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, St. Louis, MO 63110, USA

[2] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA

[3] Department of Psychological & Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106, USA

[4] Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV 26506, USA

[5] Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[6] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA

*: Equal contributions

Corresponding authors:

Runnan Cao (runnan.cao@mail.wvu.edu)

Shuo Wang (shuowang@wustl.edu)

Abstract

Processing faces is difficult for individuals with autism spectrum disorder (ASD). However, it remains unclear whether individuals with ASD are capable of making high-level social trait judgments from faces. Here, we comprehensively address this question using naturalistic face images and representatively sampled traits. Despite of intact underlying dimensions, people with ASD showed atypical judgments and reduced specificity within each trait. Deep neural networks revealed that these group differences were driven by discrepant judgments for certain types of faces and differential attention to certain features within a face. Our results were replicated in well-characterized in-lab participants and partially generalized to posed neutral faces (a preregistered study). Finally, atypical social trait judgments from faces in ASD were associated with socio-emotional experience during social interactions. Together, our results provide new insights into the computational bases and behavioral consequences of social trait judgments in ASD.

Keywords: Social trait judgment, Autism spectrum disorder, Face feature, Face perception

Introduction

People spontaneously make judgments of others' enduring dispositions upon seeing their faces: some look warm, some look competent, or feminine[1-3]. Although the accuracy of these trait judgments remains debated[4], they predict consequential behaviors in the real world, from dating and hiring decisions[5] to voting and courtroom sentencing[6,7]. Some studies have shown surprisingly high consensus between perceiver groups from different cultures and different age groups[8-10]. Others have found profound individual differences in such judgments[10-12]. For instance, a twin study has found stable differences in trustworthiness impressions, which are mainly shaped by unique personal experience[11]. A global study has shown that the different personality structures of people in one's local environment lead to different structures of one's trait judgments from faces[12]. However, little is known whether trait judgments from faces will also be different due to atypical social functioning such as that occurs in autism spectrum disorder (ASD).

ASD is a neurodevelopmental disorder that has been linked to deficits in the primary visual cortex[13,14], the mirror neuron system (e.g., in the inferior frontal gyrus)[15,16], and the network of structures implicated in social cognition (e.g., amygdala, fusiform face area)[17-21]. Individuals with ASD demonstrate multiple quantitative deficits in various aspects of face processing, including gaze processing, discriminating and memorizing different facial identities, and recognizing emotions from facial expressions[22-28]. They also spend less time engaging in social interactions and looking at faces[29-31], and of course a core part of the diagnostic criteria includes atypical social interactions. Given these two sets of findings — atypical face processing and atypical 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 towards others.

Findings from prior research remains inconclusive on this hypothesis. Studies using computer-generated faces generally find that individuals with ASD make similar trait judgments from faces as controls[32-34]. 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 ASD and controls in any of the traits[33]. In contrast, studies using photographs of real people have revealed abnormal trait judgments in ASD[32,35]. For instance, one study investigated judgments of trustworthy and approachability using

black-and-white photos of real faces in natural poses and found that individuals with ASD gave abnormally more positive ratings to these faces on both traits than controls[35]. 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 in comparison to controls. To maximize generalizability, we not only use naturalistic face stimuli of celebrities of diverse races, face angles, gaze directions, and facial expressions taken in naturalistic contexts (e.g., non-posing photos captured in the street or events)[36], but also compare our findings with more controlled face stimuli of unfamiliar individuals with neutral expressions in a preregistered study. We investigate 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)[3]. Using these rich data, we leverage deep learning techniques to characterize the specific patterns of atypical social trait judgment in ASD. We collect trait ratings from two large samples of online ASD and control participants (one sample for preregistered study), and a well-characterized in-lab sample of ASD and control participants. We compare the ASD and control data for their respective correlational structure across traits and ratings across faces within each trait. We characterize the types of faces as well as the features within a face that drive differences in the trait ratings between groups. Finally, in a case study, we show that our results have consequences for socio-emotional experience during social interactions. We do so by measuring guilt experience following interpersonal transgression, which critically depends on the transgressors' perception of the victim's reaction[37] such as that based on the victim's face.

Results

I. Intact psychological dimensions underlie trait judgments from faces in ASD

We recruited participants with self-identified ASD and controls (see **Methods**; autism confirmed with AQ and SRS; results replicated with in-lab participants with ASD who had ADOS diagnosis). Participants from each group (see **Table 1** and **Supplementary Fig. 1** 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 (PCA) on the aggregate ratings (averaged per face across participants) across the eight traits for each group. The first four PCs (without rotation) explained most of the variance in each group: 44%, 23%, 14%, 11% in online participants with ASD (total 92%) and 38%, 27%, 17%, 9% in online controls (total 92%). These results indicate that four dimensions optimally summarized the eight trait judgments of our naturalistic face stimuli.

Therefore, we extracted four PCs 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 (see **Fig. 1b** for PC loadings and **Supplementary Fig. 2a** for correlations between trait judgments), replicating the comprehensive four-dimensional space found in prior research with posed neutral white faces using a different type of face stimuli (naturalistic faces of famous people with different races and facial expressions) and different groups of participants [3]. 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 PC 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, 0.99 between ASD and controls). These results suggest that the comprehensive psychological dimensions that underlie trait judgments from faces remain intact in ASD.

II. Atypical ratings for trait judgments along all dimensions in ASD

A highly similar correlational structure across trait judgments between groups does not guarantee group similarity in the judgment of every trait. Here, we compared the judgment of each trait between the two groups. We first analyzed the inter-rater consistency for each trait judgment (**Fig. 1c, d**). We found that participants with ASD were more heterogeneous with respect to each other than controls for six of the eight traits distributed across all four dimensions (see **Fig. 1c, d** legend for statistics), consistent with the widely reported heterogeneity in ASD[38].

Trait judgments were highly consistent for different face images of the same identity for both participants with ASD and controls (**Supplementary Fig. 2b**). We next compared the mean of the aggregate ratings across participants per trait between groups. We found that participants with ASD gave atypical ratings for five of the eight traits distributed across all four dimensions (**Fig. 1e**; see figure legend for statistics): *warm*, *practical*, *feminine*, *strong*, and *youthful*. ASD and controls could be distinguished based on how they rated the faces on the eight traits (support vector machine classifier with 10-fold cross-validation and 100 repetitions, accuracy = 79.62%±4.18% [mean±SD]). These results suggest that individuals with 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 controls. We found that for the judgments of *warm*, *practical*, *strong*, and *youthful*, participants with ASD gave higher ratings for most of the face identities (**Fig. 1f**). These results showed that the atypically positive trait judgments in ASD were not merely driven by certain face identities. Interestingly, we found that for the judgments of *competent*, *practical*, and *feminine*, participants with ASD demonstrated a compressed range in their ratings across faces. That is, participants did not vary their ratings as much as controls across face identities (e.g., *competent* and *practical* in **Fig. 1f**), leading to higher ratings on the faces that controls judged low and lower ratings on the faces that controls judged high.

To formally quantify this observation, we compared the ratings between groups separately for the 10 face identities on which controls provided highest ratings (**Fig. 1g**) and the 10 face identities on which controls provided lowest ratings (**Fig. 1h**). We found that compared to controls, participants with ASD provided significantly lower ratings for the top 10 identities when judging *critical* (**Fig. 1g**; $t(806) = 3.33$, $P = 0.00090$), *competent* ($t(807) = 4.74$, $P = 2.49\times10^{-6}$), *practical* ($t(802) = 2.05$, $P = 0.041$), *feminine* ($t(736) = 6.51$, $P = 1.39\times10^{-10}$), and *youthful* ($t(829) = 2.24$, $P = 0.026$); and they provided significantly higher ratings for the bottom 10 identities for *warm* (**Fig. 1h**; $t(824) = 4.60$, $P = 4.80\times10^{-6}$), *competent* ($t(807) = 5.52$, $P = 4.66\times10^{-8}$), *practical* ($t(802) = 5.48$, $P = 5.64\times10^{-8}$), *feminine* ($t(736) = 6.17$, $P = 1.13\times10^{-9}$), *strong* ($t(822) = 3.06$, $P = 0.0023$), and *youthful* ($t(829) = 5.55$, $P = 3.93\times10^{-8}$). Therefore, the difference between the top 10 and bottom 10 identities was significantly smaller in participants with ASD compared to controls

across judgments of all four dimensions (**Fig. 1i**): *warm* ($t(824) = 4.03$, $P = 6.10\times10^{-5}$), *critical* ($t(806) = 3.30$, $P = 0.001$), *competent* ($t(807) = 9.82$, $P = 1.35\times10^{-21}$), *practical* ($t(802) = 6.85$, $P = 1.52\times10^{-21}$), *feminine* ($t(736) = 6.51$, $P = 7.01\times10^{-13}$), *strong* ($t(822) = 3.63$, $P = 0.0003$), and *youthful* ($t(829) = 5.07$, $P = 4.89\times10^{-7}$). Together, these results suggest that participants with ASD have a reduced discriminability for most social trait judgments. These findings are consistent with the reduced specificity in emotion perception [39] and noisier and more random eye movement behavior in autism in general [40-42].

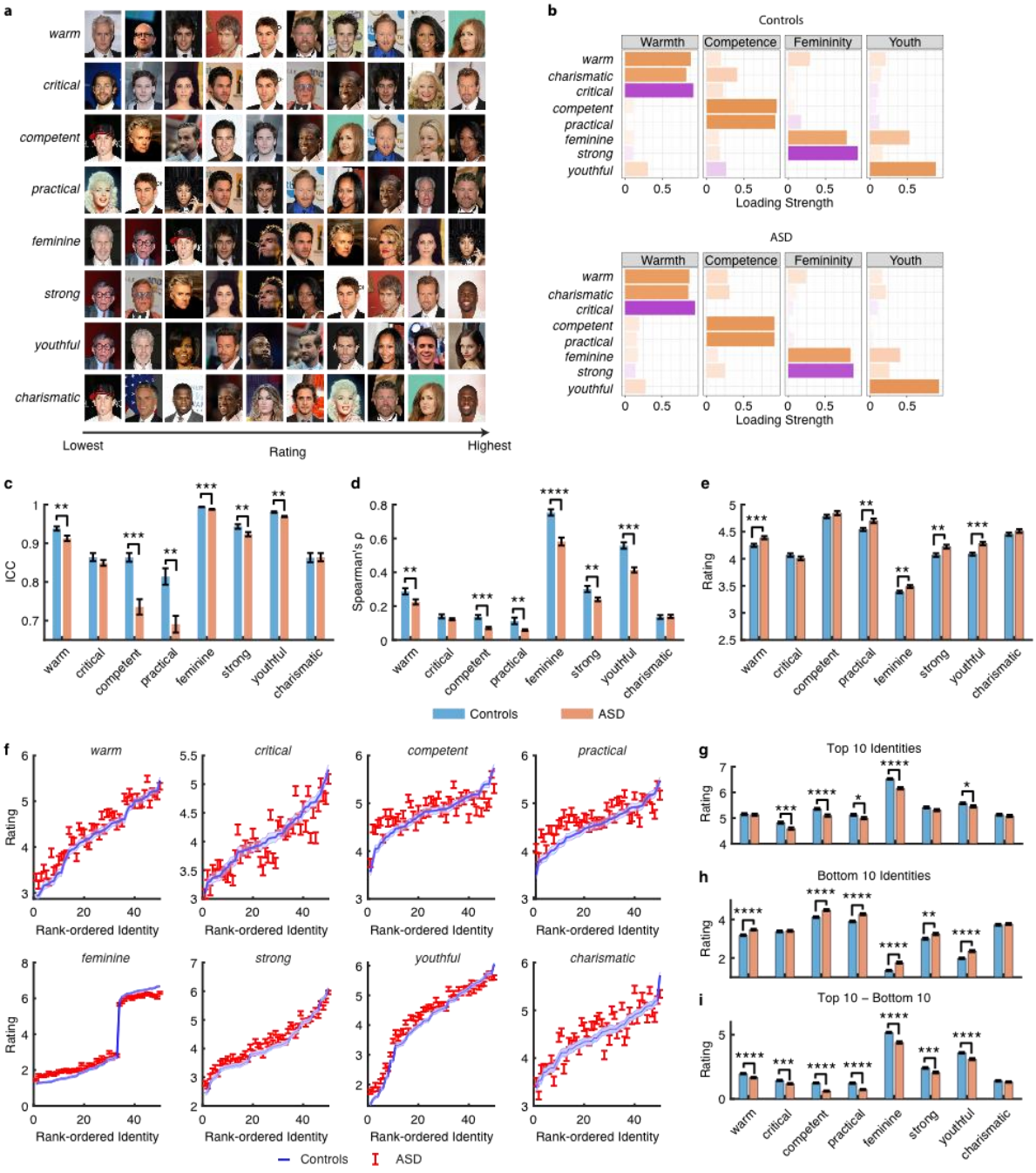


Fig. 1. Social trait judgments from participants with ASD and controls. **(a)** Example stimuli ranked by average ratings from controls for each social trait. **(b)** PCA loadings of social traits on the first four PCs. Each column plots the strength of the loadings (x-axis, absolute value) across traits (y-axis). Color coding indicates the sign of the loading (orange for positive and purple for negative). Saturated colors highlight each trait's most strongly correlated PC. **(c, d)** Inter-rater consistency. Inter-rater consistency of each trait was estimated using **(c)** the intraclass correlation coefficient (ICC) [64] and **(d)** the Spearman's correlation coefficient (ρ). Inter-rater consistency was first calculated between raters and averaged within each module, and then averaged across modules. Participants with ASD demonstrated lower inter-rater consistency for most of the traits: *warm* (two-tailed paired *t*-test across 10 rating modules; ICC: $t(9) = 3.45$, P = 0.0073; Spearman: $t(9) = 3.19$, P = 0.011), *competent* (ICC: $t(9) = 5.43$, P = 0.00042; Spearman: $t(9) = 5.78$, P = 0.00027), *practical* (ICC: $t(9) = 4.21$, P = 0.0023; Spearman: $t(9) = 3.26$, P = 0.0099), *feminine* (ICC: $t(9) = 5.19$, P = 0.00057; Spearman: $t(9) = 7.28$, $P = 4.66\times10^{-5}$), *strong* (ICC: $t(9) = 4.06$, P = 0.0029; Spearman: $t(9) = 4.09$, P = 0.0027), and *youthful* (ICC: $t(9) = 4.49$, P = 0.0015; Spearman: $t(9) = 5.76$, P = 0.00027). **(e)** Aggregate ratings. Participants with ASD gave atypical ratings for five traits (two-way repeated-measure ANOVA; main effect of participant group: $F(1,5613) = 12.82$, $P = 3.65\times10^{-4}$; main effect of trait: $F(7,5163) = 255.3$, $P = 1.22\times10^{-292}$; interaction: $F(7,5613) = 4.27$, $P = 1.03\times10^{-4}$): *warm* (two-tailed two-sample *t*-test across participants; $t(824) = 3.31$, P = 0.00097), *practical* ($t(802) = 3.13$, P = 0.0018), *feminine* ($t(736) = 2.65$, P = 0.0082), *strong* ($t(822) = 3.10$, P = 0.0020), and *youthful* ($t(829) = 4.47$, $P = 9.03\times10^{-6}$). Error bars denote ±SEM across rating modules. Asterisks indicate a significant difference between participants with ASD and controls using two-tailed two-sample *t*-test. *: P < 0.05, **: P < 0.01, ***: P < 0.001, and ****: P < 0.0001. **(f)** Ratings for each face identity rank-ordered by mean ratings from controls. Red: ASD. Blue: controls. Error bars and error shades denote ±SEM across rating modules. **(g)** Average ratings for the 10 identities with the highest ratings from controls. **(h)** Average ratings for the 10 identities with the lowest ratings from controls. **(i)** Difference in ratings between the top 10 and bottom 10 identities.

