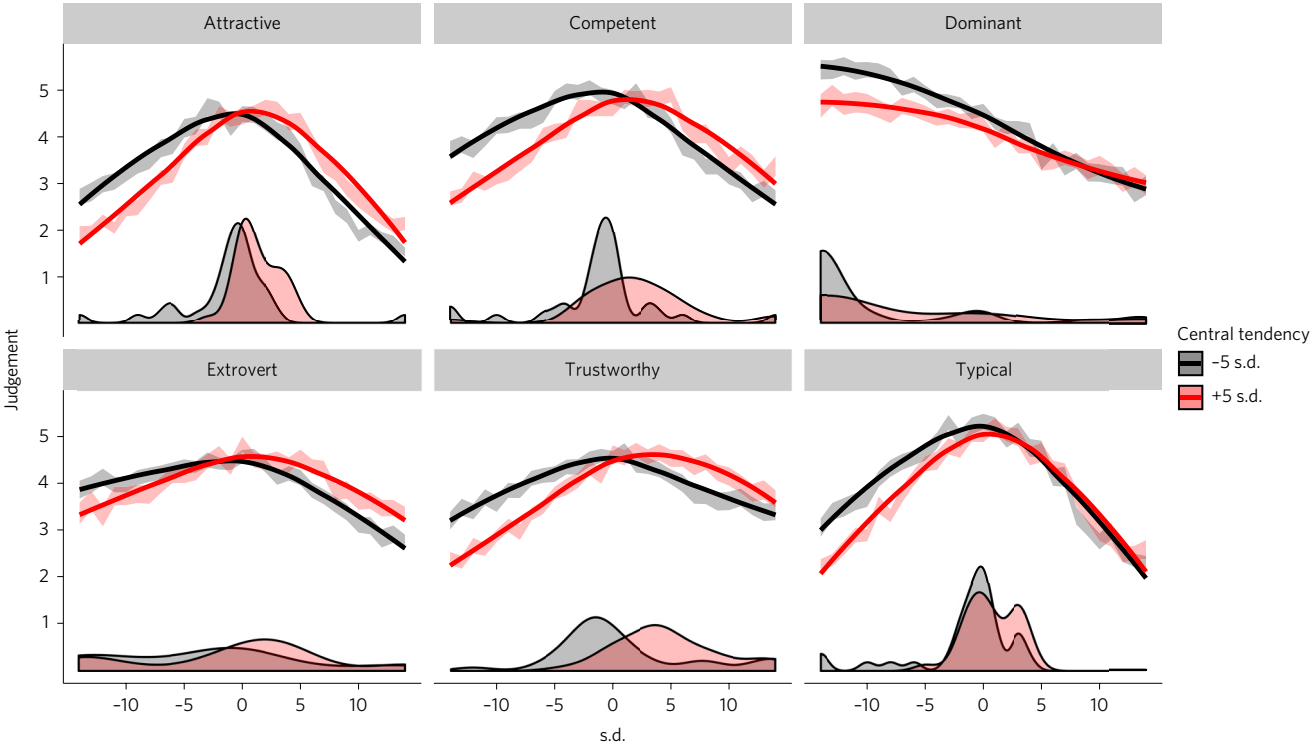

Figure 2 | Judgments study 2. Smoothed average judgments of faces in study 2 as a function of their location on the target dimension (on the *x* axis) and central tendency of the face distribution on that dimension in the learning phase (with unsmoothed between-subjects standard errors). The density plots below show the distribution of estimated peaks of the respective judgment across subjects.

Central tendency is only one determinant of distributions that is complemented by dispersion and shape. The narrower the distribution (less dispersion) on a dimension, the smaller the visual differences in the set of faces sampled from that distribution. This should make it more difficult for people to extract the dimension on which a set of faces varies systematically and should therefore reduce the effect of central tendency on evaluation. We manipulated dispersion and central tendency orthogonally in study 3

