Title

Comprehensive Social Trait Judgments From Faces in Autism Spectrum Disorder.

Permalink

https://escholarship.org/uc/item/3rg4p7nq

Journal

Psychological Science, 34(10)

Authors

Cao, Runnan
Zhang, Na
Yu, Hongbo
et al.

Publication Date

2023-10-01

DOI

10.1177/09567976231192236

Peer reviewed

Research Article

Psychological Science
2023, Vol. 34(10) 1121–1145
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976231192236
www.psychologicalscience.org/PS

Comprehensive Social Trait Judgments From Faces in Autism Spectrum Disorder

Runnan Cao[1,2], Na Zhang[2], Hongbo Yu[3], Paula J. Webster[4], Lynn K. Paul[5], Xin Li[2], Chujun Lin[6], and Shuo Wang[1,2]

[1]Department of Radiology, Washington University in St. Louis; [2]Lane Department of Computer Science and Electrical Engineering, West Virginia University; [3]Department of Psychological & Brain Sciences, University of California, Santa Barbara; [4]Department of Chemical and Biomedical Engineering, West Virginia University; [5]Division of the Humanities and Social Sciences, California Institute of Technology; and [6]Department of Psychology, University of California, San Diego

Abstract

Processing social information from faces is difficult for individuals with autism spectrum disorder (ASD). However, it remains unclear whether individuals with ASD make high-level social trait judgments from faces in the same way as neurotypical individuals. Here, we comprehensively addressed this question using naturalistic face images and representatively sampled traits. Despite similar underlying dimensional structures across traits, online adult participants with self-reported ASD showed different judgments and reduced specificity within each trait compared with neurotypical individuals. Deep neural networks revealed that these group differences were driven by specific types of faces and differential utilization of features within a face. Our results were replicated in well-characterized in-lab participants and partially generalized to more controlled face images (a preregistered study). By investigating social trait judgments in a broader population, including individuals with neurodevelopmental variations, we found important theoretical implications for the fundamental dimensions, variations, and potential behavioral consequences of social cognition.

Keywords

social trait judgment, autism spectrum disorder, facial feature, face perception, open data, open materials, preregistered

Received 12/20/22; Revision accepted 7/13/23

People spontaneously make judgments of others' enduring dispositions upon seeing their faces: Some look warm, some look competent, or some look feminine (Lin et al., 2021; Todorov et al., 2015). Although the accuracy of these trait judgments remains debated (Bonnefon et al., 2015), they predict consequential behaviors in the real world, from dating and hiring decisions (Hamermesh, 2011) to voting and courtroom sentencing (Lenz & Lawson, 2011; Wilson & Rule, 2015). Some studies have shown surprisingly high consensus between perceiver groups from different cultures and different age groups (Cogsdill et al., 2014; Hester et al., 2021; Walker et al., 2011). Other researchers have found profound individual differences in such judgments (Hester et al., 2021; Oh et al., 2022; Sutherland et al., 2020). However, it remains unclear whether trait judgments from faces will also be different because of different social functioning such as that occurs in autism spectrum disorder (ASD).

Individuals with ASD show multiple deficits in various aspects of face processing, including gaze processing, discriminating and memorizing different facial identities, and recognizing emotions from facial

Corresponding Authors:
Runnan Cao, Washington University in St. Louis, Department of Radiology
Email: r.cao@wustl.edu

Chujun Lin, University of California, San Diego, Department of Psychology
Email: chl211@ucsd.edu

Shuo Wang, Washington University in St. Louis, Department of Radiology
Email: shuowang@wustl.edu

expressions (Wang & Adolphs, 2017b). They also spend less time engaging in social interactions and looking at faces (Shic et al., 2020), and of course, a core part of the diagnostic criteria includes patterns of social interactions that are different from those of neurotypical individuals. Given these two sets of findings—different face processing and different social behavior—a common hypothesis is that they are causally related: that face processing deficits include difficulties in the kinds of social judgments from faces that drive our social behavior toward other people.

Findings from prior research remain inconclusive on this hypothesis. Studies using computer-generated faces generally have found that individuals with ASD make trait judgments from faces in a way that is similar to neurotypicals (Forgeot d'Arc et al., 2016; Latimier et al., 2019; Lindahl, 2017). For instance, one study investigated seven trait judgments (attractiveness, competence, dominance, extraversion, likeability, threat, and trustworthiness) using computer-generated faces and found no group difference between individuals with ASD and neurotypicals in any of the traits (Lindahl, 2017). In contrast, studies using photographs of real people have revealed trait judgments in ASD that are different from those of neurotypicals (Adolphs et al., 2001; Forgeot d'Arc et al., 2016). It has been shown that individuals with ASD gave more positive ratings to these faces on both traits than neurotypicals when using black-and-white photos of real faces in natural poses (Adolphs et al., 2001). Yet prior studies are limited in their conclusions by the narrow range of traits that are investigated and also by the often narrow diversity of the face stimuli, leaving their relevance to real-world social behavior unclear.

Here, we provide a comprehensive investigation of social trait judgments from faces in individuals with ASD (including both an online sample with self-reported diagnoses and a well-characterized in-lab sample with confirmed ASD diagnoses) in comparison with neurotypicals. To maximize generalizability, we used naturalistic face stimuli of celebrities of diverse races, face angles, gaze directions, and facial expressions taken in naturalistic contexts (e.g., nonposing photos captured in the street or at events; Liu et al., 2015). To reconcile discrepant findings in the literature, we also used more controlled face stimuli of unfamiliar individuals with neutral expressions, direct gaze, and a uniformed background in a preregistered study. We investigated how people make judgments of these faces for a set of eight traits that summarize the comprehensive dimensions of trait judgments from faces (two traits for each of the four dimensions; Lin et al., 2021). It is worth noting that these eight traits represent the core dimensions of social trait judgments from faces that were derived using the most comprehensive trait judgments to date (Lin et al., 2021). Therefore, our findings provide precise predictions about how individuals with ASD and neurotypicals would infer a wide range of social traits from faces. Using these rich data, we leveraged deep learning techniques to characterize the specific patterns and computational bases of the different social trait judgments between participants with ASD and neurotypicals.

Statement of Relevance

Faces are among the most important stimuli that we perceive in everyday life. The spontaneous judgments that people make of others on the basis of faces have been shown to influence consequential real-world decision making. However, existing research heavily relies on neurotypical individuals and highly controlled nondiverse face stimuli. It is important to include different populations and more naturalistic stimuli to advance a more generalizable understanding of how people make these judgments and the biases reflected in them. Here, we comprehensively characterized the similarities and differences in trait judgments from naturalistic faces between neurotypicals and people with autism spectrum disorder, who often have deficits in perceiving faces. Our findings provide new insights into how people mentally represent the relationship between different social trait judgments, why people make different social judgments from faces, and how these judgments may influence a wide range of behavior.

Method

Participants

In our main experiment, we recruited 525 participants from the Prolific platform (referred to as online participants). We included only participants who had English fluency, normal or corrected-to-normal vision, an education level above high school, and a Prolific approval rating greater than 95%. Among these participants, 113 participants had a self-reported diagnosis of ASD (SR-ASD), and 412 neurotypical participants reported no diagnosis of ASD and served as controls (see Table 1 for demographics). Self-report of ASD was probed by the following question in Prolific: "Have you received a formal clinical diagnosis of autism spectrum disorder, made by a psychiatrist, psychologist, or other qualified medical specialist? This includes Asperger's syndrome, autism disorder, high-functioning autism, or pervasive developmental disorder." We included only participants

Table 1. Summary of Participants

Group	Sex (male/female)	Age in years	Caucasian	AQ	SRS	FSIQ	ADOS		
							Communication	Social interaction	Sum
Online SR-ASD	53/59	*M* = 28.90 (8.37)	64.6%	*M* = 27.8 (8.09)	*M* = 91.7 (29.7)				
Online neurotypical	256/155	*M* = 26.34 (7.12)	67.88%	*M* = 20.3 (6.82)	*M* = 65.2 (25.2)				
In-lab ASD	23/4	*M* = 28.78 (8.55)	77.78%	*M* = 29.8 (6.53)	*M* = 85.0 (26.2)	*M* = 105.04 (15.05)	3.08	7.31	10.38
In-lab neurotypical	12/9	*M* = 30.95 (4.19)	57.14%	*M* = 11.5 (5.87)	*M* = 20.7 (16.4)	*M* = 108.50 (12.07)			
Replication online SR-ASD	116/131	*M* = 28.49 (7.32)	78.78%	*M* = 32.0 (9.37)	*M* = 105.5 (31.9)				
Replication online neurotypical	158/93	*M* = 25.88 (7.11)	68.13%	*M* = 20.1 (7.04)	*M* = 65.2 (23.6)				

Note: Standard deviations are given in parentheses. In our main experiment with naturalistic faces, we recruited online participants who self-reported a positive clinical diagnosis of ASD (SR-ASD) and online neurotypicals. In our first control/validation experiment, we recruited in-lab participants with ASD and in-lab neurotypicals. In our second control/validation experiment, we recruited another population of online participants. For all of our in-lab participants with ASD, their diagnosis was confirmed using the Autism Diagnostic Observation Schedule–2 (ADOS-2; Lord et al., 1989). We used Module 4 for adults and older adolescents and Module 3 for younger adolescents. The ADOS is a structured interaction with an experimenter, which is videotaped and scored by trained clinical staff in our laboratory, yielding scores on several scales. Scoring followed standard protocols for ADOS-2 as well as Calibrated Severity Scores (Hus & Lord, 2014). Social Responsiveness Scale–2 Adult Self-Report (SRS) raw scores are shown for online participants, and SRS T scores are shown for in-lab participants. AQ = Autism-Spectrum Quotient; FSIQ = Full-Scale Intelligence Quotient.

whose response was “Yes–as a child” or “Yes–as an adult” in the SR-ASD group (not including any participants whose response was “I am in the process of receiving a diagnosis,” “No–but I identify as being on the autism spectrum,” “No,” or “Don’t know/rather not say”). We further acquired Autism-Spectrum Quotient (AQ; Baron-Cohen et al., 2001) and Social Responsiveness Scale–2 Adult Self-Report (SRS; Constantino & Gruber, 2012) scores from the participants (92 participants with SR-ASD and 337 neurotypicals completed the questionnaires). These data confirmed that online participants with SR-ASD had significantly higher AQ scores (see Fig. S1a in the Supplemental Material available online; SR-ASD: $M = 27.76$, $SD = 8.09$; neurotypical: $M = 20.28$, $SD = 6.82$), $t(427) = 8.94$, $p = 1.15\times10^{-17}$, $d = 1.05$, 95% confidence interval (CI) = [5.83, 9.12], and SRS scores (see Fig. S1b; SR-ASD: $M = 91.73$, $SD = 29.66$; neurotypical: $M = 65.17$, $SD = 25.19$), $t(427) = 8.61$, $p = 1.38\times10^{-16}$, $d = 1.01$, 95% CI = [20.50, 32.61], than online neurotypical participants. Furthermore, online participants with SR-ASD had AQ scores (two-tailed two-sample t test), $t(112) = 0.92$, $p = .36$, $d = 0.22$, 95% CI = [−6.21, 2.28], and SRS scores, $t(109) = 1.44$, $p = .15$, $d = 0.36$, 95% CI = [−4.00, 25.45], that were comparable with in-lab participants with ASD (see below). Lastly, based on our screening criterion, online neurotypicals had no mental health conditions.

Because of a surge of female participants on the Prolific platform during our data collection for participants with SR-ASD (Charalambides, 2021), the female population of participants with SR-ASD was overrepresented in our sample (see Table 1; but see Maenner et al., 2020, for prevalence of ASD in the general population). However, we observed qualitatively the same results with male participants with SR-ASD only (see Fig. S1h) as well as participants with a balanced distribution of sexes across groups (see Fig. S1i). In addition, although the two groups of participants that we sampled differed in age (see Table 1 and Fig. S1c), $t(523) = 3.25$, $p = .0012$, $d = 0.34$, 95% CI = [1.01, 4.10], we observed similar results when we compared a subset of participants who were matched in age (see Fig. S1g).

In our first control experiment, we recruited 27 participants with ASD who had typical intellectual functioning (Full-Scale Intelligence Quotient > 80) from our laboratory’s registry and 21 neurologically and psychiatrically healthy participants with no family history of ASD as controls (referred to as in-lab participants; see Table 1 for demographics). All of our in-lab ASD participants met the criteria of the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association, 2013; confirmed by diagnostic interview with a licensed clinical psychologist) and the Autism Diagnostic Observation Schedule (ADOS; administered by a research reliable investigator and confirmed with consensus coding by a team of research reliable administrators; Hus & Lord, 2014; Lord et al., 1989) for ASD (see Table 1). We confirmed that in-lab participants with ASD had significantly higher AQ scores (ASD: $M = 29.57$, $SD = 12.06$; neurotypical: $M = 13.94$, $SD = 5.72$), two-tailed two-sample t test: $t(38) = 5.00$, $p = 1.31\times10^{-5}$, $d = 1.56$, 95% CI = [9.40, 22.17], and SRS scores (ASD: $M = 79.85$, $SD = 28.27$; neurotypical: $M = 32.29$, $SD = 25.63$), two-tailed two-sample t test: $t(38) = 5.69$, $p = 1.54\times10^{-6}$, $d = 1.76$, 95% CI = [31.37, 66.06], than in-lab neurotypical participants.

In our second control experiment, we recruited another 247 participants with SR-ASD and another 251 neurotypical participants from the Prolific platform as an independent replication sample (see Table 1). The data collection and some data analyses in this experiment were preregistered (https://osf.io/bdrty/). Only 18 participants with SR-ASD and two neurotypical participants from the main experiment participated in the second control experiment.

All participants provided written informed consent using procedures approved by the institutional review board of West Virginia University (Protocol #2012188080) and California Institute of Technology (Protocol #19-234).