Since our face stimuli were photos of celebrities, with whom participants might be familiar, we investigated how familiarity of faces might influence trait judgments (**Supplementary Fig. 2c, d**). Participants with ASD had more atypical ratings for unfamiliar identities (**Supplementary Fig. 2c**) compared to familiar identities (**Supplementary Fig. 2d**). Specifically, participants with ASD rated unfamiliar identities on *warm*, *feminine*, *strong*, and *youthful* substantially higher than controls (**Supplementary Fig. 2c**), and rated familiar identities on *practical* and *feminine* slightly higher than controls (**Supplementary Fig. 2d**; see figure legend for statistics). The distribution of familiar and unfamiliar identities was similar between participant groups. Therefore, these results suggest that individuals with ASD tend to make positive evaluations to strangers' faces.

Since prior findings showed that people are biased by racial information when making trait judgments [43,44], we capitalized on our racially diverse stimuli and participants to analyze potential cross-race effects (**Supplementary Fig. 2e, f**). We found that group differences in trait judgments

were primary driven by same-race faces as the participants (**Supplementary Fig. 2e**; see figure legend for statistics) rather than cross-race faces (**Supplementary Fig. 2f**). In addition, we found that group differences in trait judgments were primarily driven by same-sex faces as the participants (**Supplementary Fig. 2g**; see figure legend for statistics) rather than cross-sex faces (**Supplementary Fig. 2h**). We also showed that race and sex information did modify social trait judgments (see **Supplementary Fig. 2** legend for details). Together, these findings suggest that people with ASD give the most atypical trait judgments when no other biographical information is known about the face (unfamiliar) and when it is of an in-group with respect to race and sex.

III. Features across faces that contribute to atypical trait ratings in ASD

What types of faces drove the rating differences in ASD? Do the faces that participants with ASD judged most atypically share common visual features? To answer these questions, we extracted facial features from each image using a pre-trained deep neural network (DNN) VGG-16 [45] and constructed a two-dimensional face feature space using t-distributed stochastic neighbor embedding (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 various social trait judgments [46,47] (note that other DNN models could derive similar results [48]). Furthermore, this model has demonstrated correspondence to neural processing of faces in the human brain, at both single-neuron level [48] and neural population level [49].

We next 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 2D feature map; **Supplementary Fig. 3a, i**). To formally quantify the difference in feature maps between groups and identify discriminative feature map regions for each social trait (see **Supplementary Fig. 3** for illustration of detailed procedures), we estimated a continuous density map in the feature space from our sparse sampling (**Fig. 2a-d** left and **Supplementary Fig. 3b, d, j, l**) and used a permutation test (1000 runs; **Fig. 2a-d** middle and **Supplementary Fig. 3c, e, k, m**) to identify regions that had a significant group difference (**Fig. 2a-d** right and **Supplementary Fig. 3h, p**). The identified region in the feature map of each DNN layer for each trait contained faces that were

most discriminative for ratings between ASD and controls (note that an equal difference across faces could not lead to a discriminative region; in other words, a discriminative region could not be simply resulted from the gross difference in ratings).

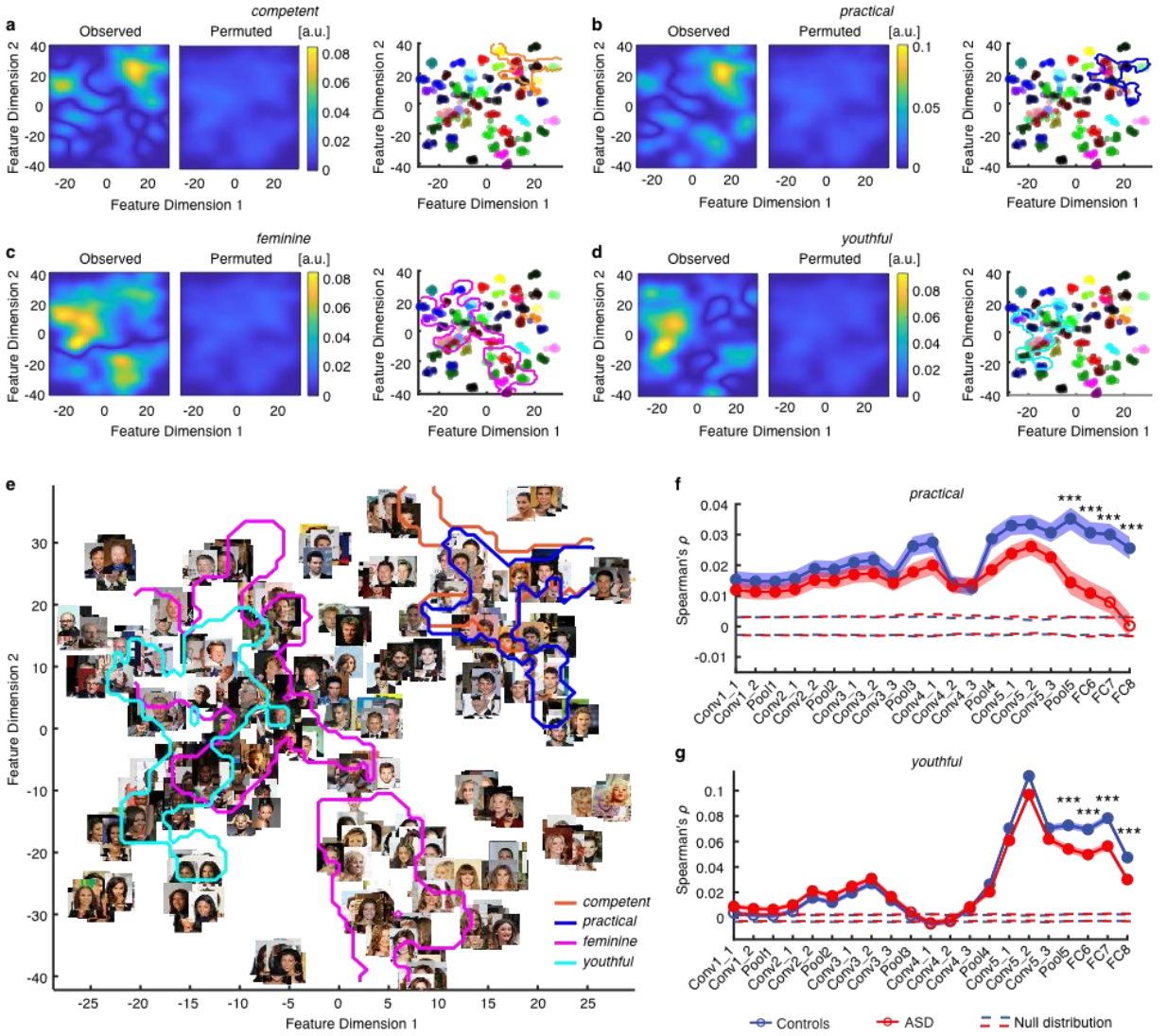


Fig. 2. Features across faces that contribute to atypical trait ratings in ASD. **(a-d)** Estimation of the rating density and identification of the discriminative regions in the feature space. By comparing observed (left) versus permuted (middle) difference in ratings between groups, we could identify 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 ASD and controls (delineated by the outlines; also shown in **(e)**). Color coding shows density in arbitrary units (a.u.). Each color in the scatter plot 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 (t-SNE) for the deep neural network (DNN) 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: $P < 0.05$, Bonferroni correction across layers). Shaded area denotes ±SD across rating modules. Dashed line denotes ±SD across permutation runs. Asterisks indicate a significant difference between participants with ASD and controls using permutation test. ***: $P < 0.001$. Red: ASD. Blue: controls. **(f)** Trait *practical*. **(g)** Trait *youthful.*

Using this approach, we identified faces that were judged most atypically by participants with ASD for each trait. For example, we found that judgments of *competent* primarily differed in young, male, Caucasian faces (**Fig. 2a, e**), whereas judgments of *youthful* primarily differed in African American faces as well as old, male, Caucasian faces (**Fig. 2d, e**; see **Supplementary Fig. 4a** for other discriminative regions across DNN layers). 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 (**Fig. 2a-e** and **Supplementary Fig. 4a**). These discriminative faces mainly appeared in the intermediate and later DNN layers where facial features are abstracted towards semantic representations (**Supplementary Fig. 4a**). These results suggest that the atypical social trait judgments in participants with ASD may stem from different representations of more abstract facial features. We further correlated the similarity across faces [50] between social trait ratings and DNN features (see **Methods**). Again, we found that the group differences in trait judgments were primarily in the later DNN layers (**Fig. 2f, g** and **Supplementary Fig. 4b**), confirming that atypical social trait judgments in ASD are driven by more abstract facial features.

IV. Features within faces that contribute to atypical trait ratings in ASD

Besides the different sensitivity to certain types of faces in ASD, the different ways that individuals with ASD take cues from an individual face may also contribute to their atypical social trait judgments. To understand which part of the face may be more informative for participants with ASD compared to controls, we trained a DNN to predict the rating from each face image for each trait (**Supplementary Fig. 5a**). A DNN was trained separately for participants with ASD and controls (see **Supplementary Fig. 5b** for model performance). We visualized the critical pixels in the face images that led to the correct prediction of social trait judgment using layer-wise relevance propagation (LRP) (see **Methods**).

We first confirmed that critical facial parts such as the eyes, mouth, and hair were important to predict social trait judgments (see **Fig. 3a, b** for examples and **Fig. 3c** for group summary). We then averaged LRP maps across face images for each participant group (**Fig. 3c**; note that critical facial parts are aligned across stimuli [51]) and identified face regions that quantitatively differed between groups (**Fig. 3d**). We found that controls relied more on information from the forehead than ASD when judging *warm*, *critical*, *practical*, and *strong*. Controls relied more on information from the eyes than ASD when judging *practical* and *strong*. Controls also relied more on information from the mouth than ASD when judging *warm*, *feminine*, and *strong*. Interestingly, participants with ASD utilized more information from the forehead and eyes when judging *youthful* and *charismatic* than controls. Together, our results reveal features within faces that participants with ASD differentially utilize to judge social traits compared to controls.

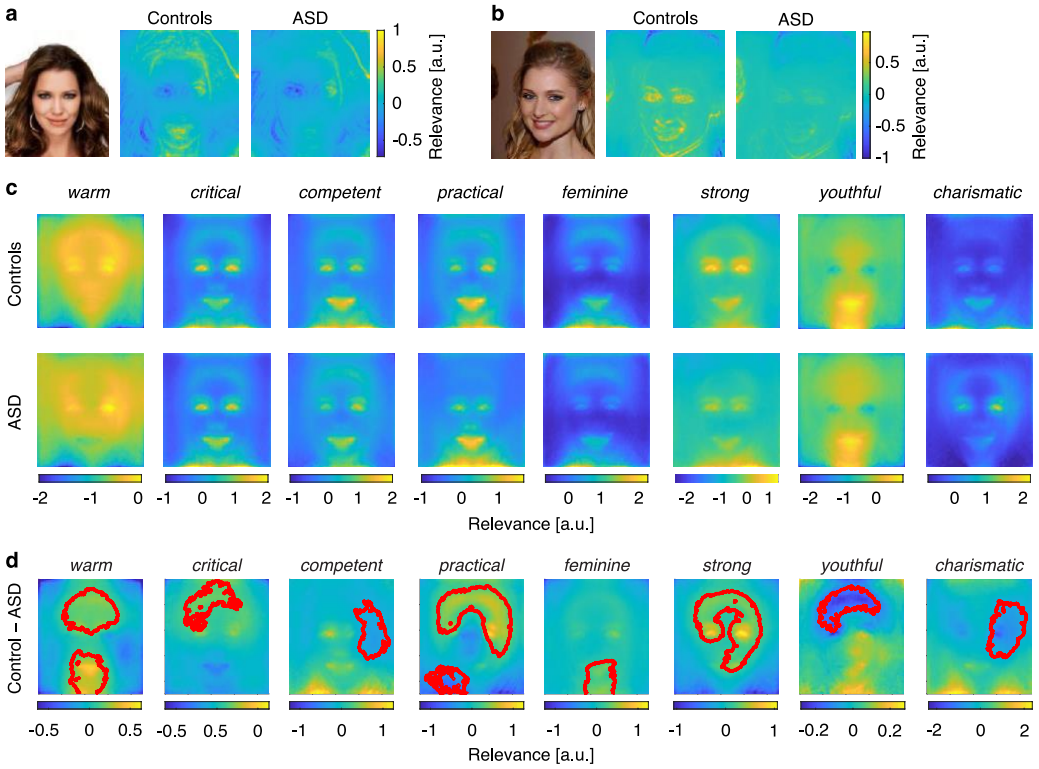


Fig. 3. Features within faces that contribute to atypical trait ratings in ASD. Relevance of each pixel to classification was revealed using layer-wise relevance propagation (LRP). Color coding shows LRP values in arbitrary units (a.u.). Yellow pixels positively contributed to the classification whereas 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. (Upper) Images from controls. (Lower) Images from participants with ASD. **(d)** Difference in LRP maps for each trait. Red contours show the regions with a significant difference between participants with ASD and controls using two-tailed paired *t*-test ($P < 10^{-18}$; cluster size > 5% of all pixels).