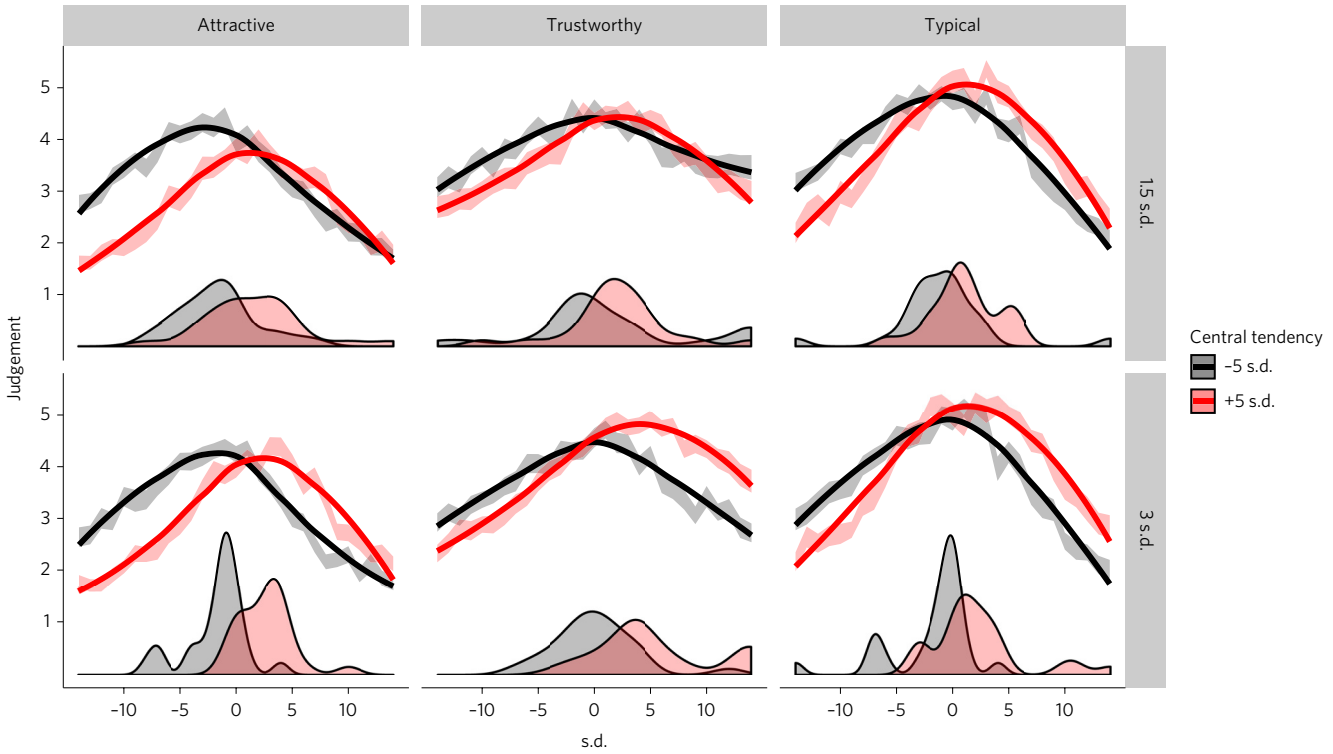

Figure 3 | Judgments study 3. Smoothed average judgments of faces in study 3 as a function of their location on the target dimension (on the *x* axis), central tendency (different colours) and dispersion (different rows) of the face distribution on that dimension in the learning phase (with unsmoothed between-subjects standard errors). The density plots below show the distribution of estimated peaks of the respective judgment across subjects.

(see Supplementary Fig. 2). Replicating the first two studies, a MANOVA on the attractiveness and trustworthiness judgment peaks (see Fig. 3) indicated that both shifted as a function of central tendency in the learning phase (Wilks' $\Lambda = 0.73$, approximate $F(2, 96) = 17.67$, $P < 0.001$, partial $\eta^2 = 0.15$; univariate analyses are reported in the Supplementary Information). As in the previous studies, participants also judged typicality. Peak typicality again shifted in the direction of the central tendency manipulation ($F(1, 98) = 15.94$, $P < 0.001$, partial $\eta^2 = 0.14$). We did not observe any interactions with dispersion for any of the judgments. Thus, even in narrower distributions, central tendency still affected social perception as in distributions with the same dispersion, as in studies 1 and 2, or the difference was too small to be detected with our sample size.

In all prior studies, participants' evaluations were affected by the manipulation of the central tendency of normally distributed faces. In normal distributions, the median face is also the modal (most frequent) face. Thus, our manipulation of central tendency so far has been driven by two factors: distance to the centre of the learned distribution and frequency at that position in the distribution. To disentangle these factors, in study 4 we manipulated—in addition to central tendency—the shape of the distribution to be either normal (as before), quasi-uniform (in which frequency is the same at each point of the distribution, and thus only distance to the centre is informative) and U-shaped (in which the most distant faces are the most frequent) (see Supplementary Fig. 2). If the effect of central tendency is frequency-dependent, it should be present when participants are exposed to a normal distribution, absent or reduced when exposed to a uniform distribution, and flipped when exposed to a U-shaped distribution (where the most frequent faces are also the most distant faces, producing the most extreme dissociation possible). If, on the other hand, the effect of central tendency is dependent on the distance to the centre of the distribution, it should be present when exposed to any of the three distribution shapes. A MANOVA on the attractiveness and trustworthiness judgment peaks (Fig. 4) indicated that both shifted as a function of central tendency in the learning phase, irrespective of its shape (Wilks' $\Lambda = 0.80$, approximate $F(2, 99) = 12.38$, $P < 0.001$, partial $\eta^2 = 0.11$; see the Supplementary Information for follow-up univariate analyses). As in the previous studies, participants also judged typicality. Peak typicality again shifted in the direction of the central tendency manipulation ($F(1, 101) = 7.31$, $P = 0.008$, partial $\eta^2 = 0.07$). We observed no interaction between central tendency and distribution shape for any of the judgments. These results indicate that distance to the centre of the distribution is sufficiently informative for evaluation based on statistical properties and that, at least in the current design, other statistical properties of the distributions do not exert a detectable effect.