Stimuli

To increase generalizability, we used naturalistic face images in our main experiment and the first control experiment. These images were ambient photos of celebrities from the CelebA data set (Liu et al., 2015). We selected 50 identities with 10 images for each identity, for a total of 500 face images. The identities were selected to include both sexes (33 male) and multiple races (40 identities were Caucasian, nine identities were African American, and one identity was biracial). The faces were of different angles and gaze directions, with diverse backgrounds and lighting. The faces showed various facial expressions, with some having accessories such as sunglasses and hats.

Some prior studies also used highly controlled face images. To reconcile prior discrepant findings that might be due to image types, we used posed neutral faces in our second control experiment. These were 50 face images of 50 different facial identities (25 female, 25 male). These faces were randomly selected from a representatively sampled set of 100 White faces from a previous study (Lin et al., 2021). They were high-resolution studio photographs of human participants from three popular databases: the Chicago Face Database (Ma et al., 2015), the Oslo Face Database (Chelnokova et al., 2014), and the Face Research Lab London (DeBruine & Jones, 2017). All face stimuli were frontal and clear, with a neutral expression, and were presented at the center of the images with the eyes aligned to the same location. All photos included the face, neck, and hair. All photos were colored, with a standard gray background, and were cropped to a standard size and shape.

Procedures

Participants rated the faces on eight social traits using a 7-point Likert-type scale through an online rating task. The social traits included *warm*, *critical*, *competent*, *practical*, *feminine*, *strong*, *youthful*, and *charismatic*. These social traits were well validated in a previous study (Lin et al., 2021). Participants also indicated whether they recognized the identity of the faces (i.e., whether they were familiar with each face identity) in the main experiment. We did not find a significant correlation between the percentage of familiar identities and AQ score, SR-ASD: $r(90) = -.11$, $p = .28$; neurotypical: $r(335) = -.05$, $p = .32$, or SRS score, SR-ASD: $r(90) = -.16$, $p = .13$; neurotypical: $r(335) = -.005$, $p = .93$, suggesting that famous face recognition was not related to autistic traits in our participants.

The celebrity faces were randomly divided into 10 modules, with each module containing one face image per face identity (totaling 50 face images per module). In each module, participants rated the faces on all eight social traits (rated in blocks). Participants completed as many modules as they wanted. In our main experiment, online participants with SR-ASD completed one to 10 modules, and online neurotypical participants completed one to two modules. In our first control experiment, in-lab participants with ASD completed four to 10 modules, and in-lab neurotypicals completed one to 10 modules. In our second control experiment, each participant rated all 50 highly controlled face images.

We applied the following three exclusion criteria:

(1) Trial-wise exclusion: We excluded trials with reaction times shorter than 100 ms or longer than 5,000 ms.
(2) Block/trait-wise exclusion: We excluded the entire block per module if more than 30% of the trials were excluded from the block per (1) above or if there were fewer than three different rating values in the block (this suggests that the participant may not have used the rating scale properly).
(3) Module-wise exclusion: We excluded a module if more than three blocks were excluded from the module per (2) above.

Interrater consistency

Interrater consistency of each trait was estimated using the intraclass correlation coefficient (ICC; two-way random-effects model for the consistency of mean ratings; McGraw & Wong, 1996) and the Spearman's correlation coefficient (ρ). The ICC and Spearman's ρ were computed between raters for each trait in each module and then averaged across modules per trait. The ICC was calculated using MATLAB (The MathWorks, Natick, MA) implementation written by Arash Salarian (https://www.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc). The Spearman's ρ was computed between each pair of raters and then averaged across all pairs of raters.

Principal component analysis

To characterize the psychological dimensions of social trait judgments from faces in each participant group, we conducted a principal component analysis, which is a statistical procedure that converts a set of high-dimensional, possibly correlated variables into a set of low-dimensional, linearly uncorrelated principal components that preserve as much of the variance in the original variables as possible. We first aggregated the rating data per trait across participants within each participant group for each face. On the basis of the aggregated data (500 faces × 8 traits), we extracted eight principal components (using *R* function *principal*, without rotation) for each participant group. We retained principal components that explained a nontrivial amount of variance (> 5%). After identifying the optimal number of principal components, we applied varimax rotation to the principal components to generate orthogonal components that were most interpretable.

Classification of participants

To examine whether social trait judgments made by participants with SR-ASD were different from those made by neurotypicals across all faces and traits, we employed a linear support vector machine, which discriminated whether a rating module was from a participant with SR-ASD or a neurotypical. We used all ratings (8 traits × 50 faces) in each module as features for model training and testing. To assess model performance, in each run, we randomly partitioned the modules into 10 equal portions and used tenfold cross-validation (i.e., each time nine portions of modules were used as the training set and the remaining one portion of modules was used as the testing set). We repeated the cross-validation 1,000 times in total.

Feature extraction and construction of feature space

To investigate the visual computational mechanism underlying social trait judgments from faces, we leveraged artificial neural networks (ANNs). ANNs have been successfully applied by prior research to advance a mechanistic understanding of face perception in ASD. For instance, a recent study used brain-tissue-mapped ANN models of primate vision to explore neural and behavioral markers of atypical facial emotion recognition in ASD (Kar, 2022). The study revealed that the image-level behavioral patterns of the ANNs matched those of neurotypical individuals more closely than individuals with ASD, and this behavioral mismatch was most prominent when the ANN behavior was decoded from units corresponding to the primate inferior temporal cortex (Kar, 2022).

Specifically, here we used the well-known deep neural network (DNN) implementation based on the VGG-16 (Visual Geometry Group, Oxford, UK) convolutional neural network architecture (Parkhi et al., 2015) to extract features for each face image. Fine-tuning was performed on the pretrained VGG-Face deep model using all images of the 50 identities in the CelebA data set (16–30 images for each identity). Features that differentiated identities (i.e., identity recognition) were extracted using this transferred model. We subsequently applied a t-distributed stochastic neighbor embedding (t-SNE) method to convert high-dimensional features into a two-dimensional feature space. t-SNE is a variation of SNE (Hinton & Roweis, 2003), a commonly used method for multiple class high-dimensional data visualization (van der Maaten & Hinton, 2008). We applied t-SNE for each layer, with the cost function parameter (Prep) of t-SNE, representing the perplexity of the conditional probability distribution induced by a Gaussian kernel, set individually for each layer. We implemented t-SNE in the MATLAB platform. Notably, neither feature extraction nor construction of feature space used any information from social trait ratings.

To identify the regions in the face feature space that elicited a significant judgment difference between participants with SR-ASD and neurotypicals (for a detailed illustration, see Fig. S4 in the Supplemental Material), we first estimated a continuous density map in the feature space by smoothing the discrete rating differences between groups using a two-dimensional Gaussian kernel (kernel size = feature dimension range * 0.05, *SD* = 2). We then estimated statistical significance for each pixel by permutation testing: In each of the 1,000 permutations, we randomly shuffled the labels of participants. We calculated the p value for each pixel by comparing the observed density value with those from the null distribution derived from permutations. We applied a mask to exclude pixels from the edges and corners of the density map where there were no faces because these regions were susceptible to false positives given our procedure. We selected the regions with significant pixels (permutation $p < .01$, false discovery rate corrected for $q < 0.01$, cluster size > 5% of the pixels within the mask; Benjamini & Hochberg, 1995).

Representational similarity between social trait ratings and DNN features

We employed a pairwise distance metric (Grossman et al., 2019) to compare representational similarity between social trait ratings and DNN features. For a given trait, we calculated the absolute difference in average ratings for each pair of face identities as the pairwise distance metric for social trait ratings, and we calculated the Euclidean distance of all DNN units from a layer for each pair of face identities as the pairwise distance metric for DNN features. We then correlated the two pairwise distance metrics using the Spearman correlation (which does not assume a linear relationship) and computed the correlation for each DNN layer. We conducted this analysis separately for each participant group. Because the consistency between face images for the same face identity in both social trait ratings and DNN features could inflate the correlation between the two distance metrics, we averaged the social trait ratings or DNN features across face images for each face identity first and then calculated the pairwise distance metrics between face identities.

To determine the statistical significance of the representational similarity between social trait ratings and DNN features, we used a nonparametric permutation test with 1,000 permutations. In each permutation, we randomly shuffled the *face identity* labels and calculated the correlation between the two distance metrics. The distribution of correlation coefficients computed with shuffling (i.e., null distribution) was compared with the one without shuffling (i.e., observed value). An observed value was deemed significant if it was greater than 95% of the values from the null distribution. A significant correlation indicated a representational similarity between social trait ratings and DNN features.

To determine the statistical significance of the representational similarity in the social trait ratings between groups (participants with SR-ASD and neurotypicals), we used a permutation test with 1,000 permutations. In each permutation, we shuffled the *participant* labels and calculated the difference in representational similarity between participant groups. We then compared the observed difference in representational similarity between participant groups with the permuted null distribution to derive statistical significance.

Visualization of critical pixels within faces for social trait judgment

We built a DNN-based regression model for each trait and each participant group. We employed transfer learning for the model. Transfer learning is a popular deep learning method in which a model developed for one task can be reused as the initial model for a second related task. Here, a VGG-16 model (a classifier), pretrained using ImageNet stimuli (Deng et al., 2009; Simonyan & Zisserman, 2015), was used as the initial model. ImageNet contains stimuli of both faces and objects. Prior research showed that DNNs trained on objects alone performed poorly on face recognition, whereas DNNs dual-trained on both faces and objects can best capture recognition of both faces and objects and their functional segregation (Dobs et al., 2022). Furthermore, our prior work using DNN models trained on faces as well as faces and objects revealed a novel region-based feature code of faces in the human amygdala and hippocampus (Cao et al., 2020).

Specifically, here we kept all convolution layers of the VGG-16 model but replaced the last two fully connected layers and the output layer with a global averaging pooling layer, a fully connected layer, and a prediction output layer (for an illustration, see Fig. S6a in the Supplemental Material). When training our regression model, we froze all convolutional layers (i.e., weights were not updated), and only the top layers (the replaced layers) were updated by training. Training was performed by the stochastic gradient descent optimizer with the base learning rate of 10^{-3}, and we used mean squared error as the loss function. The training stopped when the loss converged. Before the images were fed into our model, they were first cropped (using the dlib toolbox) and resized to 224 × 224. We cropped the faces using a bounding box that included the entire face and hair region.

We performed tenfold cross-validation in our analysis. In each training/testing run (separately for each trait and each participant group), the data set was randomly split into 10 subsets. One subset served as the test set, and the remaining nine subsets were used as the training set. To assess model performance, we calculated the correlation between the observed trait values and the predicted trait values in the testing set (note that the output was switched from classification to regression to get a continuous prediction of trait values). The correlation coefficient (Pearson's r) indicated the model prediction accuracy. Our VGG-16 network was run on the deep learning framework TensorFlow 1.15 using Python 3.6.

To explain our model's output in the domain of its input (i.e., face images), we applied layer-wise relevance propagation (LRP) to our trained regression models. LRP can use the network weights created by the forward-pass to propagate the output back through the network up to the original input image. The explanation given by LRP is a heatmap of which pixels in the original image contribute to predicting social judgments. We used the toolbox iNNvestigate (Alber et al., 2019; https://github.com/albermax/innvestigate) for implementation.

It is worth noting that in this analysis, we focused on the VGG-16-based DNN models and LRP as an image-based explanation. However, different DNN models and visualization/explanation methods may yield varying results. Prior research has demonstrated that features from both face identification DNNs and object recognition DNNs outperform facial geometry across multiple social trait judgments and out-of-sample data sets (although object recognition DNN features' predictions are susceptible to superficial cues such as color and hairstyle) and that face identification DNN features' predictions are nonspecific, meaning that models trained to predict one social judgment can also predict other social judgments (Keles et al., 2021). In line with this result, different DNN models, including object recognition models, may produce similar results for face coding (Cao et al., 2020). Different visualization/explanation methods may provide feature importance estimates that are not superior to random designations of feature importance (Hooker et al., 2019). This ambiguity poses a significant challenge to the current DNN explainability results and that different visualization/explanation methods could lead to different inferences from those presented here (Kar et al., 2022).