V. Validation with well-characterized in-lab participants

The above results were based on online participants with self-identified ASD. We next confirmed our findings by acquiring ratings from a sample of in-lab participants with confirmed ASD diagnosis (n = 27) and matched controls (n = 21). We reproduced the same PCA structure (**Fig. 4a**), reduced inter-rater consistency (**Fig. 4b, c**; see figure legend for statistics), and atypical trait ratings in ASD (**Fig. 4d**; note here controls showed more positive ratings; see **Discussion**). Importantly, we confirmed that participants with ASD had reduced specificity in their ratings (**Fig. 4e-h**). Together, we validated our main findings in participants with confirmed ASD diagnosis and matched controls.

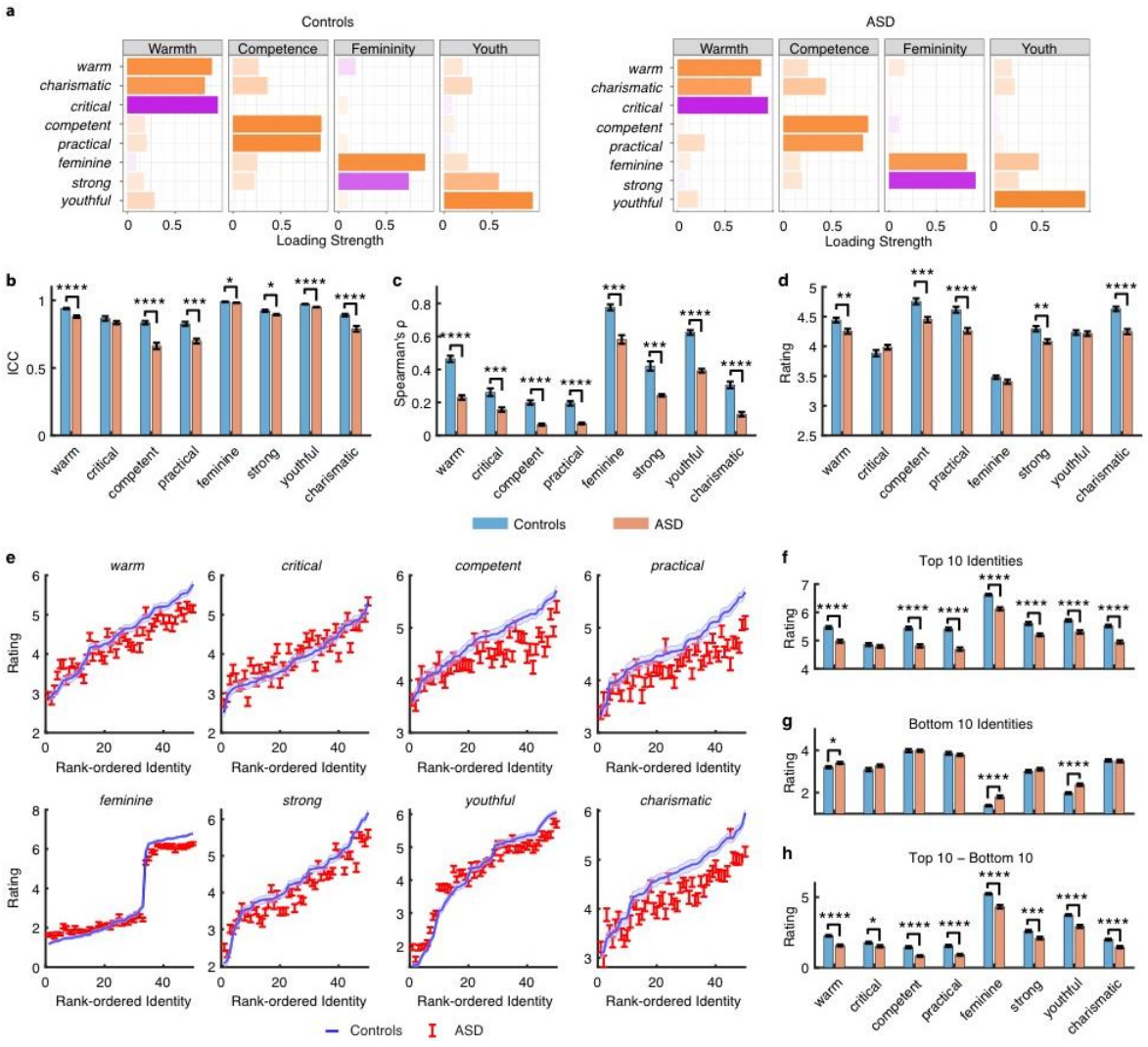


Fig. 4. Validation with in-lab participants. **(a)** PCA loadings of social traits on the first four PCs. **(b, c)** Inter-rater consistency. Participants with ASD demonstrated lower inter-rater consistency for most of the traits: *warm* (two-tailed paired t-test across 10 rating modules; ICC: $t(9) = 6.98$, $P = 6.44\times10^{-5}$; Spearman: $t(9) = 12.78$, $P = 4.48\times10^{-7}$), *critical*

(Spearman: $t(9) = 4.97$, P = 0.0007), *competent* (ICC: $t(9) = 9.15$, P = 7.44×10^{-6}; Spearman: $t(9) = 10.37$, P = 2.64×10^{-6}), *practical* (ICC: $t(9) = 6.49$, P = 0.0001; Spearman: $t(9) = 7.98$, P = 2.26×10^{-5}), *feminine* (ICC: $t(9) = 3.00$, P = 0.015; Spearman: $t(9) = 5.41$, P = 0.0004), *strong* (ICC: $t(9) = 2.69$, P = 0.025; Spearman: $t(9) = 5.73$, P = 0.0002), *youthful* (ICC: $t(9) = 7.82$, P = 2.65×10^{-5}; Spearman: $t(9) = 12.63$, P = 4.99×10^{-7}), and *charismatic* (ICC: $t(9) = 6.97$, P = 6.57×10^{-5}; Spearman: $t(9) = 12.58$, P = 5.14×10^{-7}). **(d)** Aggregate ratings. Controls had a significantly higher rating for *warm* ($t(397) = 2.76$, P = 0.006), *competent* ($t(369) = 3.87$, P = 0.0001), *practical* ($t(336) = 4.41$, P = 1.38×10^{-5}), *strong* ($t(403) = 3.27$, P = 0.001), and *charismatic* ($t(395) = 5.51$, P = 6.37×10^{-8}). **(e)** Ratings for each face identity rank-ordered by mean ratings from controls. Red: ASD. Blue: controls. Error bars and error shades denote ±SEM across rating modules. **(f)** Average ratings for the 10 identities with the highest ratings from controls. **(g)** Average ratings for the 10 identities with the lowest ratings from controls. **(h)** Difference in ratings between the top 10 and bottom 10 identities. Legend conventions as in **Fig. 1**.

VI. Comparison with posed neutral faces

We derived the above results using complex, naturalistic face stimuli. How well could individuals with ASD make social trait judgments from simpler, controlled face stimuli? To address this question, we conducted a preregistered study (see **Methods**) using posed photos of real people with neutral expressions from a previous study [3] (see **Fig. 5a** for examples). First, we replicated that the comprehensive dimensional structure underlying trait judgments remained intact in ASD (**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, 0.99). Second, we observed a reduced inter-rater consistency for *warm*, *feminine*, *strong*, and *youthful* (**Fig. 5c, d**; see figure legend for statistics), consistent with the results from complex, naturalistic face stimuli (**Fig. 1c, d**). Third, we observed a significant group difference in aggregate ratings only for the trait *critical* (**Fig. 5e**; see **Supplementary Fig. 2c** for a comparison), and reduced specificity in ratings only for *strong* and *youthful* (**Fig. 5f-h**; see **Fig. 1f-h** for a comparison). Thus, the findings from complex naturalistic faces partially generalized to posed neutral faces (our controls' ratings highly correlated with those in the previous study [3], see **Supplementary Fig. 6**). These findings suggest that, the differences between the two sets of face stimuli such as facial expressions, backgrounds, and familiarity may play an important role in shaping how individuals with ASD make social judgments from faces (see **Discussion**).

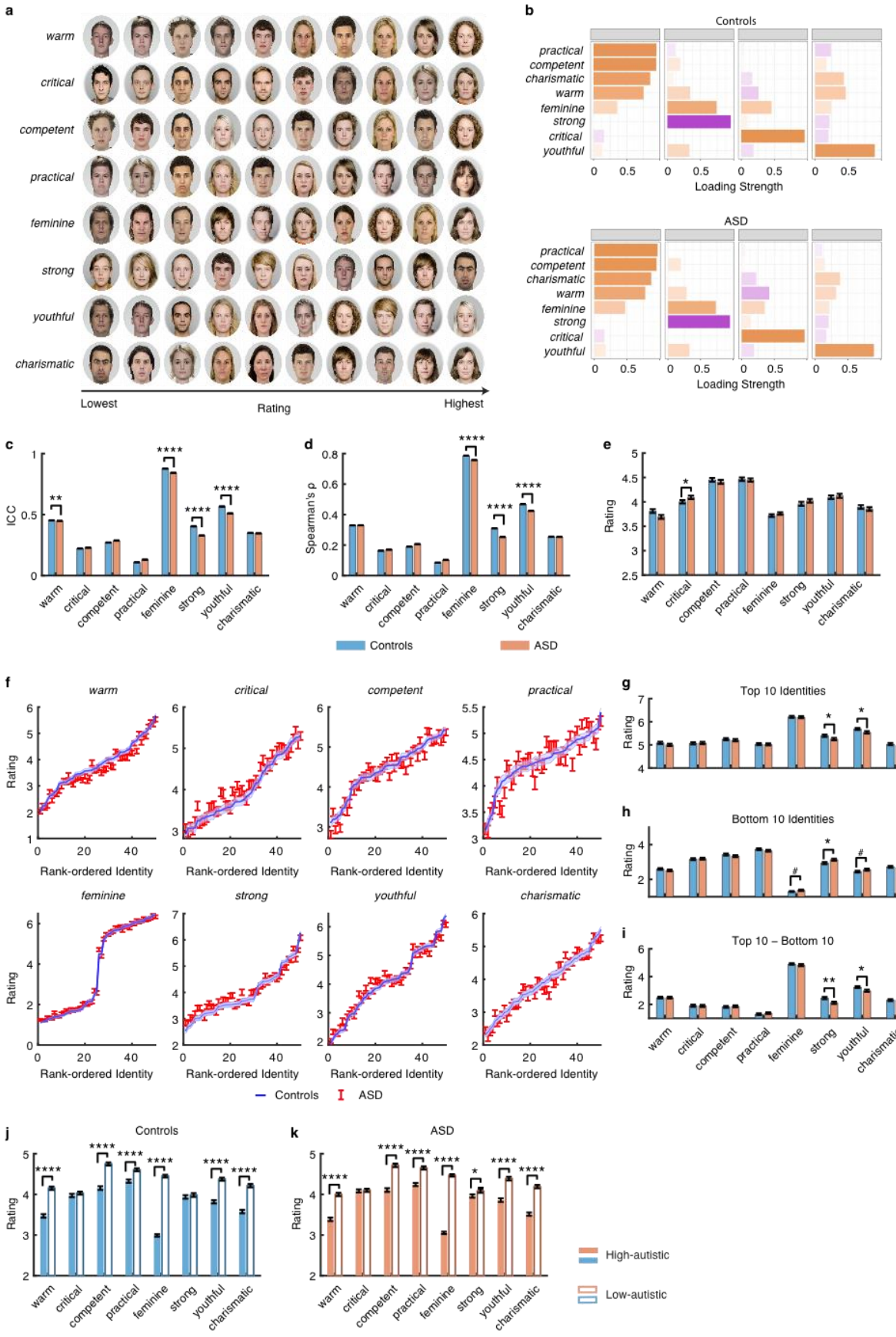

a
warm
critical
competent
practical
feminine
strong
youthful
charismatic
Lowest
Rating
Highest
b
Controls
practical
competent
charismatic
warm
feminine
strong
critical
youthful
0
0.5
Loading Strength
ASD
c
ICC
d
Spearman's ρ
e
Rating
warm
critical
competent
practical
feminine
strong
youthful
charismatic
Controls
ASD
f
Rank-ordered Identity
g
Top 10 Identities
h
Bottom 10 Identities
i
Top 10 – Bottom 10
j
Controls
k
ASD
High-autistic
Low-autistic

Fig. 5. Validation with an independent sample of participants using unfamiliar face stimuli. **(a)** Example stimuli ranked by average ratings from controls for each social trait. **(b)** PCA loadings of social traits on the first four PCs. **(c, d)** Inter-rater consistency. **(c)** Intraclass correlation coefficient (ICC). Participants with ASD demonstrated a lower ICC for *warm* (one-tailed two-sample *t*-test across participant pairs; $t(60769) = 2.40$, $P = 0.0082$), *feminine* ($t(59782) = 25.6$, $P < 10^{-10}$), *strong* ($t(61508) = 27.0$, $P < 10^{-10}$), and *youthful* ($(60516) = 15.5$, $P < 10^{-10}$). **(d)** Spearman's correlation coefficient (ρ). Participants with ASD demonstrated a lower correlation coefficient for *feminine* ($t(59782) = 21.4$, $P < 10^{-10}$), *strong* ($t(61504) = 28.7$, $P < 10^{-10}$), and *youthful* ($t(60516) = 23.2$, $P < 10^{-10}$). **(e)** Aggregate ratings. Participants with ASD had a significantly higher rating for *critical* (one-tailed two-sample *t*-test; $t(489) = 1.83$, $P = 0.034$). **(f)** Ratings for each face identity rank-ordered by mean ratings from controls. **(g)** Average ratings for the 10 identities with the highest ratings from controls. **(h)** Average ratings for the 10 identities with the lowest ratings from controls. **(i)** Difference in ratings between the top 10 and bottom 10 identities. Legend conventions as in **Fig. 1**. **(j, k)** Ratings for in-group versus out-group faces. Grouping was base on how "autistic" the faces were perceived by an independent group of participants [3]; and we used median-split to partition the stimuli into two groups. Solid bars: faces that were perceived as more autistic ("High-autistic"). Open bars: faces that were perceived as less autistic ("Low-autistic"). **(j)** Controls. **(k)** Participants with ASD. Error bars denote ±SEM across participants. Asterisks indicate a significant difference using two-tailed paired *t*-test. *: $P < 0.05$ and ****: $P < 0.0001$.