In four studies, we have shown that the location of a face in a learned distribution of facial features affects how it is evaluated: the closer the face to the central tendency, the more positively it is evaluated. We have shown that these distributions can be learned in a relatively short time (in only 500 exposures to objects sampled from those distributions); that they affect various judgments, perhaps to the extent that they contain an evaluative component; that the effects of feature distributions on evaluation are mostly dispersion invariant; and finally, that distance to the central tendency (and not frequency) is sufficiently informative for evaluation on the basis of statistical properties. This last finding contrasts our statistical learning account with a mere exposure account[26], according to which the frequency of exposure to a stimulus predicts its evaluation. Nor can our findings be explained by a perceptual adaptation account[27], given that to compute distance to the central tendency for any test face, more information about the environment is required than just the most recent or most frequent faces. The findings are more consistent with a perceptual fluency account[28], which posits that prototypical stimuli are judged more positively because they are easier to process. Indeed, research has established a correlation between face typicality and trustworthiness judgments[29]. Here, we

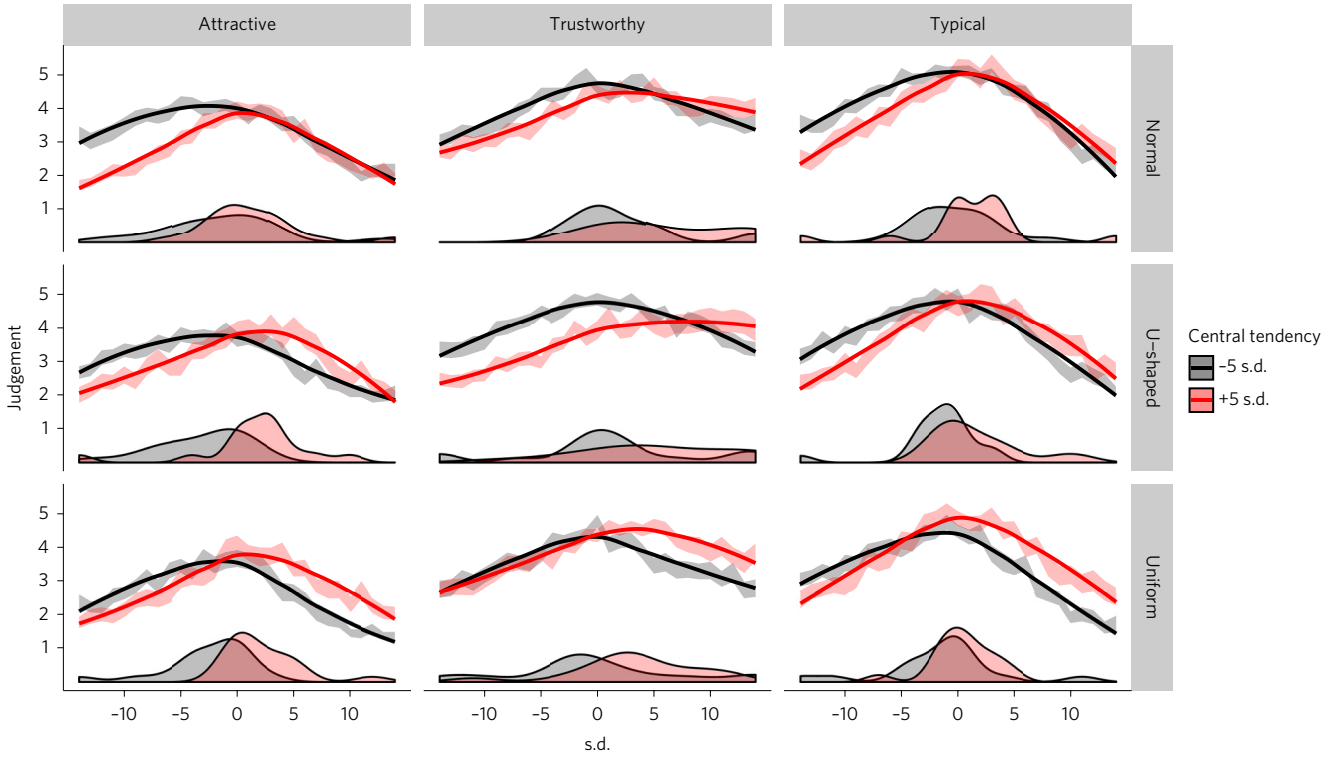

Figure 4 | Judgments study 4. Smoothed average judgments of faces in study 4 as a function of their location on the target dimension (on the *x* axis), central tendency (different colours) and distribution shape (different rows) of the face distribution on that dimension in the learning phase (with unsmoothed between-subject standard errors). The density plots below show the distribution of estimated peaks of the respective judgment across subjects

provide experimental evidence for the effect of learned statistical properties on evaluative inferences from faces. As such, our findings fit well with cognitive–ecological ideas[18–20], in that people's inferences reflect the structure of information (such as distributions of faces) in the environment.