Results

The same set of psychological dimensions underlies trait judgments in ASD and neurotypicals

We recruited participants with SR-ASD and neurotypicals online (see the Method section; results replicated with in-lab participants with ASD who had an ADOS diagnosis). Participants from each group (see Table 1 and Fig. S1 for summary) rated the faces on eight traits: *warm*, *critical*, *competent*, *practical*, *feminine*, *strong*, *youthful*, and *charismatic* (see Fig. 1a for ranking of stimuli based on ratings for each trait). To understand the overall structure of the data, we first analyzed the core dimensions that underlie the eight trait judgments in each group. To this end, we conducted a principal component analysis on the aggregate ratings (averaged

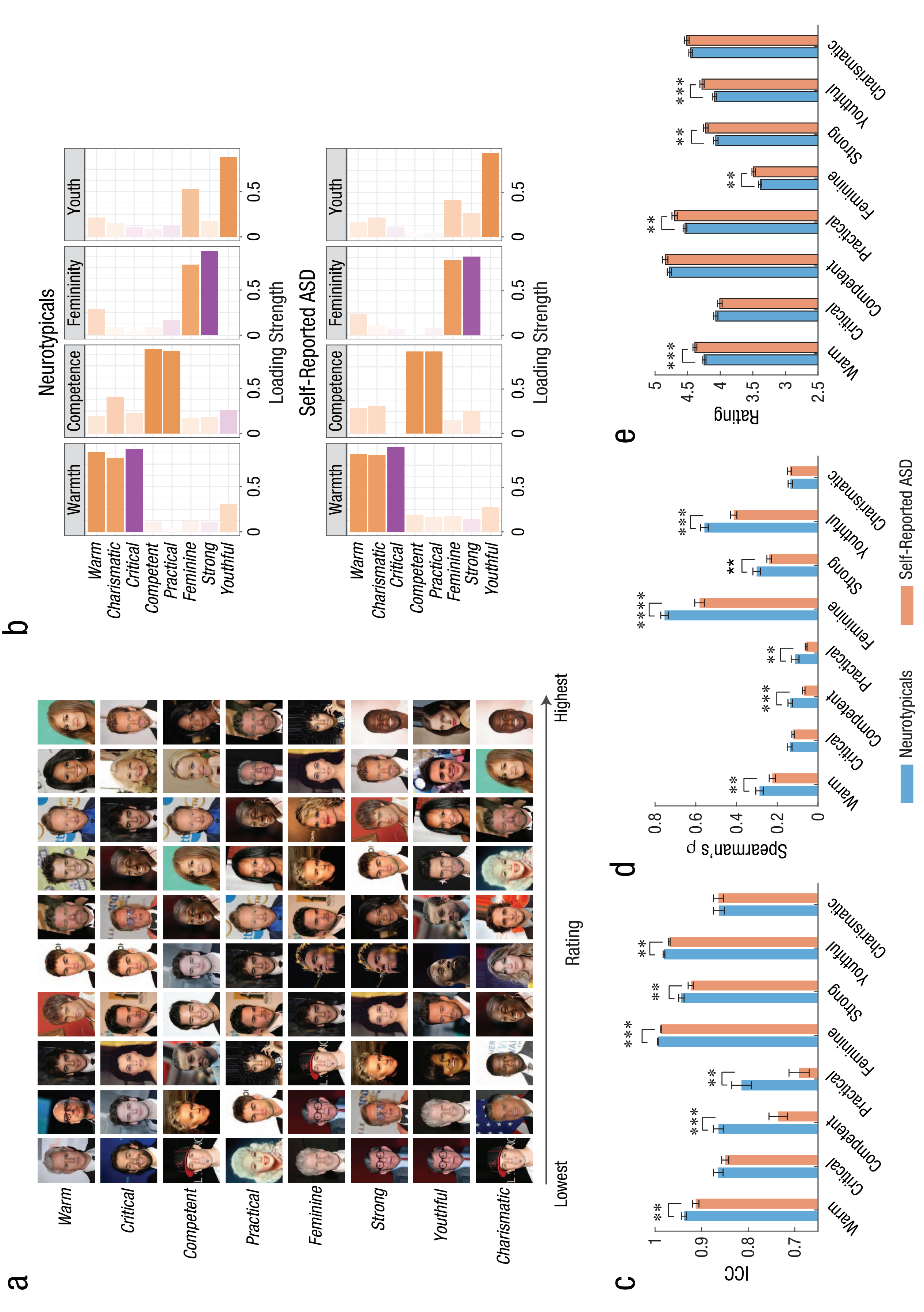

Fig. 1. *(continued on next page)*

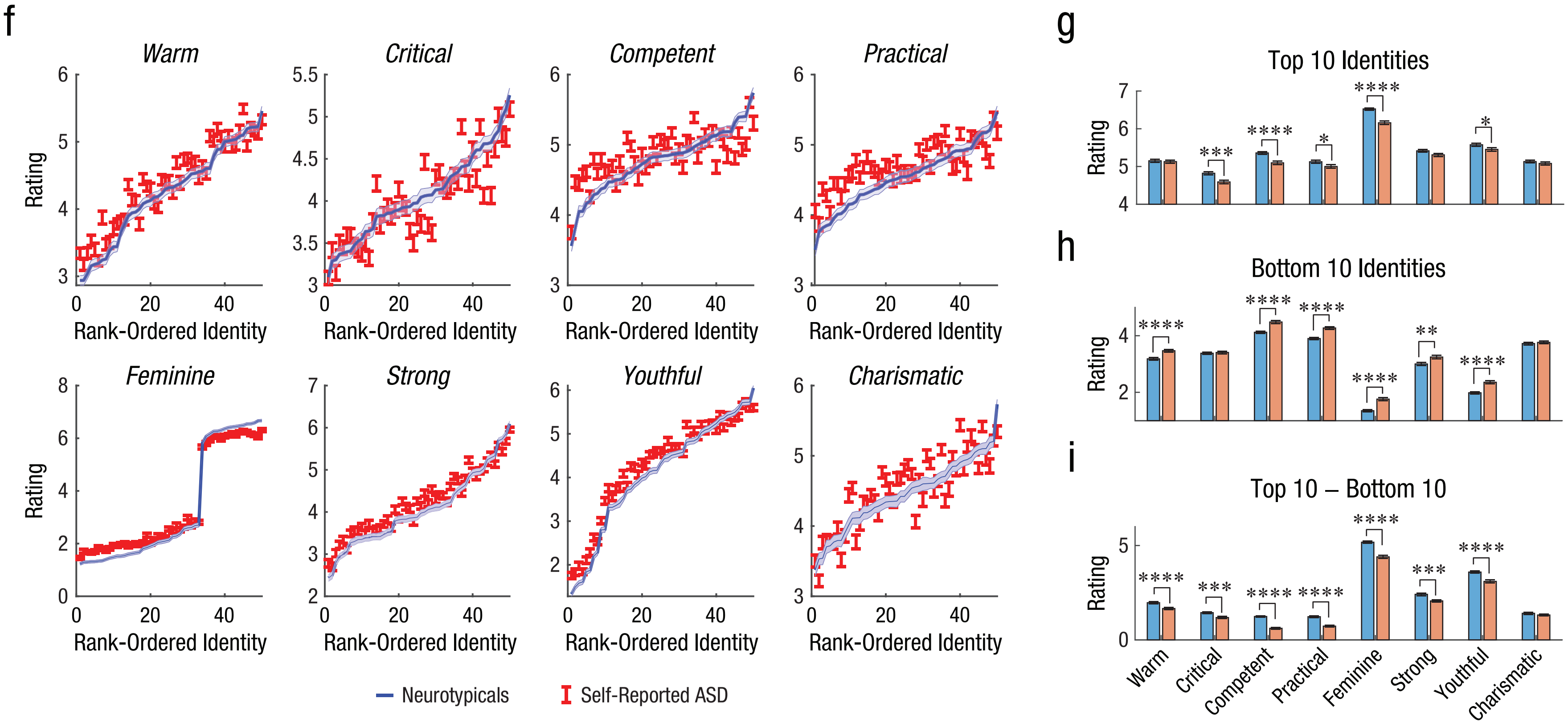

Fig. 1. Social trait judgments from participants who self-reported a positive clinical diagnosis of autism spectrum disorder (SR-ASD) and neurotypicals. (a) Example stimuli ranked by average ratings from neurotypicals for each social trait. (b) Principal component analysis loadings of social traits on the first four principal components. Each column plots the strength of the loadings (*x*-axis, absolute value) across traits (*y*-axis). Color coding indicates the sign of the loading (orange for positive, purple for negative). Saturated colors highlight each trait's most strongly correlated principal component. (c, d) Interrater consistency of each trait was estimated using (c) the intraclass correlation coefficient (ICC; McGraw & Wong, 1996) and (d) the Spearman's correlation coefficient (ρ). Interrater consistency was first calculated between raters and averaged within each module and then averaged across modules. Participants with SR-ASD demonstrated lower interrater consistency for most of the traits: *warm* (two-tailed paired-samples *t* test across 10 rating modules), ICC: $t(9) = 3.45$, $p = .0073$, $d = 1.28$, 95% confidence interval (CI) = [0.009, 0.04]; Spearman: $t(9) = 3.19$, $p = .011$, $d = 1.18$, 95% CI = [0.02, 0.11]; *competent*, ICC: $t(9) = 5.43$, $p = .00042$, $d = 2.49$, 95% CI = [0.07, 0.02]; Spearman: $t(9) = 5.78$, $p = .00027$, $d = 2.24$, 95% CI = [0.04, 0.09]; *practical*, ICC: $t(9) = 4.21$, $p = .0023$, $d = 1.81$, 95% CI = [0.06, 0.19]; Spearman: $t(9) = 3.26$, $p = .0099$, $d = 1.23$, 95% CI = [0.02, 0.09]; *feminine*, ICC: $t(9) = 5.19$, $p = .00057$, $d = 2.08$, 95% CI = [0.004, 0.009]; Spearman: $t(9) = 7.28$, $p = 4.66{\times}10^{-5}$, $d = 2.56$, 95% CI = [0.12, 0.23]; *strong*, ICC: $t(9) = 4.06$, $p = .0029$, $d = 1.11$, 95% CI = [0.009, 0.03]; Spearman: $t(9) = 4.09$, $p = .0027$, $d = 1.28$, 95% CI = [0.03, 0.10]; and *youthful*, ICC: $t(9) = 4.49$, $p = .0015$, $d = 2.09$, 95% CI = [0.006, 0.02]; Spearman: $t(9) = 5.76$, $p = .00027$, $d = 2.64$, 95% CI = [0.09, 0.20]. (e) For aggregate ratings, participants with SR-ASD gave statistically different ratings for five traits (two-way repeated measures analysis of variance)—main effect of participant group: $F(1, 5613) = 12.15$, $p = 5.20{\times}10^{-4}$, $\eta^2 = .005$; main effect of trait: $F(7, 5163) = 214.35$, $p = 6.46{\times}10^{-235}$, $\eta^2 = .24$; interaction: $F(7, 5613) = 4.27$, $p = 1.03{\times}10^{-4}$, $\eta^2 = .002$: *warm* (two-tailed two-sample *t* test across participants), $t(824) = 3.31$, $p = .00097$, $d = 0.23$, 95% CI = [0.06, 0.22]; *practical*, $t(802) = 3.13$, $p = .0018$, $d = 0.22$, 95% CI = [0.06, 0.26]; *feminine*, $t(736) = 2.65$, $p = .0082$, $d = 0.20$, 95% CI = [0.03, 0.17]; *strong*, $t(822) = 3.10$, $p = .0020$, $d = 0.22$, 95% CI = [0.06, 0.25]; and *youthful*, $t(829) = 4.47$, $p = 9.03{\times}10^{-6}$, $d = 0.31$, 95% CI = [0.11, 0.28]. Error bars denote standard errors of the mean (± *SEM*) across rating modules. Asterisks indicate a significant difference between participants with SR-ASD and neurotypicals using two-tailed two-sample *t* test. *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$. (f) Ratings for each face identity rank-ordered by mean ratings from neurotypicals (red for SR-ASD, blue for neurotypical). Error bars and error shades denote standard errors of the mean (± *SEM*) across rating modules. (g) Average ratings for the 10 identities with the highest ratings from neurotypicals. (h) Average ratings for the 10 identities with the lowest ratings from neurotypicals. (i) Difference in ratings between the top 10 and bottom 10 identities.

per face across participants) across the eight traits for each group. The first four principal components (without rotation) explained most of the variance in each group: 44%, 23%, 14%, and 11% in online participants with SR-ASD (total 92%) and 38%, 27%, 17%, and 9% in online neurotypicals (total 92%). These results indicate that four dimensions optimally summarized the eight trait judgments of our naturalistic face stimuli.

Therefore, we extracted four principal components from each group and applied the varimax rotation for maximal interpretability. The four dimensions from each group could be interpreted as warmth, competence, femininity, and youth (for principal component loadings, see Fig. 1b; for correlations between trait judgments, see Fig. S2a in the Supplemental Material). These results replicated the comprehensive trait dimensions found in prior research that used posed neutral White faces (Lin et al., 2021) with a different type of face stimuli (naturalistic faces of famous people of different races and with different facial expressions) and different groups of neurotypical participants in the present research. We confirmed that this replication was not simply due to our selection of traits: Including additional ratings on popular traits (trustworthiness, dominant) that were not representative of the four dimensions again replicated the four dimensions. We computed the Tucker index of factor congruence between the four dimensions found in each group using their principal component loadings (i.e., cosine distance between loadings). We found that the four dimensions found in both groups were highly similar (Tucker indices = 0.99, 0.97, 0.99, and 0.99 between SR-ASD and neurotypical). These results suggest that the comprehensive psychological dimensions that underlie social trait judgments from faces are similar between participants with SR-ASD and neurotypicals.

Different trait ratings along all comprehensive dimensions in individuals with ASD compared with neurotypicals

A highly similar correlational structure across trait judgments between groups does not guarantee group similarity in the rating values of every trait (e.g., correlation removes information about the mean). Here, we compared the judgment of each trait between the two groups. We first analyzed the interrater consistency for each trait judgment (see Figs. 1c and 1d). We found that participants with SR-ASD were more heterogeneous with respect to each other than neurotypicals for six of the eight traits distributed across all four dimensions (for statistics, see legends of Figs. 1c and 1d), consistent with the widely reported heterogeneity in ASD (Happe et al., 2006).

Trait judgments were highly consistent for different face images of the same identity for both participants with SR-ASD and neurotypicals (see Fig. S2b). We next compared the mean of the aggregate ratings across participants per trait between groups. We found that participants with SR-ASD gave statistically different ratings for five of the eight traits distributed across all four dimensions (see Fig. 1e; see figure legend for statistics): *warm*, *practical*, *feminine*, *strong*, and *youthful*. Participants with SR-ASD and neurotypicals could be distinguished on the basis of how they rated the faces on the eight traits (support vector machine classifier with tenfold cross-validation and 1,000 repetitions; mean accuracy across runs = 79.66%, *SD* = 0.66%). These results suggest that individuals with SR-ASD tend to evaluate faces differently across all four comprehensive dimensions.

We further zoomed into each face identity and examined which face identities led to the most discrepant ratings between groups. We rank-ordered the face identities according to the average ratings from the neurotypicals. We found that for the judgments of *warm*, *practical*, *strong*, and *youthful*—four traits that distributed across all four comprehensive dimensions—participants with SR-ASD gave higher ratings for most of the face identities (see Fig. 1f). These results showed that the more positive trait judgments in SR-ASD compared with neurotypicals were not merely driven by certain face identities. Interestingly, we found that for the judgments of *competent*, *practical*, and *feminine*, participants with SR-ASD demonstrated a compressed range in their ratings across faces. That is, participants did not vary their ratings as much as neurotypicals across face identities (see Fig. 1f for examples), leading to higher ratings on the faces that neurotypicals judged low and lower ratings on the faces that neurotypicals judged high.