Interestingly, controls in the previous study provided ratings on how *autistic* the faces looked [3]. We next explored whether participants with ASD demonstrated an in-group/out-group bias in social evaluation. We median-split the stimuli into "High-autistic" versus "Low-autistic" based on those controls' ratings, and compared the ratings provided by our participants for each trait. We found that both controls (**Fig. 5j**) and participants with ASD (**Fig. 5k**) provided consistently more positive ratings for faces that were perceived as less autistic. Although our results suggest no in-group bias in participants with ASD, they suggest that the autistic impression of a face may modulate social trait judgments in both controls participants with ASD.

VIII. Atypical link between social trait judgment and socio-emotional experience in ASD

Do the atypical social trait judgments in individuals with ASD shape their socio-emotional experience in social interactions? To answer this question, we used a dot-estimation task and measured transgressor's guilt experience following interpersonal transgression as an exemplar social interactive context (**Fig. 6a**; see **Methods**). This task has been shown to effectively induce

guilt (see **Methods**). We recruited 47 online participants with self-identified ASD and 139 controls (17 participants with ASD and 124 controls overlapped with the main study mentioned above).

We first examined the two groups' performance in this task. We focused on comparing the self-incorrect condition and the both-incorrect condition because the participants' performance feedback was the same (namely, "incorrect") in these two conditions. Therefore, participants' perception of their own performance in the dot-estimation task should not have any impact on the comparison. We showed that our experimental conditions (self-incorrect versus both-incorrect) had a significant main effect on self-reported guilt (linear mixed effect model; $B = 12.01 \pm 2.71$, 95% CI = [6.70, 17.33]; $b = 0.39$, 95% CI = [0.22, 0.56]; $t = 4.43$, $P = 1.15 \times 10^{-5}$). However, there was no main effect of group ($B = -3.18 \pm 4.59$, 95% CI = [−12.16, 5.80]; $b = -0.10$, 95% CI = [-0.39, 0.19]; $t = -0.69$, $P = 0.488$). Importantly, we observed a significant group-by-condition interaction for self-reported guilt (**Fig. 6b**; $B = 6.91 \pm 3.12$, 95% CI = [0.78, 13.02]; $b = 0.22$, 95% CI = [0.03, 0.42]; $t = 2.21$, $P = 0.027$) and, to a lesser degree (marginal effect), the tendency towards compensating the partner (i.e., compensatory tendency; **Fig. 6c**; $B = 0.37 \pm 0.21$, 95% CI = [−0.03, 0.78]; $b = 0.12$, 95% CI = [−0.01, 0.25]; $t = 1.81$, $P = 0.071$). These results suggest that for ASD participants, responsibility in causing unpleasant outcomes to a social partner had less impact on their negative self-conscious emotions (e.g., guilt) compared with controls.

We next investigated whether group differences in socio-emotional experience are associated with social trait judgments from faces (**Fig. 6d-i**). As in the main study, we acquired trait ratings of naturalistic faces from a subset of participants who completed the guilt task (34 ASD, 124 controls). We first showed that trait judgments of faces had significant main effects on self-reported guilt (controlling for the effects of conditions, groups, and group-by-condition) for *warm* (linear mixed effect: $B = 9.56 \pm 3.87$, 95% CI = [2.03, 17.09]; $t = 2.47$, $P = 0.015$), *critical* ($B = 8.17 \pm 3.37$, 95% CI = [1.61, 14.72]; $t = 2.43$, $P = 0.017$), *youthful* ($B = 9.88 \pm 3.90$, 95% CI = [2.30, 17.47]; $t = 2.54$, $P = 0.012$), and *charismatic* ($B = 9.88 \pm 3.57$, 95% CI = [2.93, 16.82]; $t = 2.77$, $P = 0.006$). Importantly, we observed significant group-by-trait-judgment interaction effects on self-reported guilt for *warm* ($B = -22.73 \pm 8.48$, 95% CI = [−39.24, −6.21]; $t = -2.68$, $P = 0.008$), *competent* ($B = -17.24 \pm 7.00$, 95% CI = [−30.87, −3.62]; $t = -2.46$, $P = 0.015$), and *charismatic* ($B = -14.42 \pm 7.13$, 95% CI = [−28.30, −0.54]; $t = -2.02$, $P = 0.045$), indicating that the associations between self-reported guilt and these social trait judgments were weaker in participants with ASD

compared to controls. These results suggest that social trait judgments from faces are associated with guilt experience in social interactions and such socio-emotional responses and consequences in individuals with ASD are weakened.

Figure 6

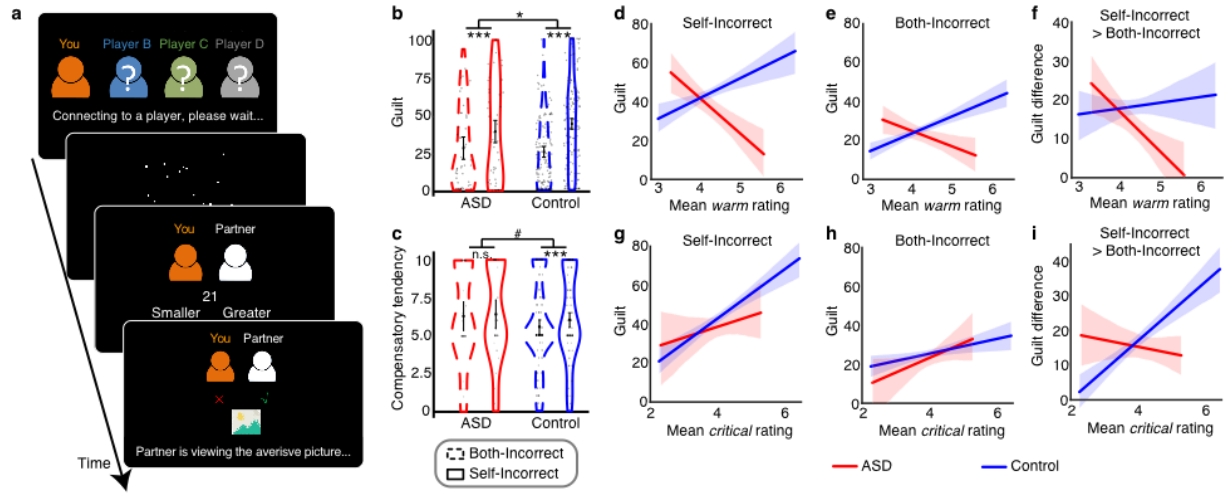


Fig. 6. Association between social trait judgment and guilt experience. **(a)** Procedure of the guilt induction task. At the beginning of each trial, the participant was randomly paired with a Player (i.e., Partner). The participant and the Partner then completed a dot estimation task, where they estimated the number of dots on the screen and compared it with a reference number (21 in this case). The participant's and the Partner's performance was then presented. Failure of either of them would lead to an unpleasant outcome for the Partner – viewing an aversive picture for 10 secs. In different trials, participants were asked to report their feelings of guilt or willingness to share the Partner's unpleasant outcome as a hypothetical measure of compensation. **(b)** Self-reported guilt and **(c)** compensation as a function of experimental condition and participant group. Our manipulation (i.e., responsibility) had stronger effects on **(b)** self-reported guilt and **(c)** compensatory tendency in the control group than in the ASD group. Each gray dot denotes a participant. The central black dot denotes the mean and the error bar denotes 95% confidence interval across participants. The contour is a kernel density estimation to show the distribution shape of the data. Significance of main effect and group-by-condition interaction: #: $P < 0.1$, *: $P < 0.05$, and ***: $P < 0.001$. n.s.: not significant. **(d-f)** Self-reported guilt in the **(d)** Self-Incorrect condition, **(e)** Both-Incorrect condition, and **(f)** their difference as a function of the participant's mean *warm* rating in the face judgment task. **(g-i)** Self-reported guilt in the **(g)** Self-Incorrect condition, **(h)** Both-Incorrect condition, and **(i)** their difference as a function of the participant's mean *critical* rating in the face judgment task. Blue line: control group. Red line: ASD group. Error shade denotes ±SEM across participants.

Discussion

We conducted a comprehensive investigation of how individuals with ASD make social trait judgments of faces compared to controls. Using a representative set of traits and naturalistic face stimuli, we showed that the dimensional structure underlying social trait judgments from faces remained intact in participants with ASD (**Fig. 1b**). However, participants with ASD showed reduced inter-rater consistency and atypical ratings for individual traits (**Fig. 1c-i**). We applied neural network modeling to show that these discrepant ratings could be explained by discrepant judgments for certain types of faces (**Fig. 2**) and discrepant utilization of features within a face (**Fig. 3**). We validated these findings using another sample of well-characterized in-lab participants (**Fig. 4** and **Supplementary Fig. 1**). We showed that these findings partially generalized to less complicated face stimuli (**Fig. 5**). Finally, we showed that social trait judgments from faces were associated with socio-emotional experience in social interactions: they were linked to guilt experience, but such effect was weakened in individuals with ASD (**Fig. 6**). Together, these findings provide a comprehensive characterization of the psychological structure and computational basis of social trait judgments from faces in individuals with ASD. These results provide initial insights into how atypical face processing in individuals with ASD may be linked to atypical social behavior.

Correlational structure between trait judgments

Prior research using the most comprehensive set of English trait words to-date showed that the hundreds of different trait judgments people make of faces could be summarized by four dimensions: warmth, competence, femininity, and youth [3]. That study was limited to posed photos of neutral faces of white individuals. Here, we replicated this four-dimensional space in both controls and individuals with ASD using diverse, naturalistic faces. 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 [43,52]. However, our results indicate that including these factors does not significantly change the correlational structure between different trait judgments.

The correlational structure between trait judgments have 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 [12,53]. Individuals with ASD are known to have impairment in verbal ability [54,55] and reduction in social interactions [29-31], even for high-functioning ASD populations. Surprisingly, we found that the correlational structure across trait judgments of faces were highly similar between controls and individuals with ASD. These findings suggest that individuals with ASD share similar understanding of the semantic relationship between different social trait descriptions and similar understanding of the personality structure in everyday life as controls.

Judgments of individual social traits

In line with prior findings on the substantial heterogeneity among individuals with ASD [38], we showed that the between-subject consensus in social trait judgments of faces among individuals with ASD was lower than that in controls. 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 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 we provided here could be flexibly applied to comparing models trained on dense individual data, which will provide insights into this possibility.

We observed atypical judgments within each trait in individuals with ASD compared to controls (**Fig. 1**). Importantly, these group differences were not merely due to baseline differences (e.g., ASD rated all faces on a trait higher than controls). Instead, atypical judgments were most salient for specific types of faces: the faces that received most extreme ratings from controls (**Fig. 1g-i**), unfamiliar faces (**Supplementary Fig. 2c-d**), same-race faces (**Supplementary Fig. 2e-f**), same-

sex faces (**Supplementary Fig. 2g-h**), and trait-dependent subsets of faces (**Fig. 2**). These findings suggest that the atypical social evaluation of faces may be a results of different conceptual associations between social traits and social groups (i.e., social stereotypes) in individuals with ASD compared to controls (e.g., the most stereotypical faces for controls received less extreme ratings in ASD; **Fig. 1i**).

We revealed that participants took different cues from faces for social trait judgments (**Fig. 3**), consistent with the large eye-tracking literature showing that people with ASD view faces differently [40,56-58]. 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 [57], but instead have an increased preference to fixate the location of the mouth [56]. Furthermore, people with ASD demonstrate active avoidance of fixating the eyes in faces, which in turn influences recognition performance of emotions [58]. In particular, our recent study has shown that the neural substrates underlying fixations on faces are related to perceived social trait judgments [51]. Therefore, atypical social trait judgment in ASD may stem from atypical eye movement patterns when viewing faces. Furthermore, atypical social trait judgment in ASD may be attributed to differential neural face representation in the amygdala and hippocampus [21].

Comparison between trait judgments of naturalistic versus posed faces

Prior studies have shown inconsistent results regarding whether individuals with ASD make social trait judgments of faces similar to controls. Earlier studies using photographs of real people have shown that people with ASD have atypical evaluation of social traits such as facial trustworthiness [32,35]. Other studies using computer-generated faces have shown that adults with ASD are as capable as controls to judge trustworthiness and dominance from faces [34] and a variety of seven different traits [33]. These findings suggest that the complexity of the face stimuli may play an important role in understanding how capable individuals with ASD are to make social trait judgments in comparison to controls.

Our results provided initial empirical evidence supporting this possibility. Besides 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 also conducted a preregistered study using an independent set

of controlled face stimuli that were neutral, frontal, with direct gaze and uniform background. We tested three preregistered hypotheses: the overall dimensional structure, inter-rater consistency, and rating specificity. We found that the trait judgments from individuals with ASD made to these less complex facial stimuli (**Fig. 5**) were more in line with controls than to naturalistic faces (**Fig. 1**): the overall dimensional structure remained intact in ASD as before; we again observed reduced inter-rater consistency in ASD but for a fewer number of traits; and that we observed atypical ratings in ASD also for a few number of traits. Thus, our findings reconcile the discrepancies in prior research by varying the face stimuli, and predicts that individuals with ASD would experience more difficulties when making social trait judgments in real-world interactions.

Advantages of our study

Our present study has the following main strengths:

First, we tested the generalizability of our conclusions using multiple sets of participants and stimuli. We not only replicated the results in our main experiment with in-lab participants with confirmed autism diagnosis but also conducted a preregistered study to show the generalizability of our results to a different type of face stimuli.

