Neurobiology of Disease

# A Computational Probe into the Behavioral and Neural Markers of Atypical Facial Emotion Processing in Autism

**Kohitij Kar**[1,2]

[1]McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 01239, and [2]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, Massachusetts 01239

**Despite ample behavioral evidence of atypical facial emotion processing in individuals with autism spectrum disorder (ASD), the neural underpinnings of such behavioral heterogeneities remain unclear. Here, I have used brain-tissue mapped artificial neural network (ANN) models of primate vision to probe candidate neural and behavior markers of atypical facial emotion recognition in ASD at an image-by-image level. Interestingly, the image-level behavioral patterns of the ANNs better matched the neurotypical subjects 'behavior than those measured in ASD. This behavioral mismatch was most remarkable when the ANN behavior was decoded from units that correspond to the primate inferior temporal (IT) cortex. ANN-IT responses also explained a significant fraction of the image-level behavioral predictivity associated with neural activity in the human amygdala (from epileptic patients without ASD), strongly suggesting that the previously reported facial emotion intensity encodes in the human amygdala could be primarily driven by projections from the IT cortex. In sum, these results identify primate IT activity as a candidate neural marker and demonstrate how ANN models of vision can be used to generate neural circuit-level hypotheses and guide future human and nonhuman primate studies in autism.**

*Key words:* **amygdala; artificial neural networks; autism; facial emotion recognition; inferior temporal cortex**

**Significance Statement**

**Moving beyond standard parametric approaches that predict behavior with high-level categorical descriptors of a stimulus (e.g., level of happiness/fear in a face image), in this study, I demonstrate how an image-level probe, using current deep-learning-based ANN models, allows identification of more diagnostic stimuli for autism spectrum disorder enabling the design of more powerful experiments. This study predicts that IT cortex activity is a key candidate neural marker of atypical facial emotion processing in people with ASD. Importantly, the results strongly suggest that ASD-related atypical facial emotion intensity encodes in the human amygdala could be primarily driven by projections from the IT cortex.**

## Introduction

The ability to recognize others' mood, emotion, and intent from facial expressions lies at the core of human interpersonal communication and social engagement. This relatively automatic, visuocognitive feature that neurotypically developed human adults take for granted shows significant differences in children and adults with autism (Adolphs et al., 2001; Golarai et al., 2006; Kennedy and Adolphs, 2012; Wang and Adolphs, 2017). Currently we lack a mechanistic and computational understanding of the underlying neural correlates of such behavioral mismatches.

There is a growing body of work on how facial identity is encoded in the primate brain, especially in the fusiform face areas in humans (Kanwisher et al., 1997; Tsao and Livingstone, 2008) and in the topographically specific face patch systems of the inferior temporal (IT) cortex of the rhesus macaques (Tsao et al., 2003, 2008; Freiwald et al., 2009). Also, previous research has linked human amygdala neural responses with recognizing facial emotions (Adolphs et al., 1994; Adolphs, 2008; Rutishauser et al., 2015). For instance, subjects who lack a functional amygdala often exhibit selective impairments in recognizing fearful faces (Broks et al., 1998; Adolphs et al., 1999). Wang et al. (2017) also demonstrated that the human amygdala parametrically encodes the intensity of specific facial emotions (e.g., fear, happiness) and their categorical ambiguity. A critical question, however, is whether the atypical facial emotion recognition broadly reported in individuals with autism spectrum disorder (ASD) arises purely from differences in sensory representations (i.e., purely perceptual alterations (Behrmann et al., 2006a; Robertson and Baron-Cohen, 2017) or because of a primary (but not mutually exclusive) variation in the

Received Nov. 8, 2021; revised May 7, 2022; accepted May 16, 2022.

Author contributions: K.K. designed research; K.K. performed research; K.K. analyzed data; and K.K. wrote the paper.

The author thanks Ralph Adolphs, Pawan Sinha (and Sinha Lab members), and James J. DiCarlo for comments and discussions; Shuo Wang for sharing the behavioral and neural datasets used in this study; and Shuo Wang, S. Sanghavi, Alina Peter, and Yoon Bai for comments on the manuscript.

The author declares no competing financial interests.

Correspondence should be addressed to Kohitij Kar at kohitij@mit.edu.

https://doi.org/10.1523/JNEUROSCI.2229-21.2022

development and function of specialized affect processing regions (e.g., atypical amygdala development leading to specific differences in encoding emotion). There are two main roadblocks in answering this question. First, heterogeneity and idiosyncrasies are commonplace across behavioral reports in autism, including facial affect processing (formal meta-analysis of recognition of emotions in autism, Uljarevic and Hamilton, 2013; Lozier et al., 2014). The inability to parsimoniously explain such heterogeneous findings prevent us from designing more efficient follow-up experiments to probe such questions further. Second, in the absence of neurally mechanistic models of behavior, it remains challenging to infer neural mechanisms from behavioral results and generate testable neural-circuit-level predictions that can be validated or falsified using neurophysiological approaches. Therefore, we need brain-mapped computational models that can predict at an image-by-image level how primates represent facial emotions across different parts of their brain and how such representations are linked to their performance in facial emotion judgment tasks like the one used in Wang and Adolphs (2017).