We have initiated a new research direction on inference from faces. By exposing individuals to controlled environments of faces sampled from specific face distributions, we were able to control the short-term learning history of participants, allowing us to experimentally test the hypothesis that social inferences from faces are shaped by learning the statistical structure of one's environment. Our approach and findings provide a potential mechanism for the development of social face perception; an explanation for individual and cultural variation in judgments, given the presence of variance in environments; and the means to experimentally test these explanations.

Methods

All face stimuli were generated using FaceGen 3.1, a multidimensional computer-generated face space with 50 symmetrical shape dimensions (which mirror the effects on both hemispheres of the face, for example, both eyes move upwards), 30 asymmetrical shape dimensions (which do not mirror the effects on both hemispheres, for example, moves the left eye upwards and the right eye downwards) and 50 symmetrical texture dimensions. Randomization of the face coordinates was performed in Python using SciPy's stats module. Orthogonalization of dimensions was performed in MATLAB. Experiments were programmed in Python using the PsychoPy module. Analyses were performed in R version 3.3.1.

The protocol for these studies was approved by the Institutional Review Board For Human Subjects of Princeton University (Protocol no. 5572). Informed consent was obtained from all participants.

Study 1. *Design and participants.* This study followed a one-factor (central tendency: −5 s.d. versus +5 s.d.) between-subjects design, with $n = 32$ in the −5 s.d. condition and $n = 31$ in the +5 s.d. condition. We prespecified the sample size to be at least 60 participants on the basis of time and resource constraints, without the use of power analysis, because no prior estimates of effect size existed for this phenomenon. (The sample sizes for the subsequent studies were based on the observed effect size in study 1.) Students from Princeton University participated for course credit ($n = 63$: 26 men, 35 women, 2 unknown; mean age (M_{age}) = 20.75, s.d. = 2.04).

Target dimension. Nine symmetrical shape dimensions were constructed using a procedure that minimizes social information in faces[25]. First, 100 symmetrical shape dimensions were randomly generated from a normal distribution. These were orthogonalized to ten previously identified social dimensions (for example, trustworthy, dominant, competent, attractive, threatening, extravert[5,7]), such that any change along the random dimension did not result in change along any of the social dimensions. Although this will hold true mathematically in the face space, psychologically, the candidate dimensions may still contain residual social variance, to the extent that the social dimensions do not fully capture the social judgments. To maximize variance in the final set of dimensions, we selected nine candidate dimensions that correlated only weakly with each other (all $|r| < 0.15$) and normalized those.

For each candidate dimension, we generated 25 faces derived from that dimension ranging from −12 s.d. to +12 s.d. (with equally spaced intervals of 1 s.d.). In a pilot study, a sample of 16 female and 14 male Princeton University students ($M_{age} = 20.57$, s.d. = 2.53) judged these faces on trustworthiness, among other judgments (see Supplementary Information), using 7-point scales. The faces were presented in random order, blocked by judgment. Inter-rater agreement of trustworthiness was high (Cronbach's $\alpha = 0.88$). The average judgments are shown in Supplementary Fig. 3. On the basis of these judgments, we selected the candidate dimension with the lowest trustworthiness variance as the to-be-learned target dimension (see Supplementary Fig. 1). The results of this pilot study have been partly reported before[2].

Stimuli. For the learning phase, we generated 500 faces per experimental condition. To mimic complex real-world environments, we generated a set of three-dimensional faces differing not only on the target dimension but also in random ways orthogonal to the target dimension. Thus, the final stimuli varied as a function of scores on the target symmetrical shape dimension, a random symmetrical shape vector that was unique to each stimulus and orthogonal to the target dimension, a random asymmetrical shape vector, and a random texture vector. The scores on the target dimension followed a normal distribution with $\mu = 0$ s.d. and $\sigma = 3$ s.d. (see Supplementary Fig. 2b, left distribution). These were translated to have means of either −5 s.d. or +5 s.d., varying between subjects.

We additionally created silhouette images for the learning phase. Each face was rotated 90° around its vertical axis to create a side view. We then replaced all colour

information with uniform grey to generate silhouette images. A random subset of these images was used during the learning task (see 'Procedure' section).