To formally quantify this observation, we compared the ratings between groups separately for the 10 face identities on which neurotypicals provided the highest ratings (see Fig. 1g) and the 10 face identities on which neurotypicals provided the lowest ratings (see Fig. 1h). We found that compared with neurotypicals, participants with SR-ASD provided significantly lower ratings for the top 10 identities when judging *critical* (see Fig. 1g), $t(806) = 3.33$, $p = .00090$, $d = 0.26$, 95% CI = [0.11, 0.35]; *competent*, $t(807) = 4.74$, $p = 2.49\times10^{-6}$, $d = 0.31$, 95% CI = [0.15, 0.38]; *practical*, $t(802) = 2.05$, $p = .041$, $d = 0.14$, 95% CI = [0.006, 0.25]; *feminine*, $t(736) = 6.51$, $p = 1.39\times10^{-10}$, $d = 0.48$, 95% CI = [0.26, 0.48]; and *youthful*, $t(829) = 2.24$, $p = .026$, $d = 0.15$, 95% CI = [0.01, 0.24]; they provided significantly higher ratings for the bottom 10 identities for *warm* (see Fig. 1h), $t(824) = 4.60$, $p = 4.80\times10^{-6}$, $d = 0.32$, 95% CI = [0.16, 0.41]; *competent*, $t(807) = 5.52$, $p = 4.66\times10^{-8}$, $d = 0.39$, 95% CI = [0.23, 0.49]; *practical*, $t(802) = 5.48$, $p = 5.64\times10^{-8}$, $d = 0.39$, 95% CI = [0.24, 0.51]; *feminine*, $t(736) = 6.17$, $p = 1.13\times10^{-9}$, $d = 0.45$, 95% CI = [0.28, 0.54]; *strong*, $t(822) = 3.06$, $p = .0023$, $d = 0.21$, 95% CI = [0.09, 0.40]; and *youthful*, $t(829) = 5.55$, $p = 3.93\times10^{-8}$, $d = 0.38$, 95% CI = [0.24, 0.50]. Therefore, the rating difference between the top 10 and bottom 10 identities was significantly smaller in participants with SR-ASD compared with neurotypicals across trait judgments along all four dimensions (see Fig. 1i): *warm*, $t(824) = 4.03$, $p = 6.10\times10^{-5}$, $d = 0.28$, 95% CI = [0.16, 0.46]; *critical*, $t(806) = 3.30$, $p = .001$, $d = 0.23$, 95% CI = [0.10, 0.41]; *competent*, $t(807) = 9.82$, $p = 1.35\times10^{-21}$, $d = 0.69$, 95% CI = [0.50, 0.75]; *practical*, $t(802) = 6.85$, $p = 1.52\times10^{-21}$, $d = 0.48$, 95% CI = [0.36, 0.64]; *feminine*, $t(736) = 6.51$, $p = 7.01\times10^{-13}$, $d = 0.52$, 95% CI = [0.56, 0.99]; *strong*, $t(822) = 3.63$, $p = .0003$, $d = 0.25$, 95% CI = [0.16, 0.54]; and *youthful*, $t(829) = 5.07$, $p = 4.89\times10^{-7}$, $d = 0.35$, 95% CI = [0.30, 0.69]. Together, these results suggest that participants with SR-ASD have a reduced discriminability for social trait judgments across all four comprehensive dimensions. These findings are consistent with ASD's reduced specificity in emotion perception (Wang & Adolphs, 2017a) and noisier and more random eye movement behavior in general (de Wit et al., 2008; Pelphrey et al., 2002; Wang et al., 2015).

Because our face stimuli were photos of celebrities, with whom participants might be familiar, we investigated how familiarity of faces might influence trait judgments (see Figs. S2c and S2d). Participants with SR-ASD rated the faces more differently from neurotypicals for unfamiliar identities (see Fig. S2c) compared with familiar identities (see Fig. S2d). Specifically, participants with SR-ASD rated unfamiliar identities on *warm*, *feminine*, *strong*, and *youthful* significantly higher than neurotypicals (see Fig. S2c), and they rated familiar identities on *practical* and *feminine* significantly higher than neurotypicals (see Fig. S2d; see figure legend for statistics). The distribution of familiar and unfamiliar identities was similar between participant groups. Therefore, these results suggest that face familiarity moderated the differences in face judgments between participants with SR-ASD and neurotypicals along all four comprehensive trait dimensions.

Because prior findings showed that people are biased by racial information when making trait judgments (Hugenberg et al., 2011; Zebrowitz et al., 2010), we capitalized on our racially diverse stimuli and participants to analyze potential cross-race effects (see Figs. S2e and S2f). We found that group differences in trait judgments between individuals with SR-ASD and neurotypicals were primarily driven by faces that were the same race as the participants (see Fig. S2e; see figure legend for statistics) rather than cross-race faces (see Fig. S2f). In addition, we found that group differences in trait judgments were primarily driven by faces that were the same sex as the participants (see Fig. S2g; see figure legend for statistics) rather than cross-sex faces (see Fig. S2h). Together, these findings suggest that people with SR-ASD give the most different trait judgments compared with neurotypicals when the face being judged belongs to an in-group member with respect to race and sex.

Finally, facial expressions of emotion may influence whether individuals with SR-ASD and neurotypicals make similar or different social trait judgments. To investigate this question, two researchers labeled the emotion of each face, and we redid our analyses using only faces that were emotionally neutral ($n = 225$). Highly similar results were derived (see Fig. S3 in the Supplemental Material). These results showed that our findings regarding the group differences between individuals with SR-ASD and neurotypicals were not merely driven by faces with nonneutral facial expressions.

Features across faces that contribute to different trait ratings in individuals with ASD compared with neurotypicals

What types of faces drove the rating differences between participants with SR-ASD and neurotypicals? Do the faces that participants with SR-ASD judged most differently from neurotypicals share common visual features? To answer these questions, we extracted facial features from each image using a pretrained DNN VGG-Face (Parkhi et al., 2015) and constructed a two-dimensional face feature space using t-SNE for each DNN layer. Although this DNN model was originally trained to classify face identities, it has been shown that its features also predict human judgments of faces on a wide range of social traits (Keles et al., 2021; Parde et al., 2019; note that other DNN models could derive similar results; Cao et al., 2020). Furthermore, this model is associated with neural processing of faces in the human brain, at both the single-neuron level (Cao et al., 2020) and the neural population level (Grossman et al., 2019). Importantly, the dimensions (or axes) of the face feature space represented interpretable variations in faces (e.g., gender). Faces that clustered in this space also shared interpretable visual features. For example, faces of the same identity were clustered, and darker skinned faces were clustered at the bottom left corner of the feature space. Therefore, identifying a region in the face feature space would reveal what types of common visual features in the faces drove most discriminative judgments between groups.

We projected the difference in rating per trait between groups for each face onto the DNN-derived face feature space per DNN layer (i.e., multiplying the difference in rating of each face to its corresponding location in the feature space to derive a rating-weighted two-dimensional feature map; see Figs. S4a and S4i). To formally quantify the difference in feature maps between groups and identify discriminative feature map regions for each social trait (see Fig. S4 for illustration of detailed procedures), we estimated a continuous density map in the feature space from our sparse sampling (see left side of Figs. 2a–2d and Figs. S4b, S4d, S4j, and S4l) and used a permutation test (1,000 runs; see middle of Figs. 2a–2d and Figs. S4c, S4e, S4k, and S4m) to identify regions that had a significant group difference (see right side of Figs. 2a–2d and Figs. S4h and S4p). The identified region in the feature map of each DNN layer for each trait contained faces that were most discriminative for ratings between individuals with SR-ASD and neurotypicals (note that an equal difference across faces could not lead to a discriminative region; in other words, a discriminative region could not simply result from the gross difference in ratings).

Using this approach, we identified faces that were judged most differently by participants with SR-ASD from neurotypicals for each trait. For example, we found that judgments of *competent* primarily differed in young, male, Caucasian faces (see Figs. 2a and 2e), whereas judgments of *youthful* primarily differed in African American faces as well as old, male, Caucasian faces (see Figs. 2d and 2e; for other discriminative regions across DNN layers, see Fig. S5a in the Supplemental Material). Therefore, this analysis systematically revealed what types of faces drove group differences in trait judgments from faces.

It is worth noting that different traits showed different discriminative faces (see Figs. 2a–2e and Fig. S5a). These discriminative faces mainly appeared in the intermediate and later DNN layers where facial features are abstracted toward semantic representations (see Fig. S5a). These results suggest that the different social trait judgments in participants with SR-ASD compared with neurotypicals may stem from different representations of more abstract facial features. We further quantified these group differences by correlating the similarity across faces (Kriegeskorte et al., 2008) between social trait ratings and DNN features (see the Method section). Again, we found that the group differences in trait judgments were primarily in the later DNN layers (see Figs. 2f and 2g and Fig. S5b), confirming that different social trait judgments in SR-ASD are driven by more abstract facial features.

We further linked this result to reduced specificity in social trait judgment (see Figs. 1f–1i). We found that faces in the discriminative regions were more likely to be at the extremes according to neurotypicals' ratings (i.e., top 10 identities and bottom 10 identities of the rank-ordered face identities; see Fig. 1f; χ^2 test: p = 3.04×10^{-6} for *competent*, p = .013 for *practical*, and $p < 10^{-20}$ for *feminine*). Therefore, the present analysis using DNN feature space further informs what types of faces are most likely to elicit a compressed range and thus reduced specificity in social trait judgment in SR-ASD.

Features within faces that contribute to different trait ratings in individuals with ASD compared with neurotypicals

Besides different types of faces, the ways in which individuals with ASD interpret cues within a face may also contribute to their different social trait judgments compared with neurotypicals. To understand which types of cues in a face may be more informative for participants with SR-ASD when making trait judgments compared with neurotypicals, we trained a DNN to predict participants' ratings of the faces using the face images as inputs (see Fig. S6a). A DNN was trained separately for each trait and participants with SR-ASD and neurotypicals, respectively (see Fig. S6b for model performance). We visualized the critical pixels in the face images that led to the correct prediction of social trait judgment using LRP (see the Method section).

We first confirmed that critical facial parts such as the eyes, mouth, and hair were important to predict social trait judgments (see Figs. 3a and 3b for examples and Fig. 3c for group summary). For example, for both groups, the eyes were important for judging *warm* and *strong*, and the mouth was important for judging *practical* and *youthful* (see Fig. 3c; note that critical facial parts are aligned across stimuli; Cao et al., 2021). We next revealed important regions of the face that were associated with group differences in the judgments for each trait (see Fig. 3d). Results showed that participants with SR-ASD relied less than neurotypicals on information (a) from the forehead when judging *warm*, *critical*, *practical*, and *strong*; (b) from the eyes when judging *practical* and *strong*; and (c) from the mouth when judging *warm*, *feminine*, and *strong*. Participants with SR-ASD used more information from the forehead and eyes when judging *youthful* and *charismatic* than neurotypicals. Together, by combining DNNs and a wide range of trait judgments, we discovered a nuanced relationship between facial features and the differences in social trait judgments between individuals with SR-ASD and neurotypicals.

Validation with well-characterized in-lab participants

The above results were based on online participants with SR-ASD. We next tested the validity of our findings with ratings from a sample of in-lab participants with