Second, we used more naturalistic stimuli. Prior studies on social trait judgments of faces have largely relied on highly controlled and artificial face stimuli, limiting the external validity of prior conclusions. The computer-generated faces used in prior research were unnatural-looking and limited to emotionally neutral and White faces [32-34]. Photographs of real people present more naturalistic-looking faces, but those used in prior studies were limited to photos of White individuals with direct gaze and neutral expressions taken in controlled lighting and backgrounds [32,35]. In real life, we see faces of diverse races, with different gaze directions, face angles, and facial expressions, and in complex contexts. In order to obtain a more generalizable understanding of trait judgments from faces in ASD, it is critical to examine more naturalistic face stimuli.

Third, we used a more comprehensive list of social traits for face judgments. As with increasing the external validity and generalizability of face stimuli, it is equally essential to broaden the trait judgments that we examine. Recent research shows that neurotypical individuals spontaneously

make hundreds of different trait judgments from faces [3]. Prior research on trait judgments in ASD has examined a decent number of traits (with a maximum of seven traits), but they may not be representative of the comprehensive space of trait judgments from faces. Understanding how individuals with ASD make judgments along all trait dimensions will allow for maximal generalizability to diverse trait judgments from faces.

Lastly, we used convergent analytic approaches (e.g., SVM, RSA, DNNs) that allowed a multifaceted and comprehensive analysis of social trait judgment in ASD. We further explored the implications of our results to individual differences in personality as well as socio-emotional experience during social interactions to understand the driving factors as well as consequences of atypical social trait judgment in ASD.

Together, the combination of diverse participant samples, stimulus sets, and analysis methods in the present study provides the most comprehensive characterization of social trait judgments of faces in ASD to date.

We further discuss possible caveats in **Supplementary Discussion**.

Future directions

In this study, we treated the ASD group as homogenous (using their group average ratings in our analyses), even though there may be considerable heterogeneity in individuals with ASD (see discussions above). Therefore, future studies using much large samples of participants will provide valuable insights into the different subtypes of social trait judgments in individuals with ASD. Furthermore, our present results may suggest potential intervention strategies: training people with ASD to look at facial features in a certain way, or training them to be careful on specific social judgments may increase their alignment with controls in evaluating faces on social dimensions. In particular, we showed that people with ASD had less atypical ratings for familiar/famous faces compared to those unfamiliar to them, suggesting that other biographical information about the targets may compensate for and mitigate the otherwise atypical face judgments in ASD. Future research using longitudinal designs will offer important empirical evidence on whether people with ASD who showed atypical social evaluation of faces and even atypical social behavior towards

unfamiliar people initially may adjust their social evaluation and behavior similar to controls once they obtain more information about the targets.

Methods

Participants

We recruited 525 participants from the Prolific platform (referred to as online participants). We only included participants who had English fluency, normal or corrected-to-normal vision, an education level above high school, and a Prolific approval rate greater than 95%. Among these participants, 113 participants had a self-reported diagnosis of ASD and 412 participants reported no diagnosis of ASD and served as controls (see **Table 1** for demographics). Self-identification 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." And we only included participants whose response was "Yes-as a child" or "Yes - as and adult" in the 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"). To further confirm their ASD demonstration, we acquired Autism Spectrum Quotient (AQ) and Social Responsiveness Scale-2 Adult Self Report (SRS) from the participants (89 participants with ASD and 307 controls completed the questionnaires) and we confirmed that online ASD participants had a significantly higher AQ (**Supplementary Fig. 1a**; ASD: 27.76±8.09 [mean±SD], controls: 20.28±6.82; $t(427) = 8.94$, $P = 1.15\times10^{-17}$) and SRS (**Supplementary Fig. 1b**; ASD: 91.73±29.66, controls: 65.17±25.19; $t(427) = 8.61$, $P = 1.38\times10^{-16}$) than online control participants. Furthermore, the online ASD participants had a comparable AQ (two-tailed two-sample *t*-test: $t(113) = 0.86$, $P = 0.39$) and SRS ($t(110) = 1.64$, $P = 0.10$) as in-lab ASD participants (see below). Lastly, based on our screening criterion, online controls had no mental health conditions.

Due to a surge of female participants in the Prolific platform during our data collection for ASD participants [59], the female population of ASD participants was over represented in our sample

(**Table 1**; but see [60] for prevalence of ASD in the general population). However, we observed qualitatively the same results with male participants with ASD only (**Supplementary Fig. 1h**). In addition, although the two groups of participants that we sampled differed in age (**Table 1** and **Supplementary Fig. 1c**; $t(523) = 3.25$, $P = 0.0012$), we observed similar results when we compared a subset of participants that matched in age (**Supplementary Fig. 1g**).

In our first control experiment, we recruited 27 high-functioning participants with ASD from our laboratory's registry and 21 neurologically and psychiatrically healthy participants with no family history of ASD as controls (see **Table 1** for demographics; referred to as in-lab participants). All of our in-lab ASD participants met DSM-V and Autism Diagnostic Observation Schedule (ADOS) criteria for ASD (**Table 1**). We confirmed that in-lab ASD participants had a significantly higher AQ (ASD: 29.57±12.06 [mean±SD], controls: 13.94±5.72; two-tailed two-sample *t*-test: $t(38) = 5.00$, $P = 1.31\times10^{-5}$) and SRS (ASD: 79.85±28.27, controls: 32.29±25.63; two-tailed two-sample *t*-test: $t(38) = 5.69$, $P = 1.54\times10^{-6}$) than in-lab control participants.

In our second control experiment, we recruited another 247 participants with self-reported ASD and another 251 control participants from the Prolific platform as an independent replication sample (**Table 1**). The data collection and some data analysis were preregistered (https://osf.io/bdrty/?view_only=089d0797257141e38564d6fffb9da4ce).

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

We used photos of celebrities from the CelebA dataset [36]. We selected 50 identities with 10 images for each identity, totaling 500 face images. The identities were selected to include both sexes (33 male) and multiple races (40 identities were Caucasian, 9 identities were African American, and 1 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.

In our second control experiment, we used 50 face stimuli 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 [3]. They were high-resolution studio photographs of human participants from three popular databases: the Chicago Face Database [61], the Oslo Face Database [62], and the Face Research Lab London [63]. All face stimuli were frontal, clear, with a neutral expression, and presented at the center of the images with the eyes aligned to the same location. All photos included the faces, necks, and hairs. All photos were colored, with a standard grey background, and were cropped to a standard size and shape.

Online rating of social traits

Participants were asked to provide judgments of social traits on a 1 to 7 scale through an online rating task. The social traits include *warm, critical, competent, practical, feminine, strong, youthful,* and *charismatic*; and these social traits were well validated in a previous study [3]. Participants also indicated whether they could recognize the identity of the face (i.e., whether they were familiar with each face identity) in the main experiment.

We divided the experiment into 10 modules, with each module containing one face image per face identity (totaling 50 face images per module). Each module included all 8 social traits (rated in blocks). In our main experiment, each online ASD participant completed 1 to 10 modules and each online control participant completed 1 to 2 modules. In our first control experiment, in-lab participants with ASD completed 4 to 10 modules and in-lab controls completed 1 to 10 modules. In our second control experiment, each online ASD participant and control participant completed one module. 50 face images (25 female, 25 male) were selected randomly from a representatively sampled set of 100 White faces from a previous study [3].

We applied the following three exclusion criteria:

(1) Trial-wise exclusion: we excluded trials with reaction times shorter than 100 ms or longer than 5000 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 3 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 3 blocks were excluded from the module per (2) above.

Inter-rater consistency

Inter-rater consistency of each trait was estimated using the intraclass correlation coefficient (ICC; two-way random-effects model for the consistency of mean ratings) [64] 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 implementation written by Arash Salarian (2020) (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 (PCA)

We conducted a principal component analysis (PCA). PCA is a statistical procedure that converts a set of high-dimensional, possibly correlated variables into a set of lower-dimensional, linearly uncorrelated PCs 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. Based on the aggregated data (500 faces x 8 traits), we extracted eight PCs (using *R* function *principal*, without rotation) for each participant group. We retained PCs that explained a nontrivial amount of variance (> 5%). After identifying the optimal number of PCs, we applied varimax rotation to the PCs to generate orthogonal components that were most interpretable.

Classification of participants

We employed a linear support vector machine (SVM) to discriminate whether a rating module was from a participant with ASD or a control. We used all ratings (8 traits x 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 10-fold cross-validation (i.e., each time 9 portions of modules were used as the training set and the remaining 1 portion of modules were used as the testing set). We had 100 runs in total (i.e., repeating the cross-validation procedure 100 times).

Feature extraction and construction of feature space

We used the well-known deep neural network (DNN) implementation based on the VGG-16 convolutional neural network (CNN) architecture [45] to extract features for each face image. Fine-tuning was performed on the pre-trained VGG-Face deep model using all images of the 50 identities in the CelebA dataset (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 stochastic neighbor embedding (SNE) [65], a commonly used method for multiple class high-dimensional data visualization [66]. 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 utilized any information from social trait ratings.

To identify the regions in the feature space showing a significant difference between groups (see **Supplementary Fig. 3** for a detailed illustration), we first estimated a continuous density map in the feature space by smoothing the discrete rating differences between groups using a 2D 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 1000 runs, we randomly shuffled the labels of participants. We calculated the p-value for each pixel by comparing the observed density value to those from the null distribution derived from permutation. 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 lastly selected the region with significant pixels (permutation $P < 0.01$, false discovery rate (FDR) [67] corrected for $Q < 0.01$, cluster size > 5% of the pixels within the mask).

Representational similarity between social trait ratings and DNN features

We employed a pairwise distance metric [49] 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 performed 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 and calculated the pairwise distance metrics between face identities.

To determine statistical significance *above chance*, we used a non-parametric permutation test with 100 runs. In each run, 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 eventually compared to the one *without* shuffling (i.e., observed value). If the correlation coefficient of the observed value was greater than 95% of the values from the null distribution, it was considered significant. A significant correlation indicated a representational similarity between social trait ratings and DNN features.

To determine statistical significance *between groups*, we further used a permutation test with 1000 runs to statistically compare the representational similarity between participants with ASD and controls. In each run, 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 where 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) pre-trained using ImageNet stimuli was used as the initial model. We kept all convolution layers, but replaced the last two fully-connected (FC) layers and the output layer with a global averaging pooling layer, a FC layer, and a prediction output layer (see **Supplementary Fig. 5a** for an illustration). When training our regression model, all convolutional layers were frozen (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 (SGD) 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.

We perform 10-fold cross-validation in our experiment. In each training/testing run (separately for each trait and each participant group), the dataset was randomly split into 10 subsets. One subset served as the test set, and the remaining 9 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*) could indicate the model's predictability. Our VGG-16 network ran on the deep learning framework TensorFlow 1.15 using Python 3.6.

To explain our model's output in the domain of its input, 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 until the original input image. The explanation given by LRP is a heatmap of which pixels in the original image contribute

to the final output. We used the toolbox iNNvestigate[68] (https://github.com/albermax/innvestigate) for implementation.

Guilt task

In this online task, participants were paired with another hypothetical participant (hereafter, "partner"). On each trial (**Fig. 6A**), the participant and the partner saw an array of dots (about 20) displayed on the screen for a short interval (1.5 s). The participant and, ostensibly, the partner indicated whether the number of the dots was larger or smaller than a reference number (e.g., 20). Afterward, their performance were presented on the screen (i.e., the outcome feedback phase). If one or both of them responded incorrectly, the partner had to watch an aversive image (i.e., unpleasant outcome), which was selected from the International Affective Picture System (IAPS) [69]. This way, we were able to manipulate the participant's responsibility in causing unpleasant outcomes to the partner. To make sure that the number of trials was balanced across conditions, unbeknownst to the participants, the outcome feedback was predetermined. There were 12 trials for each of the four possible outcomes (i.e., both-correct, partner-incorrect, self-incorrect, both-incorrect). The order of the outcome (or conditions) was randomized across participants. Our previous studies have demonstrated the validity of this task in inducing different levels of perceived guilt, negative self-conscious emotions, and compensatory behaviors [70-73]. These studies have consistently shown that participants perceive the highest level of guilt, report the highest level of self-conscious emotions, and engage in the highest degree of compensatory behaviors when they but not the partner respond incorrectly, less so when both the participant and the partner respond incorrectly, followed by the situation where the partner but not the participants respond incorrectly [70]. On the trials where the partner watched the aversive image, the participants were instructed to watch the partner's face via teleconference software after the partner ostensibly found out about the outcome (i.e., the watch-partner phase). The video, lasting for 10 seconds each, was presented such that the eye region of the partner aligns with a fixation cross on the screen, where the participants were required to fixate at the onset of the virtual interaction.

Between the outcome feedback phase and the watch-partner phase, participants were prompted to answer one of following questions: (1) how guilty they were for the partner's unpleasant outcome,

(2) their self-conscious emotions (e.g., remorseful), (3) the partner's emotions towards them with respect to their performance (e.g., anger, disappointment), (4) how much they would be willing to experience the unpleasant outcome themselves, so that the partner could experience the unpleasant outcome less. Each question was randomly presented twice in each condition (excluding the both-correct condition), and participants indicated their responses on analog scales.

Data and code availability

All data and code are publicly available on Open Science Framework (https://osf.io/e2kc6/?view_only=78f93622b0514268a6cbf660fe817981).

Acknowledgements

We thank Ralph Adolphs for valuable comments. This research was supported by the AFOSR (FA9550-21-1-0088), NSF (BCS-1945230, IIS-2114644), NIH (R01MH129426), and Dana Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

R.C., X.L., C.L., and S.W. designed research. R.C., H.Y., P.J.W., L.K.P., and C.L. performed experiments. R.C., N.Z., H.Y., C.L., and S.W. analyzed data. R.C, H.Y., X.L., C.L., and S.W. wrote the paper. All authors discussed the results and contributed towards the manuscript.

Competing Interests Statement

The authors declare no conflict of interest.

Table 1. Summary of participants. In our main experiment with naturalistic faces, we recruited online participants with ASD and online controls. In our first control/validation experiment, we recruited in-lab participants with ASD and in-lab controls. In our second control/validation experiment, we recruited another population of online participants. For all of our in-lab participants with ASD, their diagnosis has been confirmed using the Autism Diagnostic Observation Schedule-2 (ADOS-2) [74]. 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. The values are mean±SD.