For the test phase, we generated 29 faces ranging from −14 s.d. to +14 s.d. on the target dimension, with equal spaced intervals of 1 s.d. In contrast to the stimuli in the learning phase, test phase stimuli varied only as function of their position on the target dimension, affecting shape only (see Supplementary Fig. 1 for examples).

Procedure. Participants were assigned to a central tendency condition (−5 s.d. versus +5 s.d.) by the computer script on the basis of their participant number (in study 1: even versus odd). The experimenter was blind to condition. In the learning phase, participants viewed the 500 faces generated for their respective condition. To increase motivation to pay attention to the faces, participants were asked on random trials (one in five) to indicate whether a side-view silhouette matched the previously presented face. Half of the time, the silhouette matched the previously presented face. Participants were told that this task tested how well they recognized a face from its side view when they first saw that face in a frontal view. However, these responses were not analysed.

In the subsequent test phase, participants judged the 29 test faces on a 7-point scale in three separate blocks in which all faces were presented once in random order. In the three fixed-order blocks, participants judged social meaningfulness ("to what extent does the face convey any information about the person's personality?") and trustworthiness as social judgments and typicality as a manipulation check, respectively.

Finally, after the experimental tasks, participants completed an exit questionnaire asking for gender, age, eyesight (colour-blindness and normal vision), subjective ratings of performance on the experimental task and general face recognition ability, and then took part in a funnelled debriefing. Afterwards, participants were debriefed and thanked by the experimenter.

Statistics. We report two-sided tests. Before performing *t*-tests on the estimated judgment peaks, we visually checked for outliers and examined assumptions. Because we had no prespecified way of dealing with outliers, we report analyses with all data included but note here any substantial consequence of excluding outliers above and below 1.5 times the interquartile range within the respective experimental cell. In study 1, we detected one outlier for trustworthiness judgments and nine outliers for typicality judgments. Removing these outliers did not affect conclusions and substantially increased the magnitude of the central tendency effect on both judgments, respectively ($t(60) = 2.72$, $P = 0.009$, Cohen's $d = 0.69$ for trustworthiness and $t(52) = 3.20$, $P = 0.002$, Cohen's $d = 0.87$ for typicality).

All data met the assumption of homogeneity of variance, assessed with Levene's test. As can be seen from the density plots in Fig. 1, there was extreme deviation from normality for estimated social meaningfulness peaks, which also suffered from low reliability and were therefore not further analysed. Although estimated trustworthiness and typicality peaks did not strongly deviate from normality upon visual inspection, Shapiro–Wilks tests indicated significant deviation from normality of typicality peaks in both central tendency conditions. Nevertheless, a non-parametric test yielded the same results as the parametric *t*-test reported above ($W = 229$, $P < 0.001$).

Study 2. *Design and participants.* This study followed a one-factor (central tendency: −5 s.d. versus +5 s.d.) between-subjects design, with $n = 33$ in the −5 s.d. condition and $n = 31$ in the +5 s.d. condition. We prespecified sample size to be at least 60. Because we increased the number of dependent variables to five social judgments in study 2, power analysis for MANOVA showed that we would have needed a sample size of $n = 54$ to detect a central tendency effect of $d \geq 0.5$ (in study 1, d was 0.57 for trustworthiness judgments) with 80% power, assuming correlations of 0.25 between dependent variables. Princeton University students participated for course credit ($n = 66$: 19 men, 45 women, 1 female-bodied transgender, 1 unknown; $M_{age} = 20.29$, s.d. = 2.94). Two participants did not complete all trials and were excluded from data analysis.

Stimuli and procedure. Study 2 used the same stimuli as study 1. The procedure of study 2 was identical to that of study 1, with the exception of the testing phase. In the testing phase, participants judged the 29 test faces on a 7-point scale twice in each judgment block to be able to compute subject-level consistency. There were six judgment blocks, in the following fixed order: the five social judgments trustworthy, dominant, attractive, competent and extrovert; then typical as a manipulation check.

Statistics. Before performing MANOVA on the estimated judgment peaks, we checked assumptions and outliers. As in study 1, because we had no prespecified way of dealing with outliers, we report analyses with all data included but note here any substantial consequence of excluding outliers based on the same criteria as in study 1. We marked 34 data points (8.85% of all estimated peaks) as univariate outliers across all six judgments. Removing univariate outliers did not affect conclusions for social judgments and substantially increased the magnitude of the central tendency effect on social judgments (Wilks' $\Lambda = 0.57$, approximate $F(5, 40) = 5.97$, $P < 0.001$, partial $\eta^2 = 0.11$). However, the effect of the central tendency manipulation on estimated peak typicality was no longer significant ($t(56) = 1.52$, $P = 0.13$, Cohen's $d = 0.40$). We detected 29 participants who could be considered multivariate outliers (Mahalanobis distance >18.54). Nevertheless, even when excluding those, central tendency still affected social judgments (Wilks' $\Lambda = 0.66$, approximate $F(5, 29) = 3.03$, $P = 0.025$, partial $\eta^2 = 0.08$), as well as typicality judgments ($t(33) = 3.03$, $P = 0.005$, Cohen's $d = 1.61$).