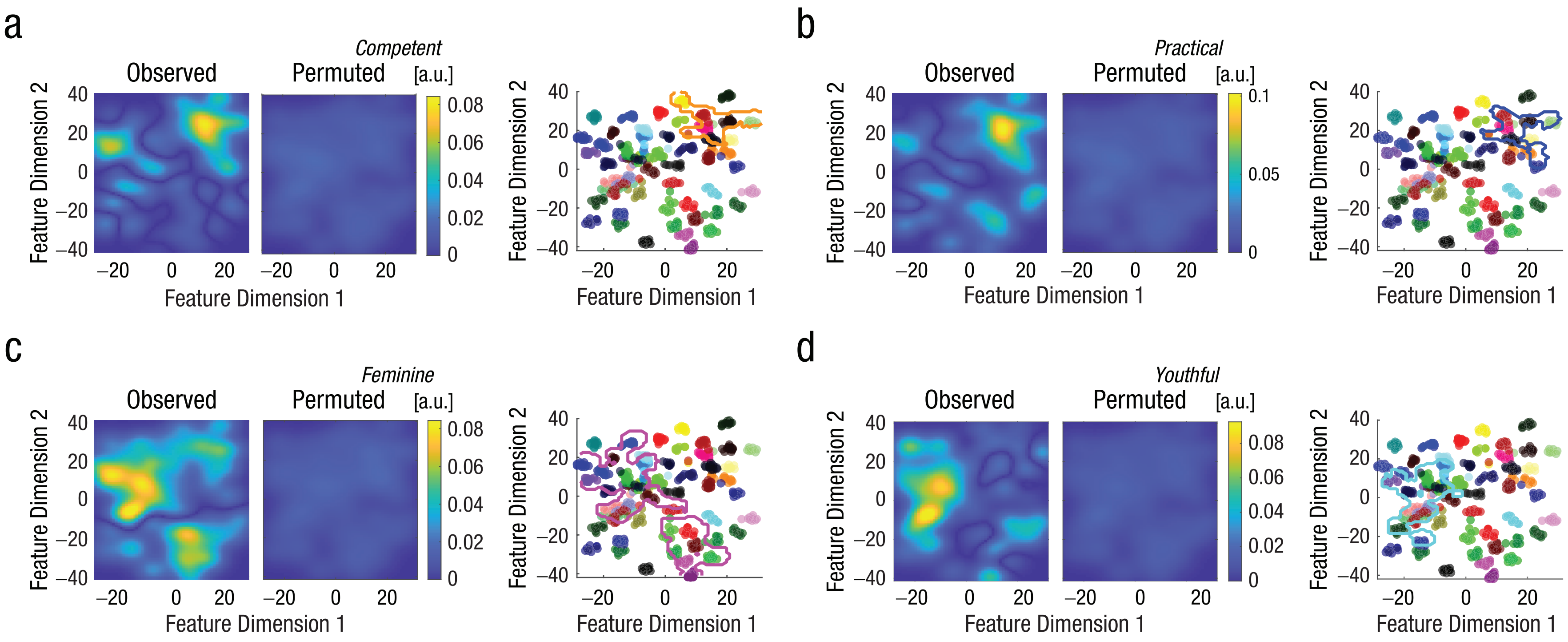

Fig. 2. *(continued on next page)*

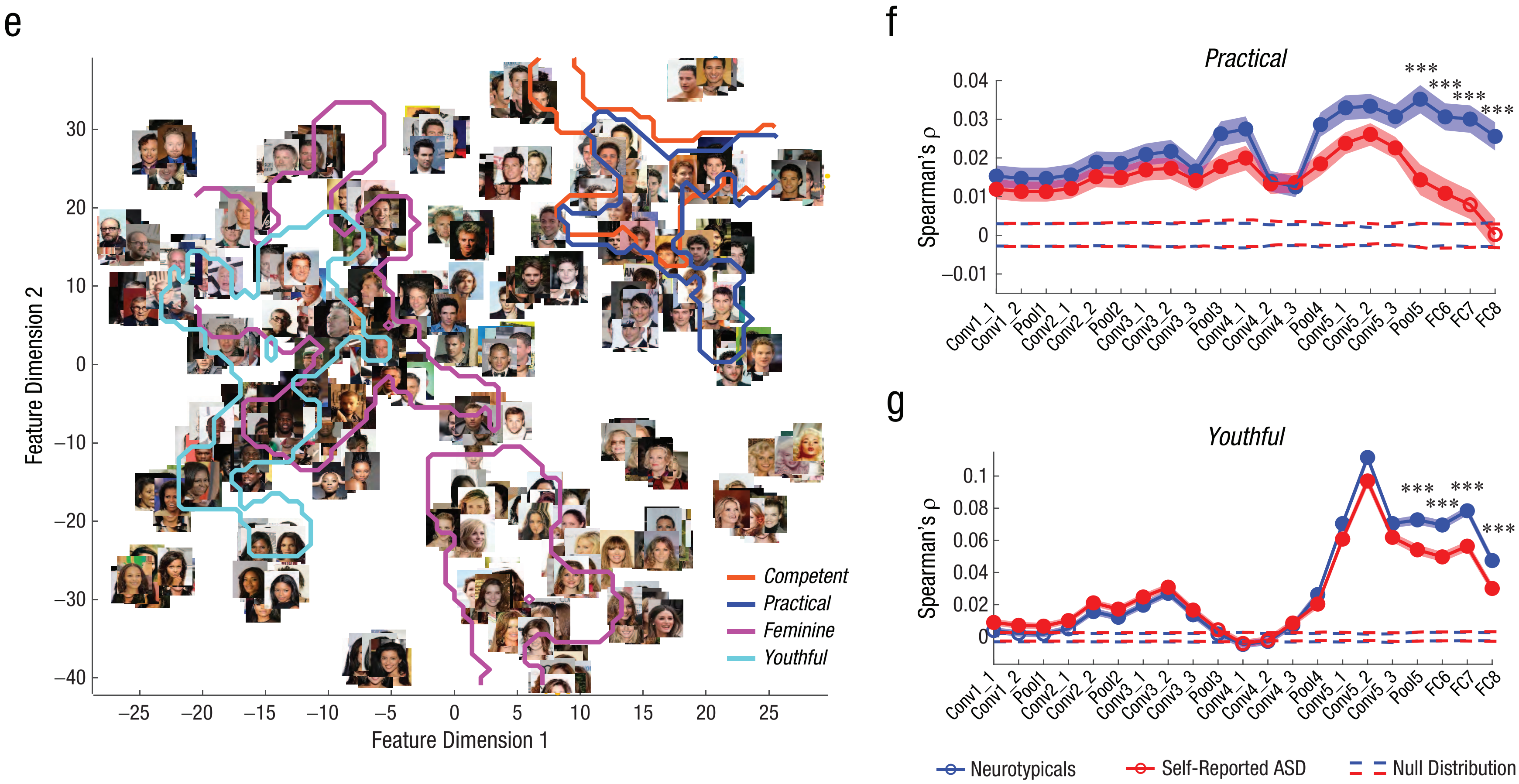

Fig. 2. Features across faces that contribute to different trait ratings in self-reported autism spectrum disorder (SR-ASD) compared with neurotypicals. (a–d) Estimation of the rating density and identification of the discriminative regions in the feature space. By comparing observed (left) and permuted (middle) difference in ratings between groups, we could identify a region in the feature space (right) where the difference in ratings was significant (discriminative regions). These regions contain faces that are most discriminative for ratings between individuals with SR-ASD and neurotypicals (delineated by the outlines; also shown in panel e). Color coding shows density in arbitrary units (a.u.). Each color in the scatterplot represents a different identity. (a) Trait *competent*. (b) Trait *practical*. (c) Trait *feminine*. (d) Trait *youthful*. (e) Discriminative regions in the face feature space constructed by t-distributed stochastic neighbor embedding for the deep neural network (DNN) fully connected layer FC6. All stimuli are shown in this space. The feature dimensions are in arbitrary units (a.u.). Outlines delineate the discriminative regions for each trait. (f, g) Representation similarity between social trait judgment ratings and DNN features for each DNN layer. Solid circles represent a significant above-chance correlation (permutation test: $p < .05$, Bonferroni correction across layers). Shaded area denotes standard deviation (± *SD*) across rating modules. Dashed line denotes standard deviation (± *SD*) across permutation runs. Asterisks indicate a significant difference between participants with SR-ASD and neurotypicals using permutation test (red for SR-ASD, blue for neurotypical). $^{***}p < .001$. (f) Trait *practical*. (g) Trait *youthful*.

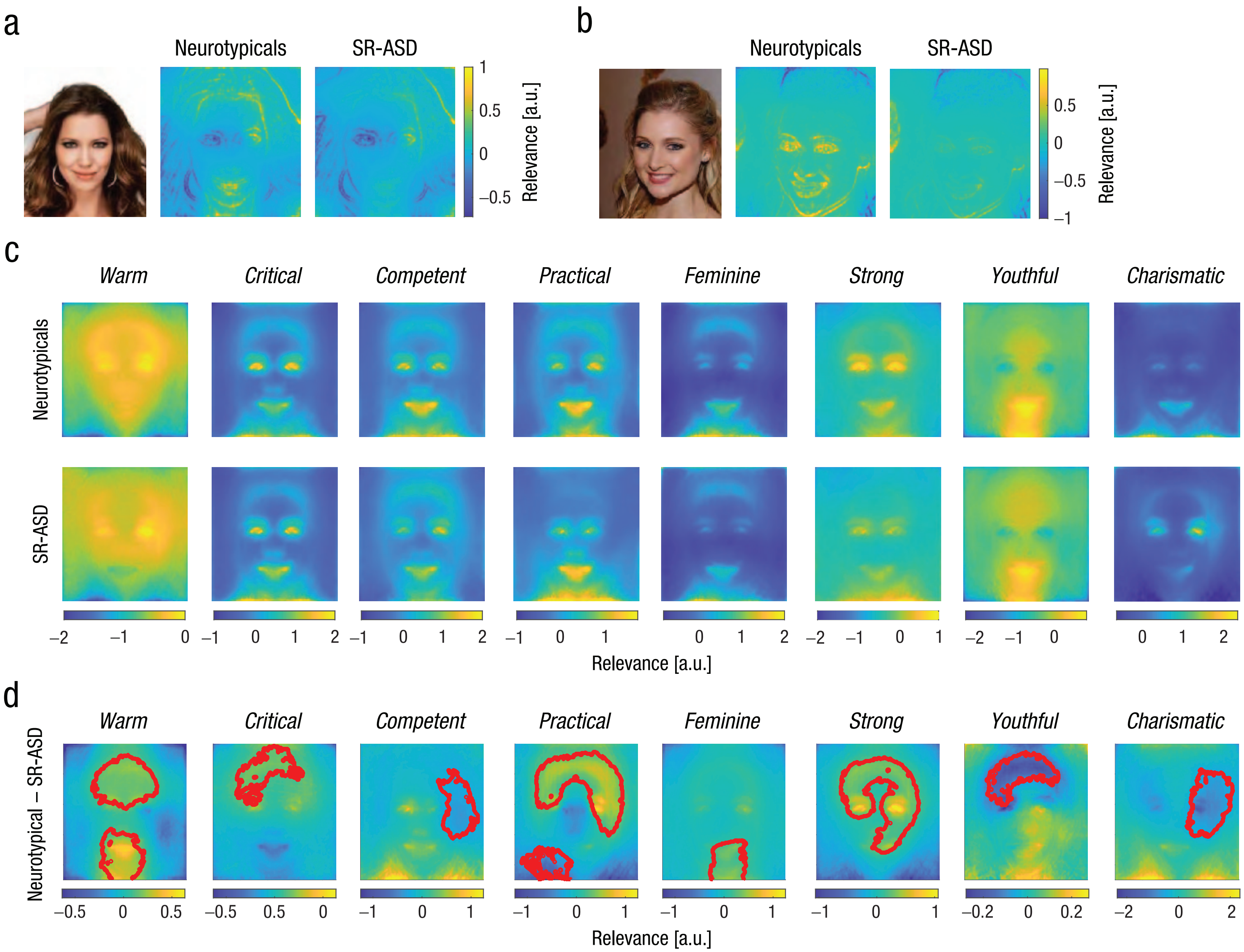

Fig. 3. Features within faces that contribute to different trait ratings in self-reported autism spectrum disorder (SR-ASD) compared with neurotypicals. Relevance of each pixel to classification was revealed using layer-wise relevance propagation (LRP). Color coding shows LRP values in arbitrary units (a.u.). Yellow pixels positively contributed to the classification, whereas blue pixels negatively contributed to the classification. (a, b) Two example faces and their corresponding LRP maps. (a) Trait *warm*. (b) Trait *strong*. (c) Average LRP maps for each trait and each group. Images from neurotypicals (upper). Images from participants with SR-ASD (lower). (d) Difference in LRP maps for each trait. Red contours show the regions with a significant difference between participants with SR-ASD and neurotypicals using two-tailed paired-samples *t* test ($p < 10^{-18}$; cluster size > 5% of all pixels).

confirmed ASD diagnosis (n = 27) and matched neurotypicals (n = 21). We replicated the findings that participants with ASD (a) shared the same psychological structure of trait judgments as neurotypicals (see Fig. 4a), (b) showed lower interrater consistency than neurotypicals (see Figs. 4b and 4c; see figure legend for statistics), and (c) rated individual traits differently from neurotypicals (see Fig. 4d; note that here neurotypicals showed more positive ratings; see the Discussion section); specifically, they showed reduced specificity in social trait judgments (see Figs. 4e–4h). Together, we validated our main findings from online participants in participants with confirmed ASD diagnosis and matched neurotypicals.

To directly compare the reduced specificity between samples (online vs. in-lab), we next used a two-way analysis of variance (participant group by sample) to test the difference between the top 10 and bottom 10 identities (note that here we used the rank order from online participants for both groups to have a direct comparison, but this rank order led to a similar result in in-lab participants; see Fig. 4h vs. Fig. 4i). We found a remarkably similar pattern of reduced specificity between samples (see Fig. 4i). Specifically, we found a main effect of sample for *charismatic*, $F(1, 1192)$ = 15.27, $p = 1.30 \times 10^{-5}$, η^2 = .016, and we confirmed the main effect of participant group for all traits: *warm*, $F(1, 1203)$ = 41.18, $p = 1.91 \times 10^{-9}$, η^2 = .03; *critical*, $F(1, 1187)$ = 8.06, p = .0046, η^2 = .006; *competent*, $F(1, 1157)$ = 64.80, $p = 2.04 \times 10^{-15}$, η^2 = .05; *practical*, $F(1, 1167)$ = 37.93, $p = 1.01 \times 10^{-9}$, η^2 = .03; *feminine*, $F(1, 1052)$ = 64.85, $p = 2.17 \times 10^{-15}$, η^2 = .06; *strong*, $F(1, 1201)$ = 9.75, p = .002, η^2 = .008; *youthful*, $F(1, 1212)$ = 56.47, $p = 1.11 \times 10^{-13}$, η^2 = .04; and *charismatic*, $F(1, 1192)$ = 13.18, $p = 2.96 \times 10^{-4}$, η^2 = .01. We also observed a significant interaction between participant group and sample for *competent*, $F(1, 1157)$ = 4.11, p = .04, η^2 = .003, and *charismatic*, $F(1, 1192)$ = 5.38, p = .02, η^2 = .004. These results indicate that judgment specificity was comparable between our online and in-lab samples. In both samples, participants with ASD demonstrated reduced specificity in their social trait judgments compared with neurotypicals.

Comparison with posed neutral faces

We derived the above results using complex, naturalistic face stimuli. How do individuals with ASD compared with neurotypicals make social trait judgments from simpler, controlled face stimuli? To address this question, we conducted a preregistered study (see the Method section) using posed photos of real people with neutral expressions from a previous study (Lin et al., 2021; see Fig. 5a for examples). First, we replicated that both participants with SR-ASD and neurotypicals shared the same comprehensive psychological dimensions underlying trait judgments (see Fig. 5b): The two groups shared the same number of optimal factors, and the four dimensions extracted from the two groups were highly similar (Tucker indices = 1.00, 0.99, 0.99, and 0.99). Second, we observed a reduced interrater consistency for *warm*, *feminine*, *strong*, and *youthful* in participants with SR-ASD (see Figs. 5c and 5d; see figure legend for statistics), consistent with the results using complex, naturalistic face stimuli (see Figs. 1c and 1d). Third, we observed a significant group difference in aggregate ratings only for the trait *critical* (see Fig. 5e; see Fig. S2c for a comparison) and reduced specificity in ratings only for *strong* and *youthful* (see Figs. 5f–5i; see Figs. 1f–1i for a comparison). These findings showed that the social trait judgments that individuals with SR-ASD made for simpler, controlled faces were less different from neurotypicals compared with judgments for more complex, naturalistic faces (note that our neurotypicals' ratings in the present study were highly correlated with those in the previous study; Lin et al., 2021; see Fig. S7 in the Supplemental Material). These findings suggest that the variations in face stimuli such as facial expressions, backgrounds, and familiarity may play an important role in shaping how individuals with SR-ASD make social judgments of others in more naturalistic contexts and may explain prior discrepant findings in the comparison between individuals with ASD and neurotypicals (see the Discussion section).