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

References

1 Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* **105**, 11087-11092 (2008). https://doi.org:10.1073/pnas.0805664105

2 Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology* **66**, 519-545 (2015). https://doi.org:10.1146/annurev-psych-113011-143831

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

4 Bonnefon, J.-F., Hopfensitz, A. & De Neys, W. Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences* **19**, 421-422 (2015). https://doi.org:10.1016/j.tics.2015.05.002

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

6 Lenz, G. S. & Lawson, C. Looking the Part: Television Leads Less Informed Citizens to Vote Based on Candidates' Appearance. *American Journal of Political Science* **55**, 574-589 (2011). https://doi.org:https://doi.org/10.1111/j.1540-5907.2011.00511.x

7 Wilson, J. P. & Rule, N. O. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science* **26**, 1325-1331 (2015). https://doi.org:10.1177/0956797615590992

8 Walker, M., Jiang, F., Vetter, T. & Sczesny, S. Universals and Cultural Differences in Forming Personality Trait Judgments From Faces. *Social Psychological and Personality Science* **2**, 609-617 (2011). https://doi.org:10.1177/1948550611402519

9 Cogsdill, E. J., Todorov, A. T., Spelke, E. S. & Banaji, M. R. Inferring Character From Faces: A Developmental Study. *Psychological Science* **25**, 1132-1139 (2014). https://doi.org:10.1177/0956797614523297

10 Hester, N., Xie, S. Y. & Hehman, E. Little Between-Region and Between-Country Variance When People Form Impressions of Others. *Psychological Science* **32**, 1907-1917 (2021). https://doi.org:10.1177/09567976211019950

11 Sutherland, C. A. M. *et al.* Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences* **117**, 10218-10224 (2020). https://doi.org:doi:10.1073/pnas.1920131117

12 Oh, D., Martin, J. D. & Freeman, J. B. Personality Across World Regions Predicts Variability in the Structure of Face Impressions. *Psychological Science*, 09567976211072814 (2022). https://doi.org:10.1177/09567976211072814

13 Sutherland, A. & Crewther, D. P. Magnocellular visual evoked potential delay with high autism spectrum quotient yields a neural mechanism for altered perception. *Brain* **133**, 2089-2097 (2010). https://doi.org:10.1093/brain/awq122

14 Robertson, C. E. *et al.* Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain* **137**, 2588-2599 (2014). https://doi.org:10.1093/brain/awu189

15 Dapretto, M. *et al.* Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* **9**, 28-30 (2006). https://doi.org:10.1038/nn1611

16 Iacoboni, M. & Dapretto, M. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience* **7**, 942-951 (2006). https://doi.org:10.1038/nrn2024

17 Pierce, K., Müller, R. A., Ambrose, J., Allen, G. & Courchesne, E. Face processing occurs outside the fusiform `face area' in autism: evidence from functional MRI. *Brain* **124**, 2059-2073 (2001). https://doi.org:10.1093/brain/124.10.2059

18 Schultz, R. T. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience* **23**, 125-141 (2005). https://doi.org:https://doi.org/10.1016/j.ijdevneu.2004.12.012

19 Pinkham, A. E., Hopfinger, J. B., Pelphrey, K. A., Piven, J. & Penn, D. L. Neural bases for impaired social cognition in schizophrenia and autism spectrum disorders. *Schizophrenia Research* **99**, 164-175 (2008). https://doi.org:https://doi.org/10.1016/j.schres.2007.10.024

20 Henry, J. D., von Hippel, W., Molenberghs, P., Lee, T. & Sachdev, P. S. Clinical assessment of social cognitive function in neurological disorders. *Nature Reviews Neurology* **12**, 28-39 (2016). https://doi.org:10.1038/nrneurol.2015.229

21 Cao, R. *et al.* A neuronal social trait space for first impressions in the human amygdala and hippocampus. *Molecular Psychiatry* (2022). https://doi.org:10.1038/s41380-022-01583-x

22 Schultz, R. T. *et al.* Abnormal Ventral Temporal Cortical Activity During Face Discrimination Among Individuals With Autism and Asperger Syndrome. *Archives of General Psychiatry* **57**, 331-340 (2000). https://doi.org:10.1001/archpsyc.57.4.331

23 Dalton, K. M. *et al.* Gaze fixation and the neural circuitry of face processing in autism. *Nat Neurosci* **8**, 519-526 (2005).

24 Dawson, G., Webb, S. J. & McPartland, J. Understanding the Nature of Face Processing Impairment in Autism: Insights From Behavioral and Electrophysiological Studies. *Developmental Neuropsychology* **27**, 403-424 (2005). https://doi.org:10.1207/s15326942dn2703_6

25 Pelphrey, K. A., Morris, J. P. & McCarthy, G. Neural basis of eye gaze processing deficits in autism. *Brain* **128**, 1038-1048 (2005).

26 Harms, M., Martin, A. & Wallace, G. Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. *Neuropsychology Review* **20**, 290-322 (2010). https://doi.org:10.1007/s11065-010-9138-6

27 Weigelt, S., Koldewyn, K. & Kanwisher, N. Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neuroscience & Biobehavioral Reviews* **36**, 1060-1084 (2012). https://doi.org:https://doi.org/10.1016/j.neubiorev.2011.12.008

28 Rutishauser, U. *et al.* Single-Neuron Correlates of Atypical Face Processing in Autism. *Neuron* **80**, 887-899 (2013).

29 Jahr, E., Eikeseth, S., Eldevik, S. & Aase, H. Frequency and latency of social interaction in an inclusive kindergarten setting: A comparison between typical children and children with autism. *Autism* **11**, 349-363 (2007). https://doi.org:10.1177/1362361307078134

30 Chawarska, K., Macari, S. & Shic, F. Context modulates attention to social scenes in toddlers with autism. *Journal of Child Psychology and Psychiatry* **53**, 903-913 (2012). https://doi.org:10.1111/j.1469-7610.2012.02538.x

31 Shic, F., Wang, Q., Macari, S. L. & Chawarska, K. 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 (2020). https://doi.org:https://doi.org/10.1111/jcpp.13118

32 Forgeot d'Arc, B. *et al.* Atypical Social Judgment and Sensitivity to Perceptual Cues in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders* **46**, 1574-1581 (2016). https://doi.org:10.1007/s10803-014-2208-5
33 Lindahl, C. (2017).
34 Latimier, A. *et al.* Trustworthiness and Dominance Personality Traits' Judgments in Adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* **49**, 4535-4546 (2019). https://doi.org:10.1007/s10803-019-04163-1
35 Adolphs, R., Sears, L. & Piven, J. Abnormal Processing of Social Information from Faces in Autism. *Journal of Cognitive Neuroscience* **13**, 232-240 (2001). https://doi.org:10.1162/089892901564289
36 Liu, Z., Luo, P., Wang, X. & Tang, X. in *Proceedings of International Conference on Computer Vision (ICCV)*.
37 Gao, X. *et al.* The mutuality of social emotions: How the victim's reactive attitude influences the transgressor's emotional responses. *NeuroImage* **244**, 118631 (2021). https://doi.org:https://doi.org/10.1016/j.neuroimage.2021.118631
38 Happe, F., Ronald, A. & Plomin, R. Time to give up on a single explanation for autism. *Nat Neurosci* **9**, 1218-1220 (2006).
39 Wang, S. & Adolphs, R. Reduced specificity in emotion judgment in people with autism spectrum disorder. *Neuropsychologia* **99**, 286-295 (2017). https://doi.org:http://dx.doi.org/10.1016/j.neuropsychologia.2017.03.024
40 Pelphrey, K. *et al.* Visual Scanning of Faces in Autism. *J Autism Dev Disord* **32**, 249-261 (2002). https://doi.org:10.1023/a:1016374617369
41 de Wit, T. C. J., Falck-Ytter, T. & von Hofsten, C. Young children with Autism Spectrum Disorder look differently at positive versus negative emotional faces. *Research in Autism Spectrum Disorders* **2**, 651-659 (2008). https://doi.org:http://dx.doi.org/10.1016/j.rasd.2008.01.004
42 Wang, S. *et al.* Atypical Visual Saliency in Autism Spectrum Disorder Quantified through Model-Based Eye Tracking. *Neuron* **88**, 604-616 (2015). https://doi.org:http://dx.doi.org/10.1016/j.neuron.2015.09.042
43 Zebrowitz, L. A., Kikuchi, M. & Fellous, J.-M. Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology* **98**, 175-189 (2010). https://doi.org:10.1037/a0017990
44 Hugenberg, K., Young, S. G., Sacco, D. F. & Bernstein, M. J. Social categorization influences face perception and face memory. *The Oxford handbook of face perception*, 245-261 (2011).
45 Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep face recognition. (2015).
46 Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S. & O'Toole, A. J. Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification. *Cognitive Science* **43**, e12729 (2019). https://doi.org:https://doi.org/10.1111/cogs.12729
47 Keles, U., Lin, C. & Adolphs, R. A Cautionary Note on Predicting Social Judgments from Faces with Deep Neural Networks. *Affective Science* (2021). https://doi.org:10.1007/s42761-021-00075-5
48 Cao, R. *et al.* Feature-based encoding of face identity by single neurons in the human medial temporal lobe. *bioRxiv*, 2020.2009.2001.278283 (2020). https://doi.org:10.1101/2020.09.01.278283

49 Grossman, S. *et al.* Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications* **10**, 4934 (2019). https://doi.org:10.1038/s41467-019-12623-6

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

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

52 Todorov, A. (Princeton University Press, 2017).

53 Stolier, R. M., Hehman, E. & Freeman, J. B. A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences* **22**, 197-200 (2018). https://doi.org:https://doi.org/10.1016/j.tics.2017.12.003

54 Spek, A., Schatorjé, T., Scholte, E. & van Berckelaer-Onnes, I. Verbal fluency in adults with high functioning autism or Asperger syndrome. *Neuropsychologia* **47**, 652-656 (2009). https://doi.org:https://doi.org/10.1016/j.neuropsychologia.2008.11.015

55 Oliveras-Rentas, R. E., Kenworthy, L., Roberson, R. B., Martin, A. & Wallace, G. L. 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 (2012). https://doi.org:10.1007/s10803-011-1289-7

56 Neumann, D., Spezio, M. L., Piven, J. & Adolphs, R. 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 (2006). https://doi.org:10.1093/scan/nsl030

57 Spezio, M. L., Adolphs, R., Hurley, R. S. E. & Piven, J. Analysis of face gaze in autism using "Bubbles". *Neuropsychologia* **45**, 144-151 (2007). https://doi.org:http://dx.doi.org/10.1016/j.neuropsychologia.2006.04.027

58 Kliemann, D., Dziobek, I., Hatri, A., Steimke, R. & Heekeren, H. R. Atypical Reflexive Gaze Patterns on Emotional Faces in Autism Spectrum Disorders. *The Journal of Neuroscience* **30**, 12281-12287 (2010). https://doi.org:10.1523/jneurosci.0688-10.2010

59 Charalambides, N. (2021).

60 Maenner, M. J., Shaw, K. A. & Baio, J. 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 (2020).

61 Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* **47**, 1122-1135 (2015). https://doi.org:10.3758/s13428-014-0532-5

62 Chelnokova, O. *et al.* Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular Psychiatry* **19**, 746-747 (2014). https://doi.org:10.1038/mp.2014.1

63 DeBruine, L. & Jones, B. (2017).

64 McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**, 30-46 (1996). https://doi.org:10.1037/1082-989X.1.1.30

65 Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. *Advances in neural information processing systems*, 857-864 (2003).

66 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
67 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
68 Alber, M. *et al.* iNNvestigate neural networks! *J. Mach. Learn. Res.* **20**, 1-8 (2019).
69 Lang, P. J., Bradley, M. M. & Cuthbert, B. N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* **1**, 3 (1997).
70 Yu, H., Hu, J., Hu, L. & Zhou, X. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc Cogn Affect Neurosci* **9**, 1150-1158 (2014). https://doi.org:10.1093/scan/nst090
71 Yu, H., Duan, Y. & Zhou, X. Guilt in the eyes: Eye movement and physiological evidence for guilt-induced social avoidance. *Journal of Experimental Social Psychology* **71**, 128-137 (2017). https://doi.org:http://doi.org/10.1016/j.jesp.2017.03.007
72 Gao, X. *et al.* Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences* (2018).
73 Li, Z., Yu, H., Zhou, Y., Kalenscher, T. & Zhou, X. Guilty by association: How group-based (collective) guilt arises in the brain. *NeuroImage* **209**, 116488 (2020). https://doi.org:https://doi.org/10.1016/j.neuroimage.2019.116488
74 Lord, C. *et al.* Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *J Autism Dev Disord* **19**, 185–212 (1989).

Supplementary Materials

Supplementary Discussion

Possible caveats

We primarily used positive (e.g., *competent*, *warm*) and neutral trait descriptions in the present study. We used a symmetrical scale (e.g., 1 for *not competent* and 7 for *competent*) and the negative traits were thus built in. Future studies using negative trait descriptions (e.g., *incompetent*, *untrustworthy*) would better dissociate whether people with ASD tend to provide more intensive (i.e., higher) or more positive ratings. Furthermore, participants with ASD demonstrated a significantly lower inter-rater consistency compared to controls in most of the traits. This suggests that participants with ASD were more variable in their ratings, consistent with the heterogeneity in their symptoms and behavior [1]. Because we did not retest participants, we could not differentiate between-participant versus within-subject variability, although it has been shown that atypical gaze patterns in autism are heterogeneous across participants but reliable within individuals [2], consistent with our present results. It is also worth noting that we further confirmed our results using non-parametric permutation tests so such heterogeneity (variance in ratings) did not affect our comparisons on the central tendency in trait judgments.

In our main experiment, we used celebrity faces, and therefore, participants' personal knowledge of the celebrities may affect how they judge social traits from the celebrity faces [3-5]. Indeed, we observed a greater difference for faces that the participants were not familiar with compared to the familiar faces (**Supplementary Fig. 2c, d**), suggesting that the group differences in ratings were due to processing of faces and facial features rather than different levels of familiarity or semantic knowledge about the people (e.g., from watching different news). We were also able to replicate our results with a different set of unfamiliar faces alone, suggesting that our results could not simply be attributed to face familiarity.