All data met the assumption of homogeneity of variance as assessed with Levene's test. As can be seen from the density plots in Fig. 2, there was deviation from normality for estimated dominance and extroversion peaks. Although the other judgment peaks did not strongly deviate from normality upon visual inspection, Shapiro–Wilks tests indicated significant deviation from normality of several judgments peaks in the −5 s.d. condition. Nevertheless, deviation from normality was attenuated for all judgments except dominance in the analysis without univariate outliers reported above, which did not affect conclusions. Moreover, robustness to non-normality in MANOVA has been shown[30,31] for samples with overall $n \geq 40$.

Study 3. *Design and participants.* This study followed a 2 (central tendency: −5 s.d. versus +5 s.d.) × 2 (dispersion: 1.5 s.d. versus 3 s.d.) between-subjects design. We prespecified the sample size to be at least 100. Power analysis for MANOVA with two dependent variables shows we would have needed a sample size of $n = 44$ to detect a central tendency effect of at least the same size as observed in study 2 with 80% power. A minimum sample size of $n = 100$ gave us 80% power to detect a medium-sized interaction effect between central tendency and dispersion ($f^2 = 0.06$). Princeton University students participated for course credit ($n = 102$: 34 male, 67 female, 1 unknown; $M_{age} = 20.22$, s.d. = 2.28).

Stimuli. Study 3 used the same target dimension as studies 1 and 2. For the learning phase, we generated four sets of 500 stimuli to match the cells of the design. We used the same procedure to generate stimuli as in studies 1 and 2 but varied σ (1.5 s.d. versus 3 s.d.) of the normal distribution from which the stimuli's positions on the target dimension were sampled (see Supplementary Fig. 2, panel a).

The test phase stimuli were the same as in studies 1 and 2.

Procedure. The procedure in study 3 was identical to that in study 2, except for the following. The testing phase comprised three judgment blocks: trustworthy, attractive and typical. Trustworthy and attractive were counterbalanced in order. The typicality judgment block, which was our manipulation check, was always last.

Statistics. Before performing MANOVA on the estimated judgment peaks, we checked assumptions and outliers. As in the previous studies, because we had no prespecified way of dealing with outliers, we report analyses with all data included but note here any substantial consequence of excluding outliers based on the same criteria as in the previous studies. We marked 28 data points (9.15% of all estimated peaks) as univariate outliers across all three judgments. Removing univariate outliers did not affect conclusions for social judgments but substantially increased the magnitude of the central tendency effect on social judgments (Wilks' $\Lambda = 0.58$, approximate $F(2, 80) = 28.63$, $P < 0.001$, partial $\eta^2 = 0.24$) and typicality ($F(1, 86) = 18.88$, $P < 0.001$, partial $\eta^2 = 0.18$). We detected 22 participants who could be considered multivariate outliers (Mahalanobis distance >10.58). Nevertheless, even when excluding those, central tendency still affected social judgments (Wilks' $\Lambda = 0.62$, approximate $F(2, 63) = 19.02$, $P < 0.001$, partial $\eta^2 = 0.21$), as well as typicality judgments ($F(1, 65) = 11.79$, $P = 0.001$, partial $\eta^2 = 0.14$). We also observed a main effect for dispersion condition on social judgments, which we do not discuss any further because the effect did not interact with the central tendency manipulation.

All data met the assumption of homogeneity of variance, assessed with Levene's test. As can be seen from the density plots in Fig. 3, the data were distributed mostly normally. However, Shapiro–Wilks tests indicated significant deviation from normality for some conditions, mostly for estimated typicality peaks. Nevertheless, deviation from normality was somewhat attenuated for all judgments in the analysis without univariate outliers reported above, which did not affect conclusions. Moreover, robustness to non-normality in MANOVA has been shown[30,31] for samples with overall $n \geq 40$.

Study 4. *Design and participants.* This study followed a 2 (central tendency: −5 s.d. versus +5 s.d.) × 3 (distribution shape: normal versus U-shaped versus uniform) between-subjects design. We prespecified the sample size to be at least 100. Power analysis for MANOVA with two dependent variables shows that we would have needed a sample size of $n = 44$ to detect a central tendency effect at least the same size as that observed in study 2 with 80% power (or $n = 30$ for the same effect size as in study 3). A minimum sample size of $n = 100$ gave us 80% power to detect a medium-sized interaction effect between central tendency and distribution shape ($f^2 = 0.06$). Princeton University students participated for course credit ($n = 108$: 43 men, 63 women, 1 genderqueer-born male, 1 unknown; $M_{age} = 20.07$, s.d. = 3.40). One participant did not complete all trials and was excluded from data analysis.