Discussion

We conducted a comprehensive investigation of how individuals with ASD make social trait judgments of faces compared with neurotypicals. Using a representative set of traits and naturalistic face stimuli, as well as large samples of online participants and well-characterized in-lab participants, we showed that individuals with ASD and neurotypicals shared the same psychological structure underlying social trait judgments from faces. However, participants with ASD showed reduced interrater consistency and different ratings for individual traits compared with neurotypicals. We applied neural network modeling to show that these discrepant ratings were explained by discrepant judgments for certain types of faces and discrepant utilization of features within a face. These group differences persisted but were less severe when social trait judgments were made for simpler and more controlled face stimuli. Together, these findings advance a comprehensive understanding of the psychological structure and computational basis of social trait judgments from faces in individuals with ASD. These results provide initial insights into how different face processing in individuals with ASD may be linked to their different social behaviors.

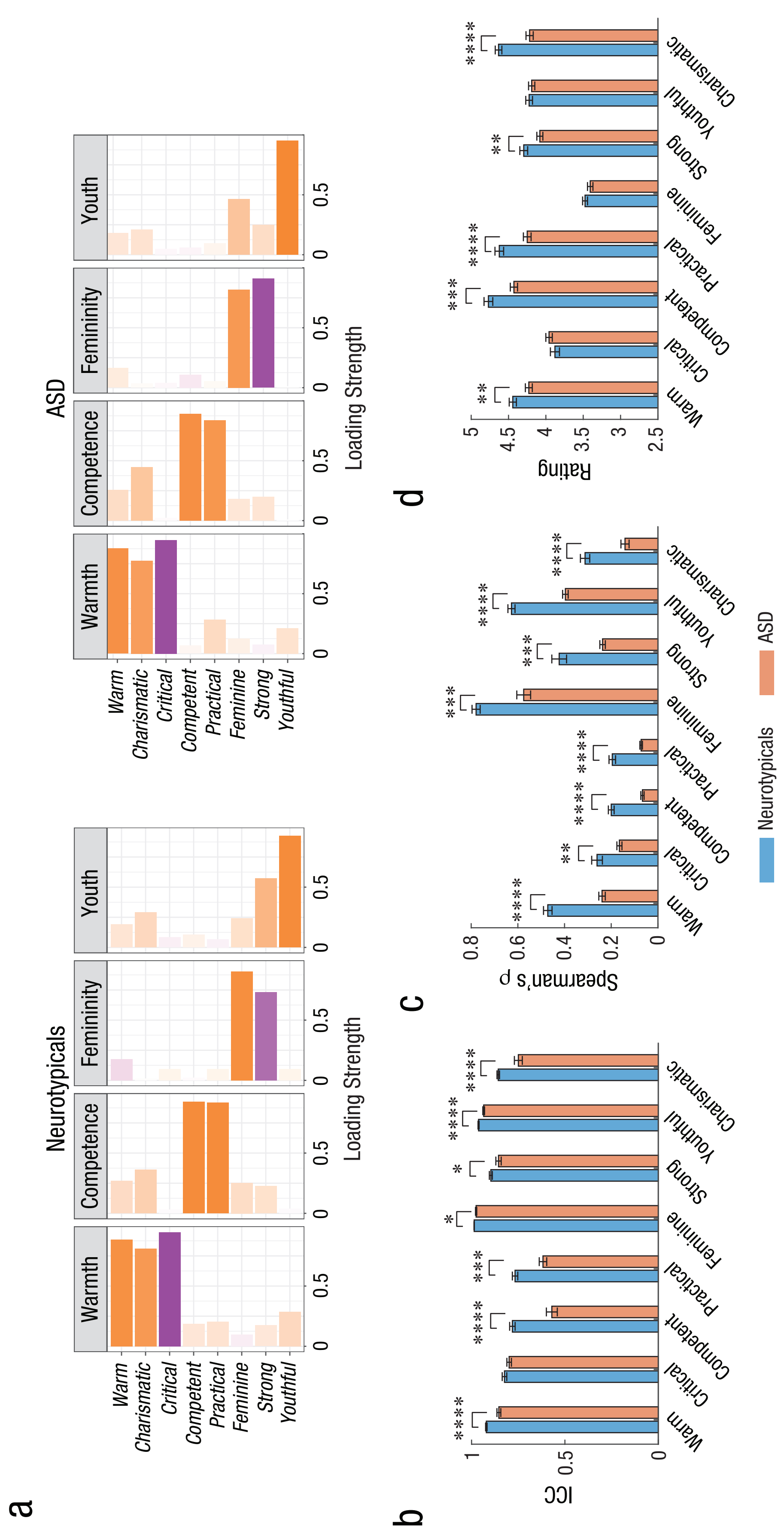

Fig. 4. *(continued on next page)*

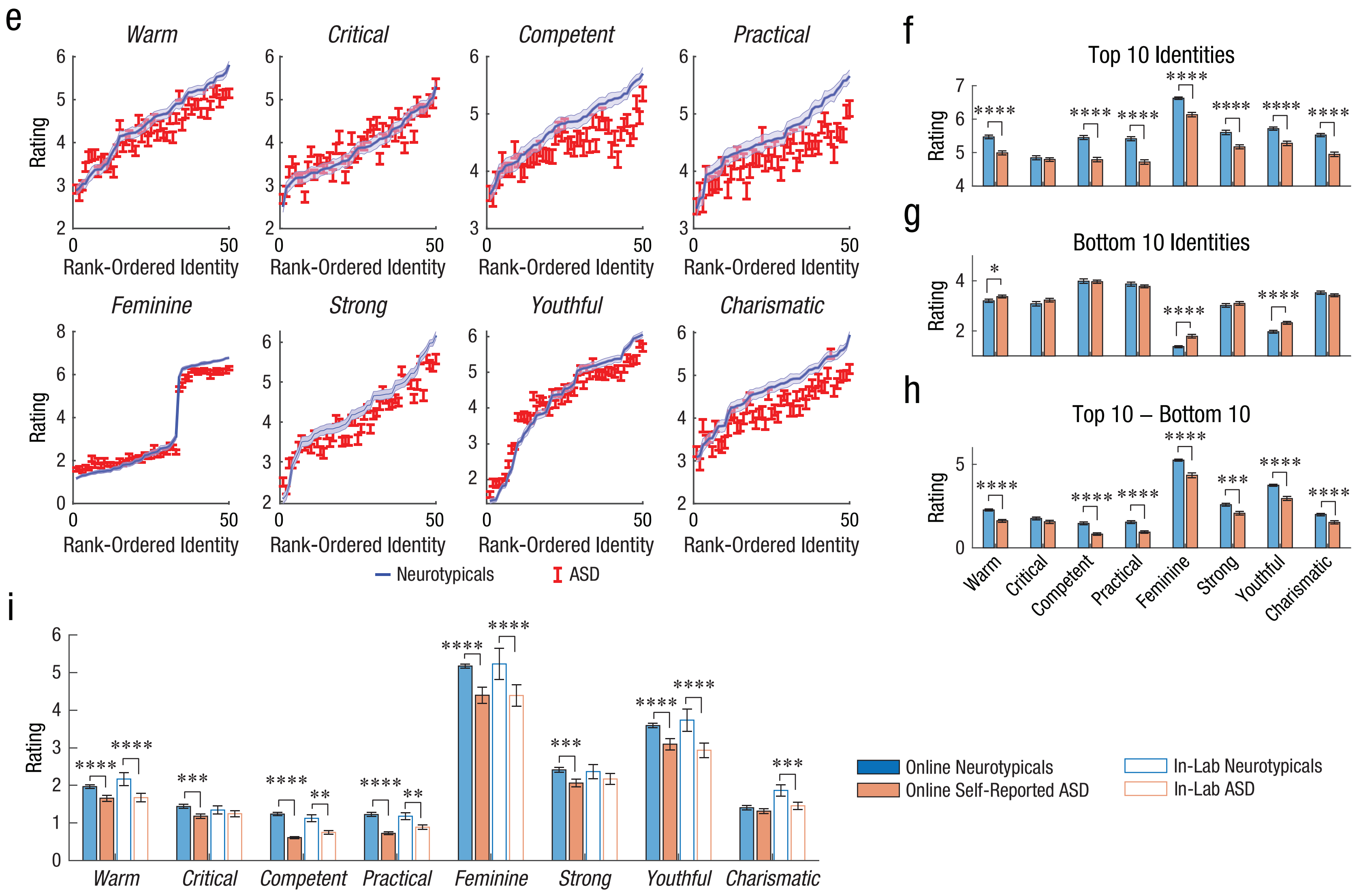

Fig. 4. Validation with in-lab participants. (a) Principal component analysis loadings of social traits on the first four principal components. (b, c) Interrater consistency. Participants with ASD demonstrated lower interrater consistency for most of the traits: *warm* (two-tailed paired-samples *t* test across 10 rating modules), intraclass correlation coefficient (ICC): $t(9) = 8.77$, $p = 1.05{\times}10^{-5}$, $d = 2.67$, 95% confidence interval (CI) = [0.05, 0.08]; Spearman: $t(9) = 13.55$, $p = 2.72{\times}10^{-7}$, $d = 4.50$, 95% CI = [0.19, 0.27]; *critical*, Spearman: $t(9) = 4.57$, $p = .001$, $d = 1.68$, 95% CI = [0.05, 0.14]; *competent*, ICC: $t(9) = 7.63$, $p = 3.22{\times}10^{-5}$, $d = 2.87$, 95% CI = [0.15, 0.27]; Spearman: $t(9) = 10.50$, $p = 2.39{\times}10^{-6}$, $d = 4.17$, 95% CI = [0.10, 0.16]; *practical*, ICC: $t(9) = 7.12$, $p = 5.54\ {\times}10^{-5}$, $d = 2.68$, 95% CI = [0.10, 0.20]; Spearman: $t(9) = 8.08$, $p = 2.05{\times}10^{-5}$, $d = 3.64$, 95% CI = [0.09, 0.16]; *feminine*, ICC: $t(9) = 2.80$, $p = .02$, $d = 1.32$, 95% CI = [0.002, 0.02]; Spearman: $t(9) = 5.21$, $p = 5.78{\times}10^{-4}$, $d = 2.64$, 95% CI = [0.12, 0.29]; *strong*, ICC: $t(9) = 2.90$, $p = .02$, $d = 1.11$, 95% CI = [0.009, 0.07]; Spearman: $t(9) = 5.32$, $p = 4.80{\times}10^{-4}$, $d = 2.45$, 95% CI = [0.11, 0.26]; *youthful*, ICC: $t(9) = 6.76$, $p = 8.26{\times}10^{-5}$, $d = 2.21$, 95% CI = [0.02, 0.04]; Spearman: $t(9) = 12.34$, $p = 6.05{\times}10^{-7}$, $d = 5.30$, 95% CI = [0.19, 0.27]; and *charismatic*, ICC: $t(9) = 6.95$, $p = 6.67{\times}10^{-5}$, $d = 2.14$, 95% CI = [0.07, 0.14]; Spearman: $t(9) = 12.47$, $p = 5.53{\times}10^{-7}$, $d = 2.86$, 95% CI = [0.14, 0.20]. (d) For aggregate ratings, neurotypicals had a significantly higher rating for *warm*, $t(379) = 3.12$, $p = .002$, $d = 0.32$, 95% CI = [0.08, 0.35]; *competent*, $t(350) = 4.28$, $p = 2.41{\times}10^{-5}$, $d = 0.47$, 95% CI = [0.18, 0.50]; *practical*, $t(365) = 4.41$, $p = 4.79{\times}10^{-6}$, $d = 0.49$, 95% CI = [0.22, 0.53]; *strong*, $t(316) = 3.30$, $p = .001$, $d = 0.34$, 95% CI = [0.04, 0.18]; and *charismatic*, $t(373) = 5.99$, $p = 4.95{\times}10^{-9}$, $d = 0.62$, 95% CI = [0.28, 0.55]. (e) Ratings for each face identity rank-ordered by mean ratings from neurotypicals (red for ASD, blue for neurotypical). Error bars and error shades denote standard errors of the mean (± *SEM*) across rating modules. (f) Average ratings for the 10 identities with the highest ratings from neurotypicals. (g) Average ratings for the 10 identities with the lowest ratings from neurotypicals. (h, i) Difference in ratings between the top 10 and bottom 10 identities. Legend conventions as in Fig. 1.