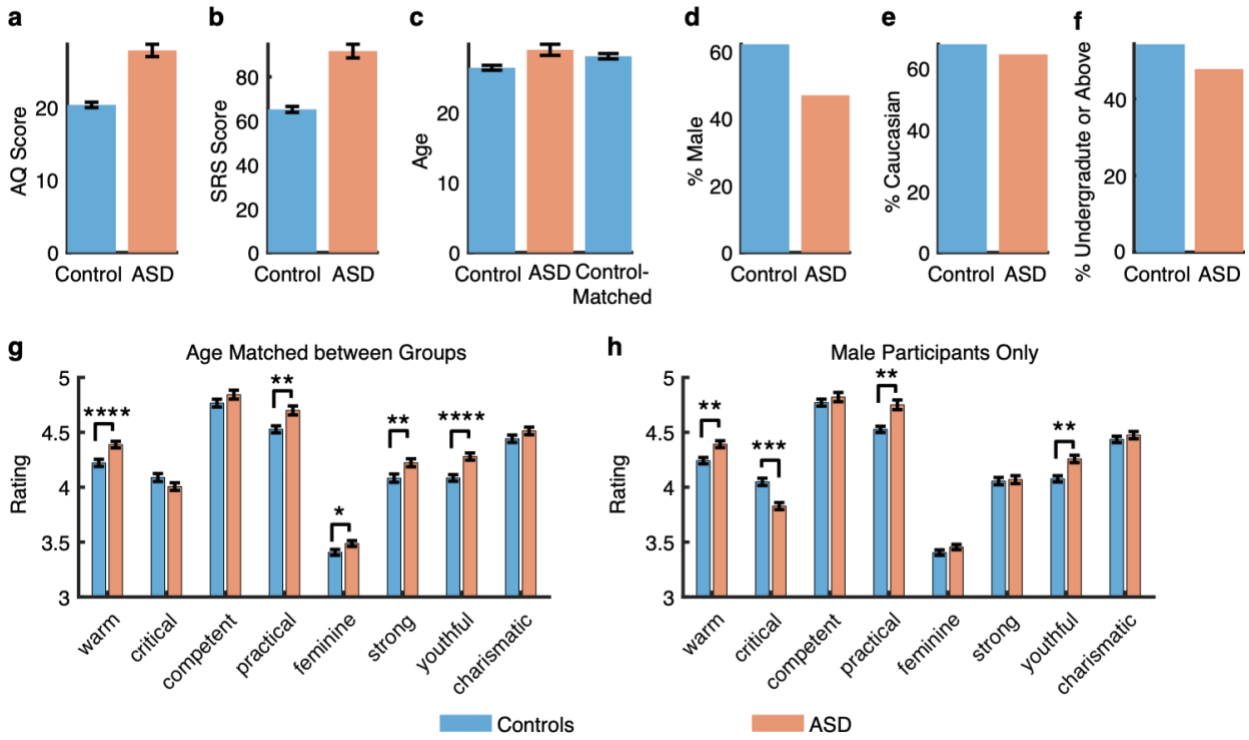
In addition, we used a DNN pre-trained for face identity recognition to construct face feature spaces for social trait comparisons. Such DNN can predict a wide range of social traits reliably [6]. Therefore, this DNN can extract facial information that is relevant to social judgment. In particular, this DNN shows an organized structure for face representation, which can facilitate our interpretation of what types of faces drive atypical social trait judgment in ASD. However, our

results may be specific to the DNN architecture. Future research comparing feature spaces from different pre-trained DNNs and using out-of-sample prediction (e.g., projecting novel faces to the discriminative regions [7]) will be useful for testing the generalizability of our results.

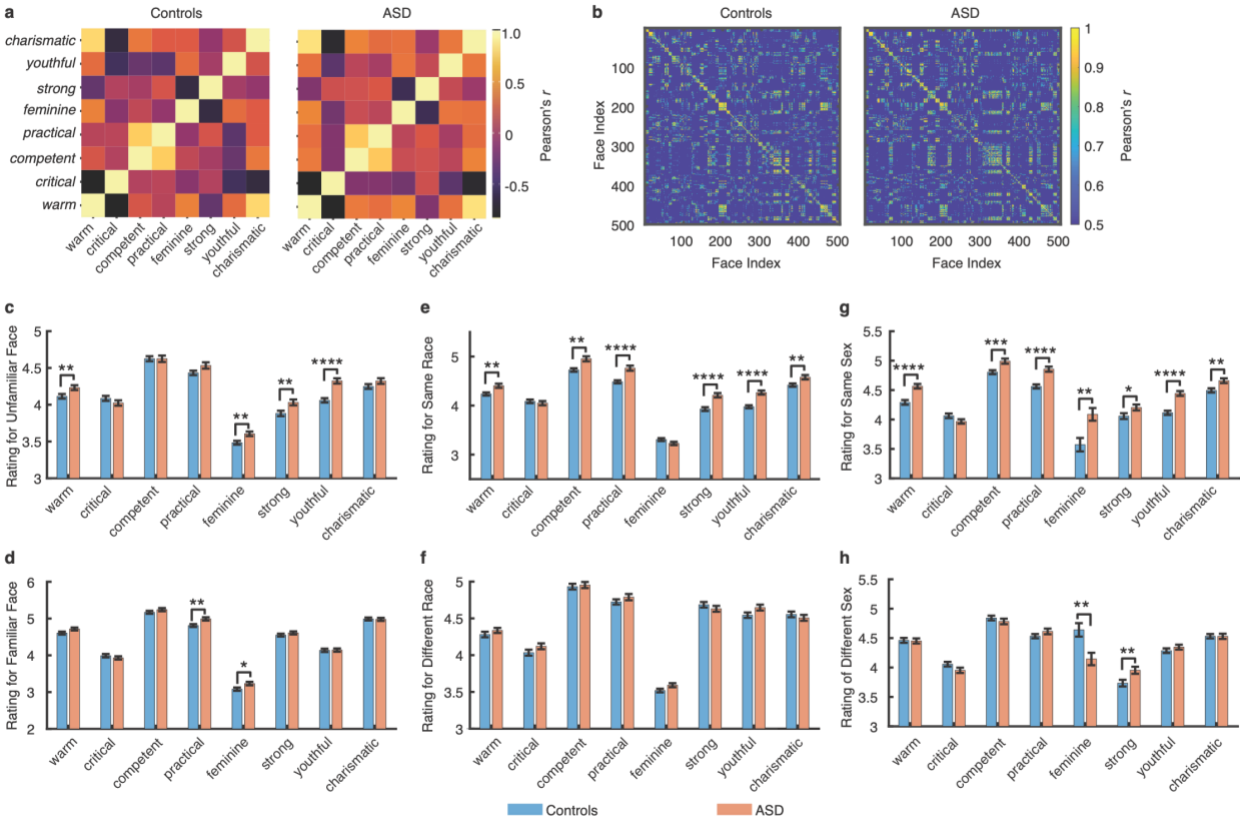
We found an intact PCA structure in participants with ASD but differences in individual ratings, which could be further explained by facial features derived using DNNs and t-SNE. Furthermore, we could accurately classify the raters. This was likely because PCA preserves the global structure of the data while t-SNE preserves the local structure (i.e., the difference was between dimensions versus clusters); PCA is sensitive to outliers while t-SNE is not; and PCA is linear and unique (except for rotation) while t-SNE is nonlinear and not unique. This was also likely because PCA computed a linear combination of features whereas the difference in ratings was disproportional for different traits.

With an independent sample of in-lab participants, we replicated three key results: the intact overall dimensional structure, the reduced inter-rater consistency, and the reduced rating specificity in ASD. Although in the main experiment online participants with ASD had more positive ratings whereas in the replication experiment in-lab participants with ASD had more negative ratings, such differences were likely resulted from reduced rating specificity: depending on the differences at the extremes, grand average could show either a positive or a negative difference between groups. In other words, the grand average might be an oversimplified metric to compare between groups. In addition, the differences in these results may be attributed to different compositions of participants (see **Methods** and **Table 1**) and sex differences in social trait judgment [8,9].

Supplementary Figures

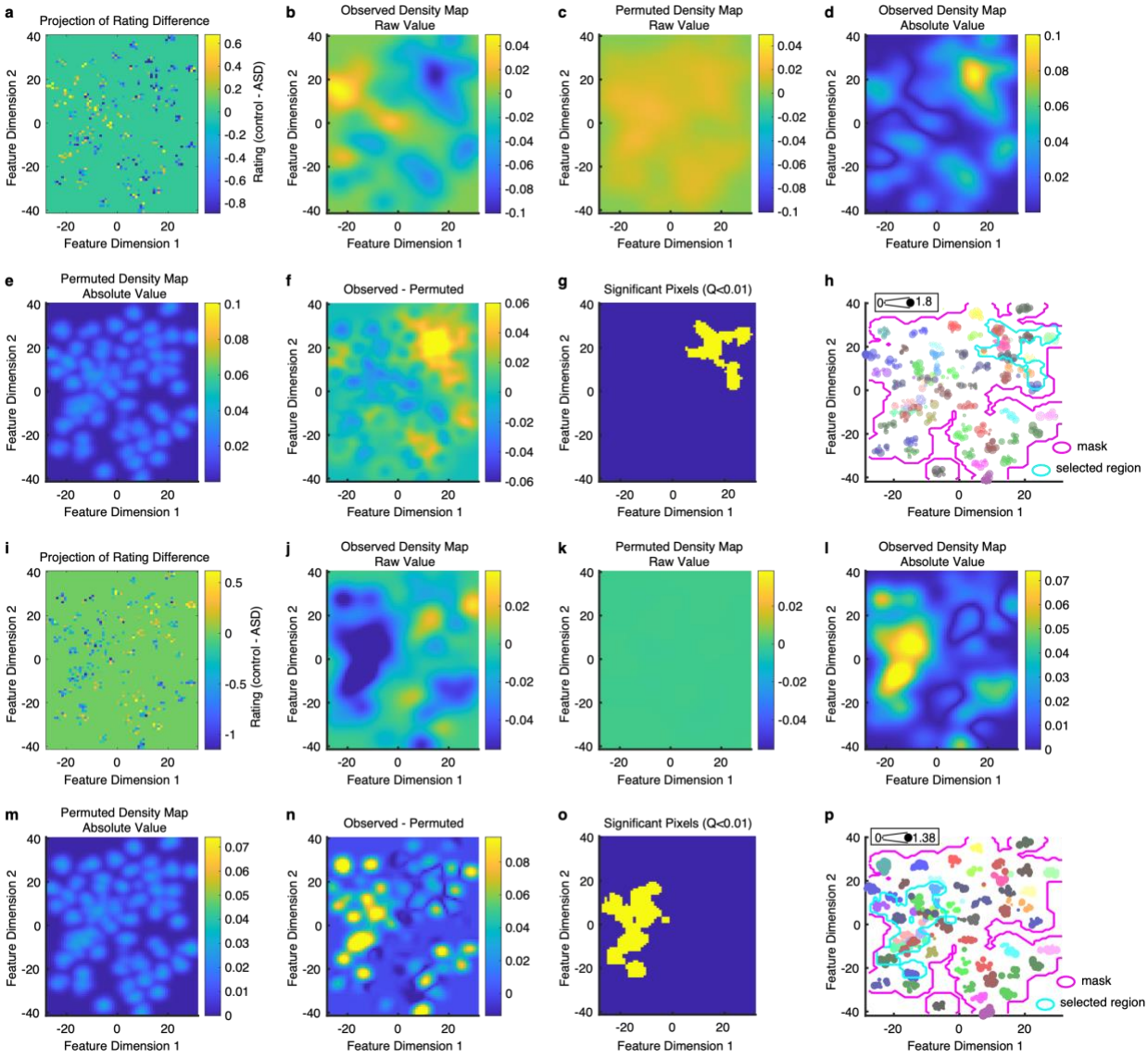


Supplementary Fig. 1. Characterization of participants. **(a)** Autism Spectrum Quotient (AQ). **(b)** Social Responsiveness Scale-2 Adult Self Report (SRS-A-SR). **(c)** Age. **(d)** Percentage of male participants. **(e)** Percentage of Caucasian participants. **(f)** Percentage of participants whose education level is undergraduate or above. Error bars denote ±SEM across participants. **(g)** Aggregate ratings from a subset of participants that matched in age. Significant difference was observed for traits *warm* (two-tailed two-sample *t*-test; $t(744) = 3.73$, $P = 0.00021$), *practical* ($t(723) = 3.14$, $P = 0.0017$), *feminine* ($t(658) = 1.99$, $P = 0.047$), *strong* ($t(744) = 2.64$, $P = 0.0084$), and *youthful* ($t(749) = 4.29$, $P = 2.04\times10^{-5}$). **(h)** Aggregate ratings from male participants only. Significant difference was observed for traits *warm* ($t(458) = 2.62$, $P = 0.0090$), *critical* ($t(447) = 3.49$, $P = 0.00053$), *practical* ($t(450) = 3.19$, $P = 0.0015$), and *youthful* ($t(463) = 3.09$, $P = 0.0021$). Error bars denote ±SEM across rating modules. Asterisks indicate a significant difference between participants with ASD and controls using two-tailed two-sample *t*-test. *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$, and ****: $P < 0.0001$.