Stimuli. Study 4 used the same target dimension as in studies 1–3. For the learning phase, we generated six sets of 500 stimuli to match the cells of the design. We used the same procedure to generate stimuli as in studies 1–3 but varied the distribution shape from which the stimuli's positions on the target dimension were sampled (see Supplementary Fig. 2, panel b). For the normal distribution, we used the same stimuli as in study 1 and 2, with $\mu = 0$ s.d. and $\sigma = 3$ s.d.

For the U-shaped distribution, we first generated a sample from a normal distribution with $\sigma = 3$ s.d., split that distribution across the median, appended the left tail of the distribution (the half that was smaller than the median) to the right of the right tail of the distribution (the half that was greater than the median) and centred the resulting distribution around 0.

For the (quasi-)uniform distribution, we simply sampled from a uniform distribution within the same range of scores that was spanned by the normal and U-shaped distributions.

To create the final learning phase stimuli for the six conditions, the scores of all these distributions were translated to have means of either +5 s.d. or −5 s.d.

The test phase stimuli were the same as in studies 1–3.

Procedure. The procedure for study 4 was identical to that of study 3. Note that chronologically, study 4 preceded study 3.

Statistics. Before performing MANOVA on the estimated judgment peaks, we checked assumptions and outliers. As in the previous studies, because we had no prespecified way of dealing with outliers, we report analyses with all data included but note here any substantial consequence of excluding outliers on the basis of the same criteria as in the previous studies. We marked 26 data points (8.10% of all estimated peaks, equally distributed across conditions) as univariate outliers across all three judgments. Removing univariate outliers did not affect conclusions for social judgments but substantially increased the magnitude of the central tendency effect on social judgments (Wilks' $\Lambda = 0.64$, approximate $F(2, 84) = 23.61$, $P < 0.001$, partial $\eta^2 = 0.20$) and typicality ($F(1, 92) = 13.38$, $P < 0.001$, partial $\eta^2 = 0.13$). We detected 18 participants who could be considered multivariate outliers (Mahalanobis distance >9.66). Nevertheless, even when excluding those, central tendency still affected social judgments (Wilks' $\Lambda = 0.64$, approximate $F(2, 69) = 19.83$, $P < 0.001$, partial $\eta^2 = 0.20$) and typicality judgments ($F(1, 74) = 15.09$, $P < 0.001$, partial $\eta^2 = 0.11$).

All data met the assumption of homogeneity of variance, assessed with Levene's test. As can be seen from the density plots in Fig. 4, the data were distributed mostly normally. However, Shapiro–Wilks tests indicated significant deviation from normality for some conditions, mostly for estimated typicality peaks. Nevertheless, there was almost no deviation from normality for all judgments in the analysis without univariate outliers reported above, which did not affect conclusions. Moreover, robustness to non-normality in MANOVA has been shown[30,31] for samples with overall $N \geq 40$.

Code availability. R code for data processing, analysis and visualization is publicly available in figshare with identifier https://dx.doi.org/10.6084/m9.figshare.3563472 (ref. 32). Python code for generating stimuli and running the experiments is available from R.D.

Data availability. All data that support the findings of this study are publicly available in figshare with identifier https://dx.doi.org/10.6084/m9.figshare.3563472 (ref. 32). The stimuli used in this study are publicly available in figshare with identifier https://dx.doi.org/10.6084/m9.figshare.3563514 (ref. 33).