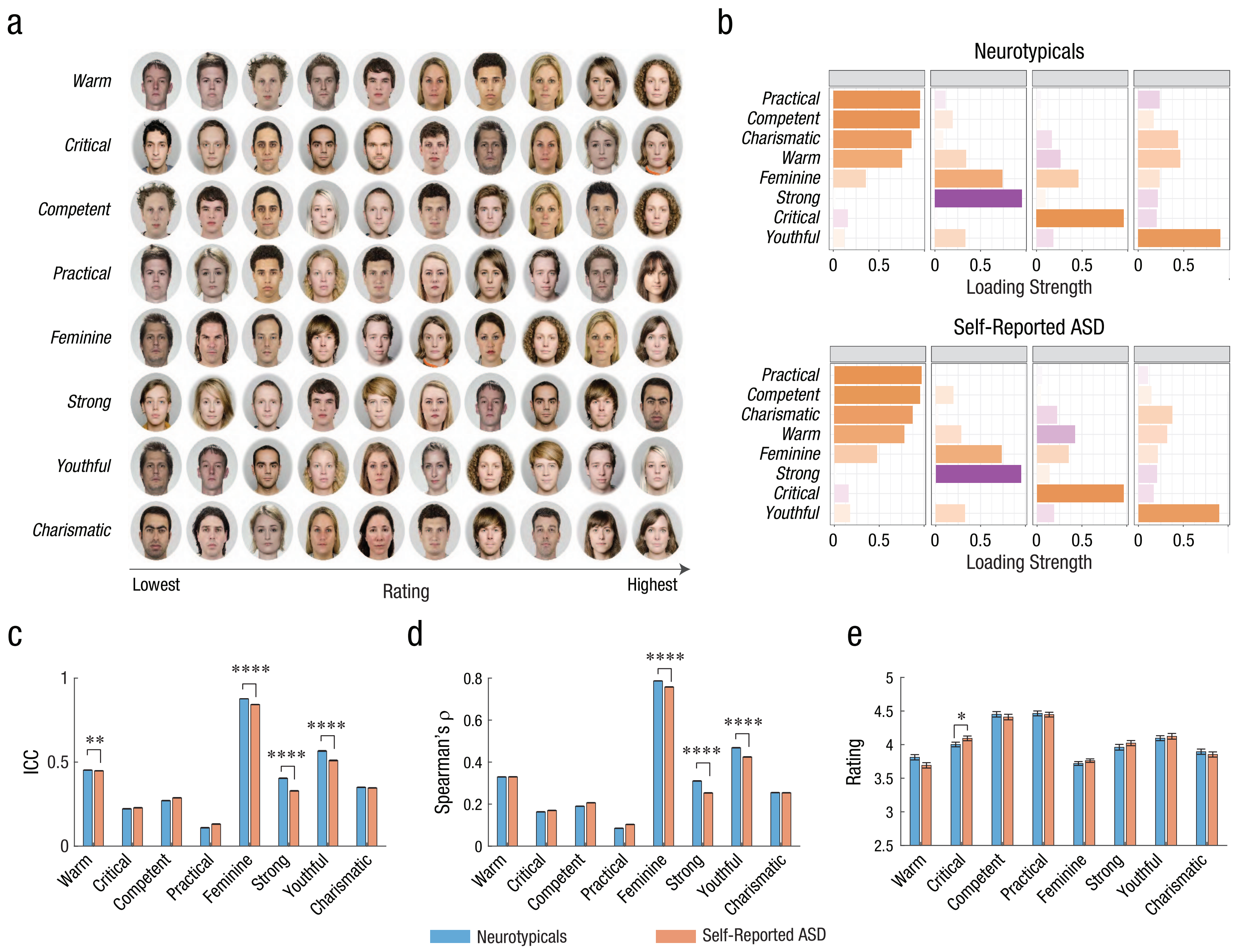

Fig. 5. *(continued on next page)*

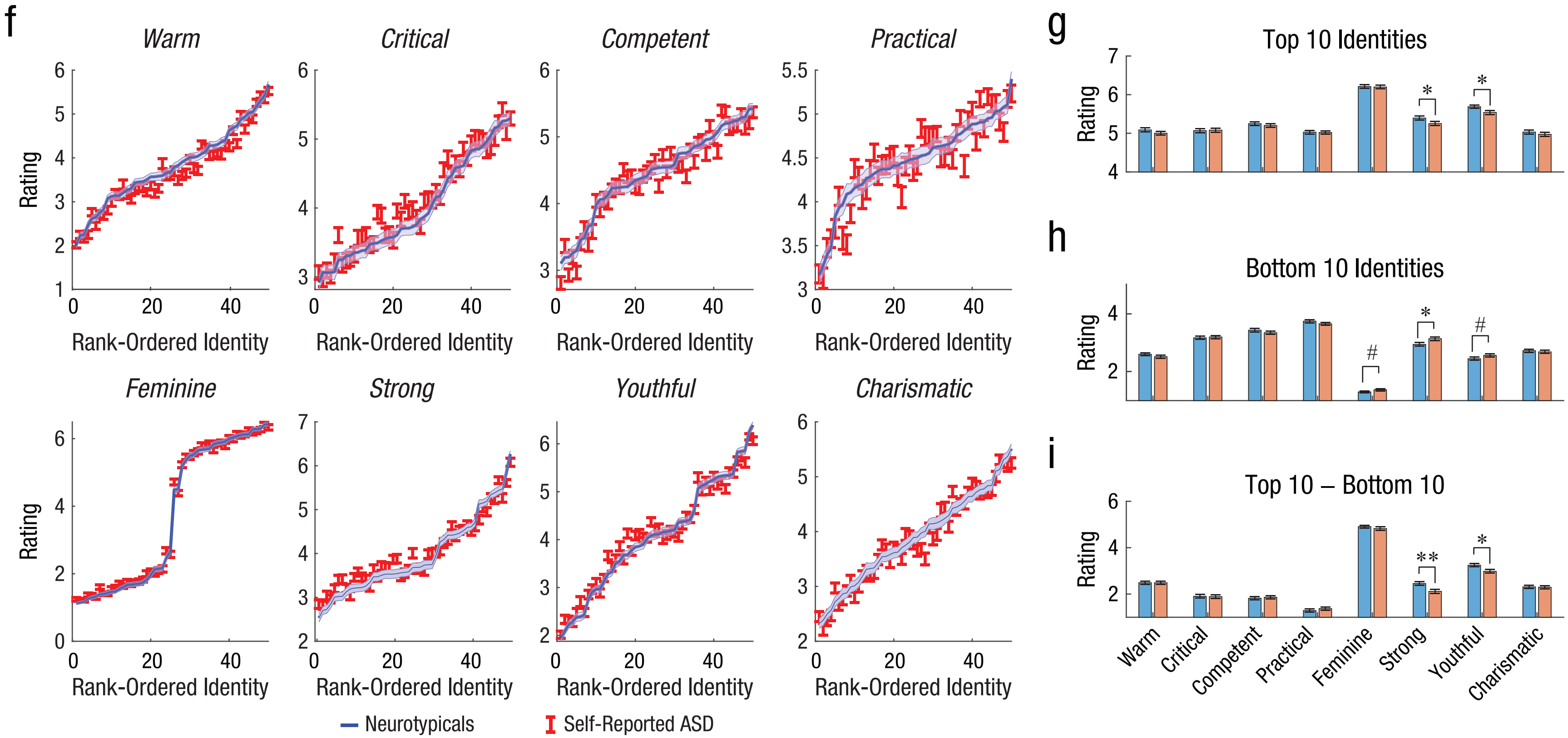

Fig. 5. Validation with an independent sample of participants using unfamiliar face stimuli. (a) Example stimuli ranked by average ratings from neurotypicals for each social trait. (b) Principal component analysis loadings of social traits on the first four principal components. (c, d) Interrater consistency. (c) Intraclass correlation coefficient (ICC). Participants with self-reported autism spectrum disorder (SR-ASD) demonstrated a lower ICC for *warm* (one-tailed two-sample *t* test across participant pairs), *t*(60769) = 2.40, *p* = .0082, *d* = 0.020, lower bound of 95% confidence interval (CI) = 0.0014; *feminine*, *t*(59782) = 25.6, $p < 10^{-10}$, *d* = 0.21, lower bound of 95% CI = 0.03; *strong*, *t*(61508) = 27.0, $p < 10^{-10}$, *d* = 0.22, lower bound of 95% CI = 0.07; and *youthful*, *t*(60516) = 15.5, $p < 10^{-10}$, *d* = 0.13, lower bound of 95% CI = 0.05. (d) Spearman's correlation coefficient (ρ). Participants with SR-ASD demonstrated a lower correlation coefficient for *feminine*, *t*(59782) = 21.4, $p < 10^{-10}$, *d* = 0.17, lower bound of 95% CI = 0.03; *strong*, *t*(61504) = 28.7, $p < 10^{-10}$, *d* = 0.23, lower bound of 95% CI = 0.05; and *youthful*, *t*(60516) = 23.2, $p < 10^{-10}$, *d* = 0.19, lower bound of 95% CI = 0.04. (e) For aggregate ratings, participants with SR-ASD had a significantly higher rating for *critical* (one-tailed two-sample *t* test), *t*(489) = 1.83, *p* = .034, *d* = 0.09, higher bound of 95% CI = 0.009. (f) Ratings for each face identity rank-ordered by mean ratings from neurotypicals. (g) Average ratings for the 10 identities with the highest ratings from neurotypicals. (h) Average ratings for the 10 identities with the lowest ratings from neurotypicals. (i) Difference in ratings between the top 10 and bottom 10 identities. Legend conventions as in Fig. 1.

Generalizable psychological dimensions underlying social trait judgments

It remains debated what fundamental psychological dimensions underlie social cognition. Whereas some researchers have argued that warmth and competence summarize most of the variance in social cognition (Fiske et al., 2007), a range of other theories has been proposed (Sutherland & Young, 2022; Tamir et al., 2016). One prior study using the most comprehensive set of English trait words to date showed that the hundreds of different trait judgments that people make from faces could be summarized by four dimensions: warmth, competence, femininity, and youth (Lin et al., 2021). But even that study was limited to posed photos of neutral faces of White individuals. Here, using naturalistic face images of diverse individuals, we replicated the four-dimensional framework in both neurotypicals and individuals with ASD. Importantly, much research has shown that factors such as facial expressions, race, and contexts play an important role in shaping how people make social trait judgments from faces (Todorov, 2017; Zebrowitz et al., 2010). Our results indicate that these factors do not significantly change the correlational structure between different trait judgments.

The correlational structure between trait judgments has been shown to be extremely flexible. It is shaped by factors such as perceivers' understanding of the conceptual relations between the trait words and the perceivers' experiential sampling of the personality structure in their local environment (Oh et al., 2022; Stolier et al., 2018). Individuals with ASD are known to have impairment in verbal ability (i.e., speech, verbal IQ, communication ability, and verbal fluency; Oliveras-Rentas et al., 2012; Spek et al., 2009) and reduction in social interactions (Chawarska et al., 2012; Jahr et al., 2007; Shic et al., 2020), even for participants with ASD who have typical intellectual functioning. Surprisingly, we found that the correlational structure across trait judgments of faces was highly similar between neurotypicals and individuals with ASD. These findings suggest that individuals with ASD and neurotypicals may share similar understanding of the semantic relationship between different social trait descriptions and similar understanding of the personality structure in everyday life. Altogether, our results suggest that all of the four dimensions—warmth, competence, femininity, and youth—may be fundamental to social cognition.

Within- and between-groups variations of social trait judgments

In line with prior findings on the substantial heterogeneity among individuals with ASD (Happe et al., 2006), we showed that the between-subjects consensus in social trait judgments of faces among individuals with ASD was lower than that in neurotypicals. At least three factors may contribute to this increased heterogeneity. First, there may be increased perceptual heterogeneity in ASD, such as more diverse patterns of feature utilization among individuals with ASD, which could be formally tested in future research with dense individual data using the critical pixel analysis pipeline that we provided here. Second, there may be increased conceptual heterogeneity in ASD, such as more diverse understanding of the trait words among individuals with ASD, although in our study, we have provided a one-sentence definition of the trait word for every participant. Third, there may be increased mapping heterogeneity in ASD, such as different mappings between facial features and social trait impressions. The analysis pipeline of DNN features in regressions that we provided here could be flexibly applied to comparing models trained on dense individual data, which will provide insights into this possibility.

Prior research on the variation of social trait judgments from faces focused on three factors: targets, perceivers, and contexts (Hehman et al., 2017). It has been shown that these three factors mainly influence social trait judgments independently (Xie et al., 2022). However, our findings indicate that these factors may interact. For instance, the group differences that we found between participants with SR-ASD and neurotypicals (see Fig. 1) were not merely due to baseline perceiver differences (e.g., participants with SR-ASD rated all faces on a trait higher than neurotypicals). Instead, perceiver differences were most prominent for specific types of faces (targets): the faces that received the most extreme ratings from neurotypicals (see Figs. 1g–1i), unfamiliar faces (see Figs. S2c and S2d), same-race faces (see Figs. S2e and S2f), same-sex faces (see Figs. S2g and S2h), and trait-dependent subsets of faces (see Fig. 2). These findings suggest that the different social evaluation of faces may be a result of different conceptual associations between social traits and social groups (i.e., social stereotypes) in individuals with SR-ASD compared with neurotypicals (e.g., the most stereotypical faces for neurotypicals received less extreme ratings in SR-ASD; see Fig. 1i).

We revealed that social trait judgments from participants with SR-ASD were associated with different critical pixels of the face compared with neurotypicals (see Fig. 3). These findings are consistent with the large eye-tracking literature showing that people with ASD view faces differently (Kliemann et al., 2010; Neumann et al., 2006; Pelphrey et al., 2002; Spezio et al., 2007). For example, people with ASD show an increased tendency to saccade away from the eye region of faces when information is present in those regions (Spezio et al., 2007) but instead have an increased preference to fixate the location of the mouth (Neumann et al., 2006). Furthermore, people with ASD demonstrate active avoidance of fixating the eyes in faces, which in turn influences recognition performance of emotions (Kliemann et al., 2010). In particular, our recent study has shown that the neural substrates underlying fixations on faces are related to perceived social trait judgments (Cao et al., 2021). Therefore, different social trait judgments in ASD may stem from different eye movement patterns when viewing faces. Furthermore, different social trait judgments in ASD compared with neurotypicals may be attributed to differential neural face representation in the amygdala and hippocampus (Cao et al., 2022).

Reconciling prior literature comparing social trait judgments between individuals with ASD and neurotypicals

Prior studies have shown inconsistent results regarding whether individuals with ASD make social trait judgments of faces similar to neurotypicals. Earlier studies using photographs of real people have shown that people with ASD evaluated social traits such as facial trustworthiness (Adolphs et al., 2001; Forgeot d'Arc et al., 2016) differently from neurotypicals. Other studies using computer-generated faces have shown that adults with ASD judge trustworthiness and dominance (Latimier et al., 2019) as well as a variety of seven different traits (Lindahl, 2017) similarly to neurotypicals. Without a clear understanding of whether individuals with ASD make different social trait judgments from faces, we are uncertain whether these judgments may be linked to the different social behaviors observed in ASD.