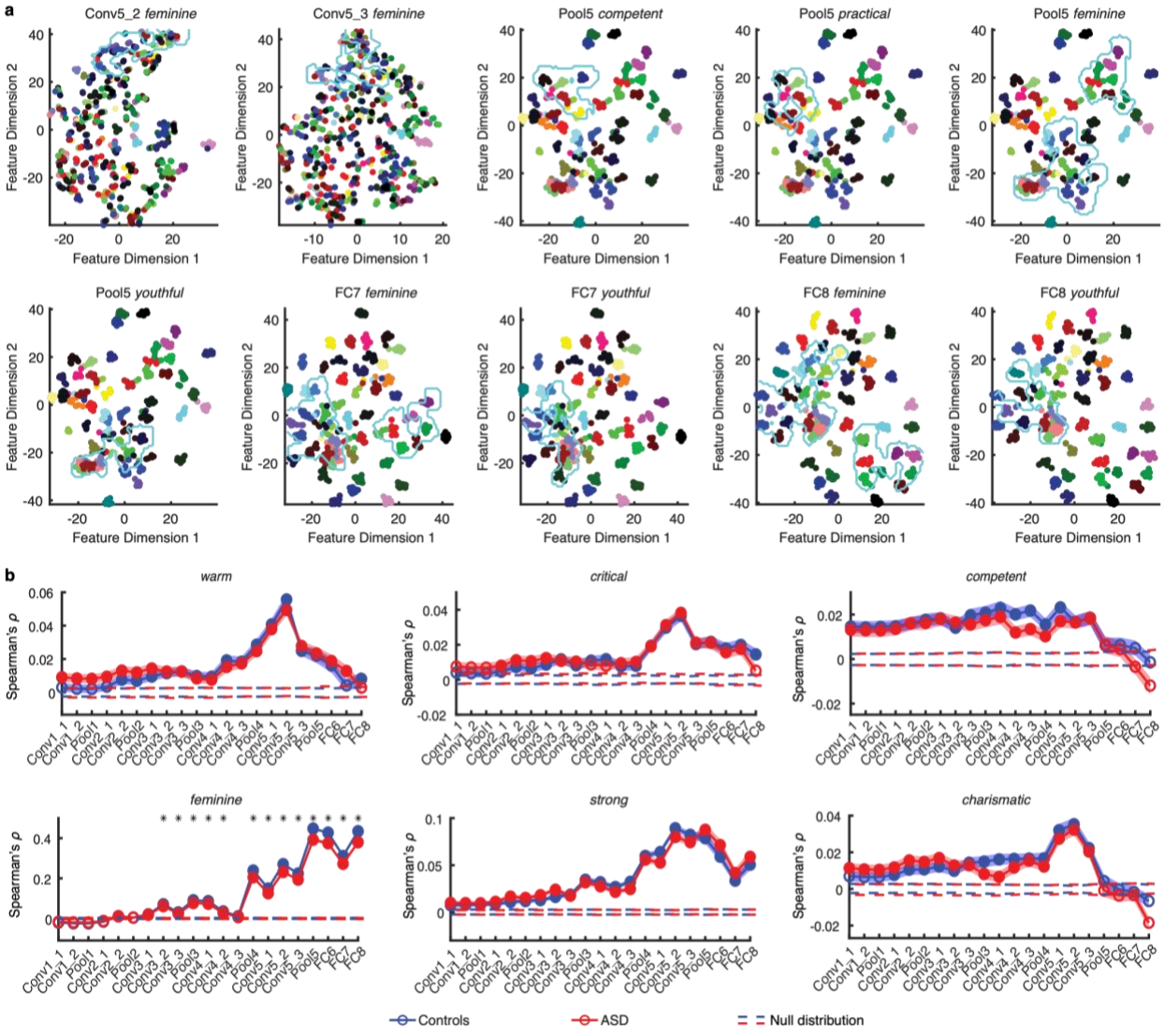
Supplementary Fig. 2. Additional analyses for social trait judgments **(a)** Pearson correlations between aggregate social trait judgments. **(b)** Similarity matrices calculated using face images. Color coding shows similarity values. Pearson correlation was calculated for each pair of faces (500 faces in total) across 8 social traits. Faces are organized according to identities (e.g., face indices 1 to 10 correspond to face identity 1, face indices 11 to 20 correspond to face identity 2, and so on). **(c, d)** Social trait judgment for **(c)** unfamiliar versus **(d)** familiar faces. Participants with ASD had a slightly higher rating when judging familiar identities on *practical* (two-tailed two-sample *t*-test; $t(794) = 2.95$, P = 0.0033) and *feminine* ($t(729) = 2.17$, P = 0.030) but had a substantially higher rating when judging unfamiliar identities on four traits: *warm* ($t(813) = 2.55$, P = 0.011), *feminine* ($t(727) = 2.83$, P = 0.0048), *strong* ($t(812) = 2.73$, P = 0.0065), and *youthful* ($t(818) = 5.58$, $P = 3.27\times10^{-8}$). **(e, f)** Social trait judgment for **(e)** same-race versus **(f)** different-race faces. Social traits primarily differed between groups in same-race faces (two-tailed two-sample *t*-test; *warm*: $t(719) = 2.93$, P = 0.0034; *competent*: $t(701) = 3.38$, P = 0.00077; *practical*: $t(697) = 4.23$, $P = 2.62\times10^{-5}$; *strong*: $t(712) = 4.38$, $P = 1.36\times10^{-5}$; *youthful*: $t(721) = 5.23$, $P = 2.21\times10^{-7}$; *charismatic*: $t(711) = 2.63$, P = 0.0086) rather than different-race faces. We also found

that race information could modulate social trait judgments for *competent*, *practical*, *feminine*, *strong*, *youthful*, and *charismatic* in controls, and *competent*, *practical*, *feminine*, *strong*, and *youthful* in participants with ASD (two-tailed paired *t*-test: all Ps < 0.05). **(g, h)** Social trait judgment for **(g)** same-sex versus **(h)** different-sex faces. Social traits primarily differed between groups in same-sex faces (two-tailed two-sample *t*-test; *warm*: $t(813) = 4.87$, $P = 1.34\times10^{-6}$; *competent*: $t(796) = 3.36$, $P = 0.00082$; *practical*: $t(791) = 5.17$, $P = 2.93\times10^{-7}$; *feminine*: $t(725) = 3.07$, $P = 0.0022$; *strong*: $t(811) = 2.08$, $P = 0.038$; *youthful*: $t(818) = 6.04$, $P = 2.33\times10^{-9}$; *charismatic*: $t(808) = 3.00$, $P = 0.0027$) rather than different-sex faces (*feminine*: $t(725) = 2.97$, $P = 0.0030$; *strong*: $t(811) = 2.64$, $P = 0.0083$). We also found that sex information could modulate social trait judgments for *warm*, *feminine*, *strong*, and *youthful* in controls, and *warm*, *competent*, *practical*, *strong*, *youthful*, and *charismatic* in participants with ASD (two-tailed paired *t*-test: all Ps < 0.05). Error bars denote ±SEM across rating modules. Asterisks indicate a significant difference between participants with ASD and controls using two-tailed two-sample *t*-test. *: P < 0.05, **: P < 0.01, ***: P < 0.001, and ****: P < 0.0001.

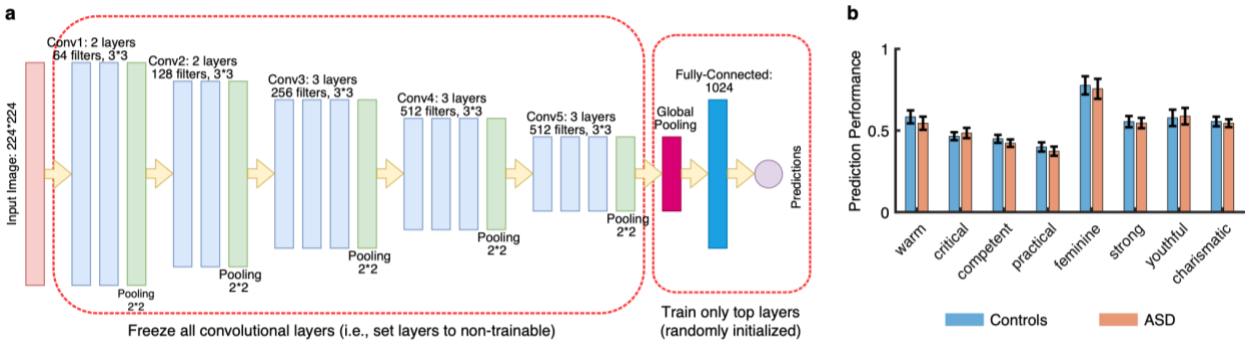


Supplementary Fig. 3. Illustration of the procedure for identifying discriminative regions in the face feature space. **(a-h)** Trait *practical* in the layer FC6. **(i-p)** Trait *youthful* in the layer FC6. **(a, i)** Projection of the rating difference (control − ASD) onto the feature space. **(b, j)** Raw values were smoothed by a Gaussian kernel to derive density maps for observed distributions. **(c, k)** Density maps for permuted distributions. **(d, e, l, m)** Density maps were transformed to absolute values for statistical comparisons. **(d, l)** Density maps for observed distributions. **(e, m)** Density maps for permuted distributions. **(f, n)** The difference maps between observed and permuted distributions. **(g, o)** Statistically significant pixels identified by comparing the observed distribution to the permuted distribution (permutation test for each pixel: $P < 0.01$, corrected by false discovery rate [FDR] [10]). **(h, p)** Identified regions after thresholding for the minimum number

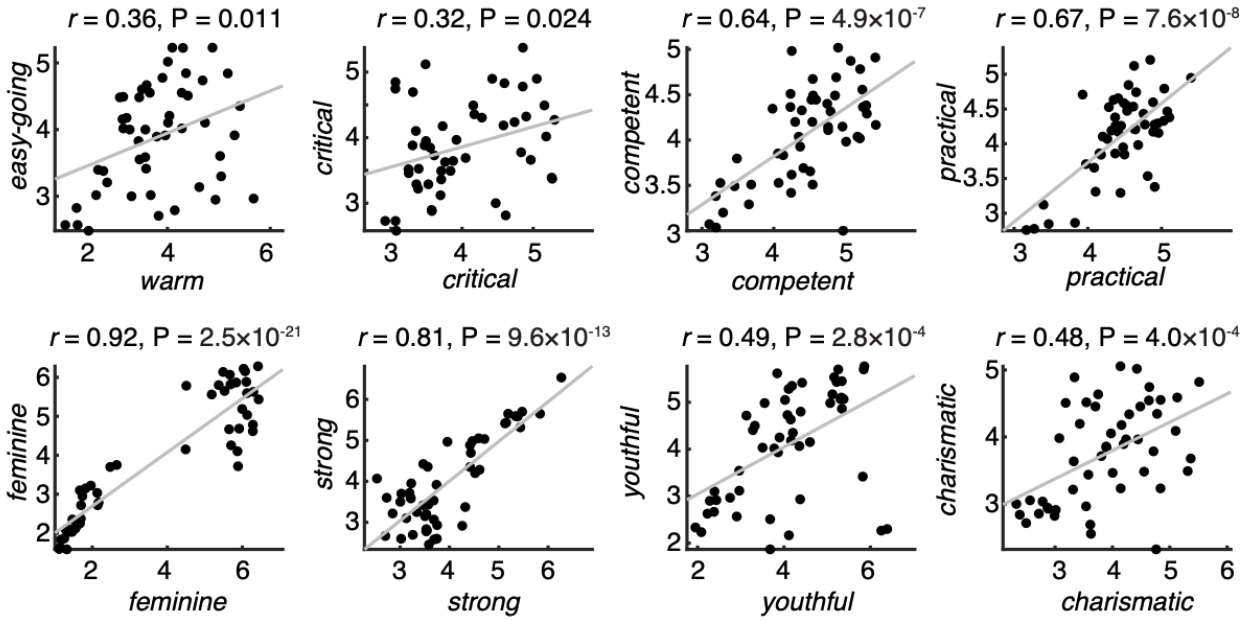
of pixels within the cluster. A mask (shown in magenta) was first applied to exclude pixels from the edges and corners where there were no faces because the regions with a small number of faces (i.e., the samples were sparse) were susceptible to false positives. Cluster size must be greater than 5% of the total number of pixels of the face space within the mask because small clusters were likely to be false positive. Each color represents a different identity. The size of the dot indicates the absolute value of the rating difference between groups.



Supplementary Fig. 4. Additional results for features across faces that contribute to atypical trait ratings in ASD. **(a)** Discriminative regions for social traits across deep neural network (DNN) layers. **(b)** Representation similarity between social trait judgment ratings and DNN features for each DNN layer. Legend conventions as in **Fig. 2**. *: $P < 0.05$.



Supplementary Fig. 5. Model structure and performance for social trait prediction. **(a)** Model structure. We applied transfer learning to a VGG-16 deep neural network (DNN) to build a DNN-based regression model for each trait and each participant group. We kept all convolution layers, but replaced the last two fully-connected (FC) layers and the output layer with a global averaging pooling layer, a FC layer, and a prediction output layer. During model training, all convolutional layers were frozen and only the top layers were updated by training. **(b)** Model performance. Model performance was assessed by the correlation between the observed trait values and the predicted trait values in the testing set. The correlation coefficient (Pearson's *r*) could indicate the model's predictability. Error bars denote ±SEM across cross-validation runs.



Supplementary Fig. 6. Correlation of ratings between controls from the present study (the second control experiment) and controls from a previous study [11] for the unfamiliar faces. Each dot represents a face, and the gray line denotes the linear fit.

Supplementary References

1 Happe, F., Ronald, A. & Plomin, R. Time to give up on a single explanation for autism. *Nat Neurosci* **9**, 1218-1220 (2006).

2 Keles, U. *et al.* Atypical gaze patterns in autism are heterogeneous across subjects but reliable within individuals. *bioRxiv*, 2021.2007.2001.450793 (2021). https://doi.org:10.1101/2021.07.01.450793

3 Gordon, I. & Tanaka, J. W. The role of name labels in the formation of face representations in event-related potentials. *British Journal of Psychology* **102**, 884-898 (2011). https://doi.org:https://doi.org/10.1111/j.2044-8295.2011.02064.x

4 Schwartz, L. & Yovel, G. The roles of perceptual and conceptual information in face recognition. *Journal of Experimental Psychology: General* **145**, 1493-1511 (2016). https://doi.org:10.1037/xge0000220

5 Oh, D., Walker, M. & Freeman, J. B. Person knowledge shapes face identity perception. *Cognition* **217**, 104889 (2021). https://doi.org:https://doi.org/10.1016/j.cognition.2021.104889

6 Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S. & O'Toole, A. J. Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification. *Cognitive Science* **43**, e12729 (2019). https://doi.org:https://doi.org/10.1111/cogs.12729

7 Cao, R. *et al.* Feature-based encoding of face identity by single neurons in the human medial temporal lobe. *bioRxiv*, 2020.2009.2001.278283 (2020). https://doi.org:10.1101/2020.09.01.278283

8 Wallach, M. A. & Kogan, N. Sex differences and judgment processes1. *Journal of Personality* **27**, 555-564 (1959). https://doi.org:https://doi.org/10.1111/j.1467-6494.1959.tb01883.x

9 Bosak, J., Sczesny, S. & Eagly, A. H. The Impact of Social Roles on Trait Judgments: A Critical Reexamination. *Personality and Social Psychology Bulletin* **38**, 429-440 (2011). https://doi.org:10.1177/0146167211427308

10 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

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