Received 6 April 2016; accepted 6 September 2016; published 14 November 2016

References

1. Hassin, R. & Trope, Y. Facing faces: studies on the cognitive aspects of physiognomy. *J. Pers. Soc. Psychol.* **78,** 837–852 (2000).
2. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Ann. Rev. Psychol.* **66,** 519–545 (2015).
3. Zebrowitz, L. A. *Reading Faces: Window to the Soul?* (Westview, 1999).
4. Olivola, C. Y., Funk, F. & Todorov, A. Social attributions from faces bias human choices. *Trends Cogn. Sci.* **18,** 566–570 (2014).
5. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl Acad. Sci. USA* **105,** 11087–11092 (2008).
6. Dotsch, R. & Todorov, A. Reverse correlating social face perception. *Soc. Psychol. Pers. Sci.* **3,** 562–571 (2012).
7. Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N. & Falvello, V. B. Validation of data-driven computational models of social perception of faces. *Emotion* **13,** 724–738 (2013).
8. Said, C., Sebe, N. & Todorov, A. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* **9,** 260–264 (2009).
9. Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M. & Young, A. W. Social inferences from faces: ambient images generate a three-dimensional model. *Cognition* **127,** 105–118 (2013).
10. Zebrowitz, L. A. & Montepare, J. M. Social psychological face perception: why appearance matters. *Soc. Personal. Psychol. Compass* **2,** 1497–1517 (2008).
11. Boothroyd, L. G., Jones, B. C., Burt, D. M. & Perrett, D. I. Partner characteristics associated with masculinity, health and maturity in male faces. *Pers. Individ. Dif.* **43,** 1161–1173 (2007).
12. Montepare, J. M. & Zebrowitz, L. A. Person perception comes of age: the salience and significance of age in social judgments. *Adv. Exp. Soc. Psychol.* **30,** 93–161 (1998).
13. Secord, P. F., Dukes, W. F. & Bevan, W. Personalities in faces. I. An experiment in social perceiving. *Genet. Psychol. Monogr.* **49,** 231–279 (1954).
14. Zebrowitz, L. A. in *Handbook of Face Perception* (eds Calder, A. *et al.*) 31–50 (Oxford Univ. Press, 2011).
15. Todorov, A. & Oosterhof, N. N. Modeling social perception of faces. *IEEE Signal Process. Mag.* **28,** 117–122 (2011).
16. Walker, M. & Vetter, T. Portraits made to measure: manipulating social judgments about individuals with a statistical face model. *J. Vis.* **9,** 1–13 (2009).
17. Walker, M. & Vetter, T. Changing the personality of a face: perceived big two and big five personality factors modeled in real photographs. *J. Pers. Soc. Psychol.* **110,** 609-624 (2016).
18. Fiedler, K. & Wänke, M. The cognitive-ecological approach to rationality in social psychology. *Soc. Cogn.* **27,** 699–732 (2009).
19. De Houwer, J., Gawronski, B. & Barnes-Holmes, D. A functional-cognitive framework for attitude research. *Eur. Rev. Soc. Psychol.* **24,** 252–287 (2013).
20. Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M. & Danner, D. Why positive information is processed faster: the density hypothesis. *J. Pers. Soc. Psychol.* **95,** 36–49 (2008).
21. Reber, A. S. Implicit learning and tacit knowledge. *J. Exp. Psychol. Gen.* **118,** 219–235 (1989).
22. Gordon, P. C. & Holyoak, K. J. Implicit learning and generalization of the "mere exposure" effect. *J. Pers. Soc. Psychol.* **45,** 492–500 (1983).
23. Bruce, V., Doyle, T., Dench, N. & Burton, M. Remembering facial configurations. *Cognition* **38,** 109–144 (1991).
24. Cabeza, R., Bruce, V., Kato, T. & Oda, M. The prototype effect in face recognition: extension and limits. *Mem. Cognit.* **27,** 139–151 (1999).
25. Said, C. P., Dotsch, R. & Todorov, A. The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia* **48,** 3596–3605 (2010).
26. Zajonc, R. B. Feeling and thinking: preferences need no inferences. *Am. Psychol.* **35,** 151–175 (1980).
27. Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G. & Nakayama, K. Fitting the mind to the world: face adaptation and attractiveness aftereffects. *Psychol. Sci.* **14,** 558–566 (2003).
28. Reber, R., Schwarz, N. & Winkielman, P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Pers. Soc. Psychol. Rev.* **8,** 364–382 (2004).
29. Sofer, C., Dotsch, R., Wigboldus, D. H. J. & Todorov, A. What is typical is good: the influence of face typicality on perceived trustworthiness. *Psychol. Sci.* **26,** 39–47 (2015).
30. Tabachnick, B. G. & Fidell, L. S. *Using Multivariate Statistics* 6th edn (Pearson, 2013).
31. Seo, T., Kanda, T. & Fujikoshi, Y. The effects of nonnormality on tests for dimensionality in canonical correlation and MANOVA models. *J. Multivar. Anal.* **52,** 325–337 (1995).
32. Dotsch, R., Hassin, R. & Todorov, A. Statistical learning shapes face evaluation: raw data, processed data, and analysis R code. *figshare* https://dx.doi.org/10.6084/m9.figshare.3563472.v1 (2016).
33. Dotsch, R., Hassin, R. & Todorov, A. Statistical learning shapes face evaluation: stimuli. *figshare* https://dx.doi.org/10.6084/m9.figshare.3563514.v1 (2016).

Acknowledgements

The authors are grateful to V. Falvello for her help in data collection, to A. Sklar for early discussions about the work, and to H. Aarts for commenting on a previous version of the manuscript. This research was supported by NWO Rubicon grant no. 446-10-014 awarded to R.D. and United States–Israel Binational Science Foundation grant no. 2013417 awarded to R.R.H. and A.T. The funders had no role in the study design, the data collection and analysis, the decision to publish or the preparation of the manuscript.

Author contributions

R.D. programmed the studies, analysed data and wrote the manuscript. All authors were involved in study design, discussed the results and edited the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.D.

How to cite this article: Dotsch, R., Hassin, R.R. & Todorov, A. Statistical learning shapes face evaluation. *Nat. Hum. Behav.* **1,** 0001 (2016).

Competing interests

The authors declare no competing interests.