Our results provided initial empirical evidence that the complexity of the face stimuli plays an important role in determining whether individuals with ASD and neurotypicals make different social trait judgments from faces. In addition to using a diverse set of naturalistic face stimuli that varied in factors such as facial expressions, pose, gaze, and background in our main study, we conducted a preregistered study using an independent set of controlled face stimuli that were neutral and frontal, with a direct gaze and a uniform background. We tested three preregistered hypotheses: the overall psychological structure, interrater consistency, and rating specificity. We found that the trait judgments made by individuals with SR-ASD in response to the more controlled facial stimuli (see Fig. 5) were more similar to those of neurotypicals than to judgments made in response to naturalistic faces (see Fig. 1), although differences persisted for some traits. Our findings suggest that the discrepant results in prior literature may be due to the variation of face stimuli: More differences between individuals with ASD and neurotypicals will be observed when more complex, naturalistic stimuli are used. These results suggest that the greater differences in social trait judgments from faces between individuals with ASD and neurotypicals in naturalistic contexts may help explain the differences in social behavior observed between these two groups in real-world interactions.

Limitations of our study

Although our study had advantages (see the Supplemental Material), several limitations of our designs constrained the generalizability of our conclusions.

First, we collected data from a large online sample and confirmed that participants with SR-ASD had significantly high AQ and SRS scores. This large amount of data makes it possible to train more complicated models (e.g., allowing nonlinearity) to provide new insights into how features within and across faces contribute to atypical trait judgment in ASD. However, the validity of AQ and SRS as indications of clinical ASD is still under debate (Hadad & Yashar, 2022). The online sample with SR-ASD may also include a wide range of autism severity, although it is most likely that our participants have typical intellectual functioning. Specifically, the mean AQ for online participants with SR-ASD was lower than that of in-lab participants with ASD and might be below the cutoff score of AQ for ASD (note that the cutoff score of AQ is debatable), whereas the mean AQ for online neurotypicals was higher than that of in-lab neurotypicals. Therefore, online participants may exhibit less variation in autism severity compared with in-lab participants, although a greater difference in ratings was observed. Future research is needed to investigate social trait judgments as a function of autism diagnoses and stratify participants on the basis of autism severity.

Second, using an independent sample of in-lab participants, we successfully replicated three key results: the intact overall dimensional structure, the reduced interrater consistency, and the reduced rating specificity in SR-ASD. However, the results regarding the magnitude of trait ratings in ASD compared with neurotypicals show less consistency between the two samples. Specifically, in the main online experiment, participants with SR-ASD had more positive ratings, whereas in the replication in-lab experiment, participants with ASD had more negative ratings. These discrepancies in the results may be attributed to various factors, including the different participant compositions (see the Method section and Table 1), potential sex differences in social trait judgment (Bosak et al., 2011; Wallach & Kogan, 1959), differing functioning levels of participants, the specificity of ASD diagnoses (as mentioned above), and variations in survey modes. Furthermore, these differences were likely influenced by reduced rating specificity, meaning that depending on the variations at the extremes, the grand average could show either a positive or a negative difference between the groups. In other words, relying solely on the grand average might oversimplify the comparison between the groups. Future research that quantifies comprehensive differences between samples may provide further insights into the extent and nature of the differences in the magnitude of various trait judgments between individuals with ASD and neurotypicals.

Third, we attempted to improve generalizability by sampling social traits comprehensively along core dimensions and using both ambient face stimuli and photos taken in more controlled conditions. However, these designs do not account for all possible factors that may be associated with trait judgment differences between individuals with ASD and neurotypicals. For instance, all of our face stimuli were static, whereas in real life, people usually see others move their faces dynamically. Future research using dynamic face stimuli will inform how atypical trait judgments from faces might have behavioral consequences for real-life social interactions in ASD.

We further discuss possible caveats of our study as well as future directions in the Supplemental Material.

Transparency

Action Editor: Daniela Schiller
Editor: Patricia J. Bauer
Author Contributions

Chujun Lin and Shuo Wang contributed equally to this study.

Runnan Cao: Conceptualization; Formal analysis; Visualization; Writing – original draft.

Na Zhang: Formal analysis; Software.

Hongbo Yu: Formal analysis; Writing – original draft; Writing – review & editing.

Paula J. Webster: Data curation.

Lynn K. Paul: Data curation.

Xin Li: Formal analysis; Supervision.

Chujun Lin: Data curation; Formal analysis; Writing – original draft; Writing – review & editing.

Shuo Wang: Conceptualization; Supervision; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the McDonnell Center for Systems Neuroscience, Air Force Office of Scientific Research (FA9550-21-1-0088), National Science Foundation (BCS-1945230, IIS-2114644), and National Institutes of Health (R01MH129426). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Open Practices

All data and code have been made publicly available via the Open Science Framework and can be accessed at osf.io/e2kc6/. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Na Zhang https://orcid.org/0000-0002-8845-5055
Hongbo Yu https://orcid.org/0000-0002-3384-7772
Shuo Wang https://orcid.org/0000-0003-2562-0225

Acknowledgments

We thank Ralph Adolphs for valuable comments.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976231192236

References

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience*, *13*, 232–240.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P.-J. (2019). iNNvestigate neural networks! *Journal of Machine Learning Research*, *20*, 1–8.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B: Methodological*, *57*, 289–300.

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences*, *19*, 421–422.

Bosak, J., Sczesny, S., & Eagly, A. H. (2011). The impact of social roles on trait judgments: A critical reexamination. *Personality and Social Psychology Bulletin*, *38*, 429–440.

Cao, R., Li, X., Brandmeir, N. J., & Wang, S. (2021). Encoding of facial features by single neurons in the human amygdala and hippocampus. *Communications Biology*, *4*, Article 1394. https://doi.org/10.1038/s42003-021-02917-1

Cao, R., Lin, C., Hodge, J., Li, X., Todorov, A., Brandmeir, N. J., & Wang, S. (2022). A neuronal social trait space for first impressions in the human amygdala and hippocampus. *Molecular Psychiatry*, *27*, 3501–3509.

Cao, R., Wang, J., Lin, C., De Falco, E., Peter, A., Rey, H. G., DiCarlo, J., Todorov, A., Rutishauser, U., Li, X., Brandmeir, N. J., & Wang, S. (2020). *Feature-based encoding of face identity by single neurons in the human medial temporal lobe*. bioRxiv. https://doi.org/10.1101/2020.09.01.278283

Charalambides, N. (2021). We recently went viral on TikTok—here's what we learned. https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned

Chawarska, K., Macari, S., & Shic, F. (2012). Context modulates attention to social scenes in toddlers with autism. *Journal of Child Psychology and Psychiatry*, *53*, 903–913.

Chelnokova, O., Laeng, B., Eikemo, M., Riegels, J., Løseth, G., Maurud, H., Willoch, F., & Leknes, S. (2014). Rewards of beauty: The opioid system mediates social motivation in humans. *Molecular Psychiatry*, *19*, 746–747.

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, *25*, 1132–1139.

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale: SRS-2*. Western Psychological Services.

DeBruine, L., & Jones, B. (2017). Face Research Lab London Set (Version 5). figshare. https://doi.org/10.6084/m9.figshare.5047666.v5

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference* on *Computer Vision* and *Pattern Recognition* (pp. 248–255). IEEE.

de Wit, T. C. J., Falck-Ytter, T., & von Hofsten, C. (2008). Young children with autism spectrum disorder look differently at positive versus negative emotional faces. *Research in Autism Spectrum Disorders*, *2*, 651–659.

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*, Article eabl8913. https://doi.org/10.1126/sciadv.abl8913

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*, 77–83.

Forgeot d'Arc, B., Ramus, F., Lefebvre, A., Brottier, D., Zalla, T., Moukawane, S., Amsellem, F., Letellier, L., Peyre, H., Mouren, M.-C., Leboyer, M., & Delorme, R. (2016). Atypical social judgment and sensitivity to perceptual cues in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *46*, 1574–1581.

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., Khuvis, S., Herrero, J. L., Irani, M., Mehta, A. D., & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*, Article 4934. https://doi.org/10.1038/s41467-019-12623-6

Hadad, B.-S., & Yashar, A. (2022). Sensory perception in autism: What can we learn? *Annual Review of Vision Science*, *8*, 239–264.

Hamermesh, D. S. (2011). *Beauty pays*. Princeton University Press.

Happe, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience*, *9*, 1218–1220.

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

Hester, N., Xie, S. Y., & Hehman, E. (2021). Little between-region and between-country variance when people form impressions of others. *Psychological Science*, *32*, 1907–1917.

Hinton, G. E., & Roweis, S. T. (2002). Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems* (pp. 857–864). MIT Press. https://dl.acm.org/doi/10.5555/2968618.2968725

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 9737–9748). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3454287.3455160

Hugenberg, K., Young, S. G., Sacco, D. F., & Bernstein, M. J. (2011). Social categorization influences face perception and face memory. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *Oxford Handbook of Face Perception* (pp. 245–262). Oxford University Press.

Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, Module 4: Revised algorithm and standardized severity scores. *Journal of Autism and Developmental Disorders*, *44*, 1996–2012.

Jahr, E., Eikeseth, S., Eldevik, S., & Aase, H. (2007). Frequency and latency of social interaction in an inclusive kindergarten setting: A comparison between typical children and children with autism. *Autism*, *11*, 349–363.

Kar, K. (2022). A computational probe into the behavioral and neural markers of atypical facial emotion processing in autism. *The Journal of Neuroscience*, *42*, 5115–5126.

Kar, K., Kornblith, S., & Fedorenko, E. (2022). Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence*, *4*, 1065–1067.

Keles, U., Lin, C., & Adolphs, R. (2021). A cautionary note on predicting social judgments from faces with deep neural networks. *Affective Science*, *2*, 438–454.

Kliemann, D., Dziobek, I., Hatri, A., Steimke, R., & Heekeren, H. R. (2010). Atypical reflexive gaze patterns on emotional faces in autism spectrum disorders. *The Journal of Neuroscience*, *30*, 12281–12287.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. https://doi.org/10.3389/neuro.06.004.2008

Latimier, A., Kovarski, K., Peyre, H., Fernandez, L. G., Gras, D., Leboyer, M., & Zalla, T. (2019). Trustworthiness and dominance personality traits' judgments in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *49*, 4535–4546.

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*, 574–589.

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications*, *12*, Article 5168. https://doi.org/10.1038/s41467-021-25500-y

Lindahl, C. (2017). *Judgments of social dimensions of faces in individuals with high-functioning autism*. [Master's thesis, Stockholm University]. http://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1111755&dswid=1460

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 3730–3738).

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, *19*, 185–212.

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

Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., Christensen, D. L., Wiggins, L. D., Pettygrove, S., Andrews, J. G., Lopez, M., Hudson, A., Baroud, T., Schwenk, Y., White, T., Rosenberg, C. R., Lee, L.-C., Harrington, R. A., Huston, M., . . . Dietz, P. M. (2020). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016. *MMWR Surveillance Summaries*, *69*, 1–12.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Neumann, D., Spezio, M. L., Piven, J., & Adolphs, R. (2006). Looking you in the mouth: Abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Social Cognitive and Affective Neuroscience*, *1*, 194–202.

Oh, D., Martin, J. D., & Freeman, J. B. (2022). Personality across world regions predicts variability in the structure of face impressions. *Psychological Science*, *23*, 1240–1256.

Oliveras-Rentas, R. E., Kenworthy, L., Roberson, R. B., Martin, A., & Wallace, G. L. (2012). WISC-IV profile in high-functioning autism spectrum disorders: Impaired processing speed is associated with increased autism communication symptoms and decreased adaptive communication abilities. *Journal of Autism and Developmental Disorders*, *42*, 655–664.

Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S., & O'Toole, A. J. (2019). Social trait information in deep convolutional neural networks trained for face identification. *Cognitive Science*, *43*, Article e12729. https://doi.org/10.1111/cogs.12729

Parkhi, O., Vedaldi, A., & Zisserman, A. (2019). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 1–12). British Machine Vision Association.

Pelphrey, K., Sasson, N., Reznick, J. S., Paul, G., Goldman, B., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders*, *32*, 249–261.

Shic, F., Wang, Q., Macari, S. L., & Chawarska, K. (2020). The role of limited salience of speech in selective attention to faces in toddlers with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *61*, 459–469.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Spek, A., Schatorjé, T., Scholte, E., & van Berckelaer-Onnes, I. (2009). Verbal fluency in adults with high functioning autism or Asperger syndrome. *Neuropsychologia*, *47*, 652–656.

Spezio, M. L., Adolphs, R., Hurley, R. S. E., & Piven, J. (2007). Analysis of face gaze in autism using "Bubbles." *Neuropsychologia*, *45*, 144–151.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, *22*, 197–200.

Sutherland, C. A. M., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., Collova, J. R., & Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences, USA*, *117*, 10218–10224.

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology*, *113*, 1056–1078.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, USA*, *113*, 194–199.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

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

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science*, *2*, 609–617.

Wallach, M. A., & Kogan, N. (1959). Sex differences and judgment processes. *Journal of Personality*, *27*, 555–564.

Wang, S., & Adolphs, R. (2017a). Reduced specificity in emotion judgment in people with autism spectrum disorder. *Neuropsychologia*, *99*, 286–295.

Wang, S., & Adolphs, R. (2017b). Social saliency. In Q. Zhao (Ed.), *Computational and cognitive neuroscience of vision* (pp. 171–193). Springer.

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., & Zhao, Q. (2015). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, *88*, 604–616.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*, 1325–1331.

Xie, S. Y., Thai, S., & Hehman, E. (2022). Everyday perceiver-context influences on impression formation: No evidence of consistent effects. *Personality and Social Psychology Bulletin*, *49*, 955–968.

Zebrowitz, L. A., Kikuchi, M., & Fellous, J.-M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology*, *98*, 175–189.