The differences in facial emotion judgments between neurotypical adults and individuals with autism are often interpreted with inferential models (e.g., psychometric functions) that base their predictions on high-level categorical descriptors of the stimuli (e.g., overall facial expression level of happiness, fear, and other primary emotions (Ekman and Keltner, 1997). Such modeling efforts are likely to ignore an important source of variance produced by the image-level sensory representations of each stimulus being tested. To interpret this source of variance, it is necessary to develop models that are image computable. Recent progress in computer vision and computational neuroscience has led to the development of artificial neural network (ANN) models that can both perform human-like object recognition (Rajalingham et al., 2015, 2018) and contain internal components that match human and macaque visual systems (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). Such image-computable ANNs can generate testable neural hypotheses (Bashivan et al., 2019; Kar et al., 2019) and help design experiments that leverage the image-level variance to guide us beyond the standard parametric approaches.

In this study, I have used a family of brain-tissue-mapped ANN models of primate vision to generate testable hypotheses and identify candidate neural and behavior markers of atypical facial emotion recognition in people with ASD. Specifically, I have compared the predictions of ANN models with behavior measured in neurotypical adults and people with autism (Wang and Adolphs, 2017) and facial emotion decodes from neural activity measured in the human amygdala (Wang et al., 2017). I observed that the ANNs could accurately predict the human facial emotion judgments at an image-by-image level. Interestingly, the image-level behavioral patterns of the models better matched the neurotypical human subjects' behavior than those measured in individuals with autism. This behavioral mismatch was most remarkable when the model behavior was constructed from units that correspond to the primate IT cortex. Interestingly, I also observed this behavioral mismatch when comparing neural decodes from a distinct population of visually facilitated neurons in the human amygdala with *Control* and ASD behavior. However, ANN-IT activation patterns could fully account for the image-level behavioral predictivity of the human amygdala population responses that has been previously implicated in autism-related facial emotion processing differences (Rutishauser et al., 2015; Wang et al., 2017). Furthermore, *in silico* experiments revealed that learning the facial emotion discrimination task with noisier ANN-IT representations (i.e., with higher response variability per unit) results in weaker synaptic connections between the model-IT and the downstream decision unit that increase the match of the model to the image-level behavioral patterns measured in the ASD population. In sum, these results argue that atypical sensory representations in the primate inferior temporal cortex that drive a distinct population of neurons in the human amygdala is a key candidate mechanism of atypical facial emotion processing in individuals with autism, a testable neural hypothesis for future human and nonhuman primate studies.

## Materials and Methods

### *Human behavior*

In this study, I have reanalyzed behavioral data previously collected and used in a study by Wang and Adolphs (2017; raw behavioral dataset from S. Wang and R. Adolphs, personal communication).

#### *Participants*

In the original study (Wang and Adolphs, 2017), 18 high-functioning participants with ASD (15 male) were recruited. All ASD participants met *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, and *International Classification of Diseases*, Revision 10, diagnostic criteria for ASD, and they met the cutoff scores for ASD on the Autism Diagnostic Observation Schedule Second Edition revised scoring system for Module 4 and the Autism Diagnostic Interview Revised or the Social Communication Questionnaire when an informant was available. The ASD group had a full-scale IQ (FSIQ) of 105 $\pm$ 13.3 (from the Wechsler Abbreviated Scale of Intelligence, Second Edition), a mean age of 30.8 $\pm$ 7.40 years, a mean Autism Spectrum Quotient (AQ) of 29.3 $\pm$ 8.28, a mean Social Responsiveness Scale Adult Self Report (SRS-A-SR) of 84.6 $\pm$ 21.5, and a mean Benton score of 46.1 $\pm$ 3.89 (Benton scores 41–54 were in the normal range). The Autism Diagnostic Observation Schedule (ADOS) item scores were not available for two participants, so we were unable to use the revised scoring system. But these individuals' original ADOS algorithm scores all met the cutoff scores for ASD.

Fifteen neurologically and psychiatrically healthy participants with no family history of ASD (11 male) were recruited as *Controls*. *Controls* had a comparable FSIQ of 107 $\pm$ 8.69 (two-tailed *t* test, $p$ = 0.74) and a comparable mean age of 35.1 $\pm$ 11.4 years ($p$ = 0.20) but a lower AQ (17.7 $\pm$ 4.29, $p = 4.62 \times 10^{-5}$) and SRS-A-SR (51.0 $\pm$ 30.3, $p$ = 0.0039) as expected. Participants gave written informed consent, and all original experiments were approved by the California Institute of Technology Institutional Review Board. All participants had normal or corrected-to-normal visual acuity. No enrolled participants were excluded for any reasons.

#### *Facial emotion judgment task*

During the task, Wang and Adolphs (2017) asked participants to discriminate between two emotions, fear and happiness. The image set includes faces of four individuals (two female) from the Stoichiometric Traits of Organisms in Their Chemical Habitats database (Roy et al., 2007), each showing fear and happiness expressions, which are highly recognizable emotions. To generate the morphed expression continua for the experiments, Wang and Adolphs (2017) interpolated pixel value and location between fearful exemplar faces and happy exemplar faces using a piecewise cubic-spline transformation over a Delaunay tessellation of manually selected control points. They created five levels of fear-happy morphs, ranging from 30% fear and 70% happy to 70% fear and 30% happy in steps of 10% (Fig. 1*B*). Low-level image properties were equalized using the SHINE (spectrum, histogram, and intensity normalization and equalization) toolbox (Willenbockel et al., 2010). In each trial, a face was presented for 1 s followed by a question prompt asking participants to make the best guess of the facial emotion (Fig. 1*A*). After stimulus offset, participants had 2 s to respond, otherwise the trial was aborted and discarded. Participants were instructed to

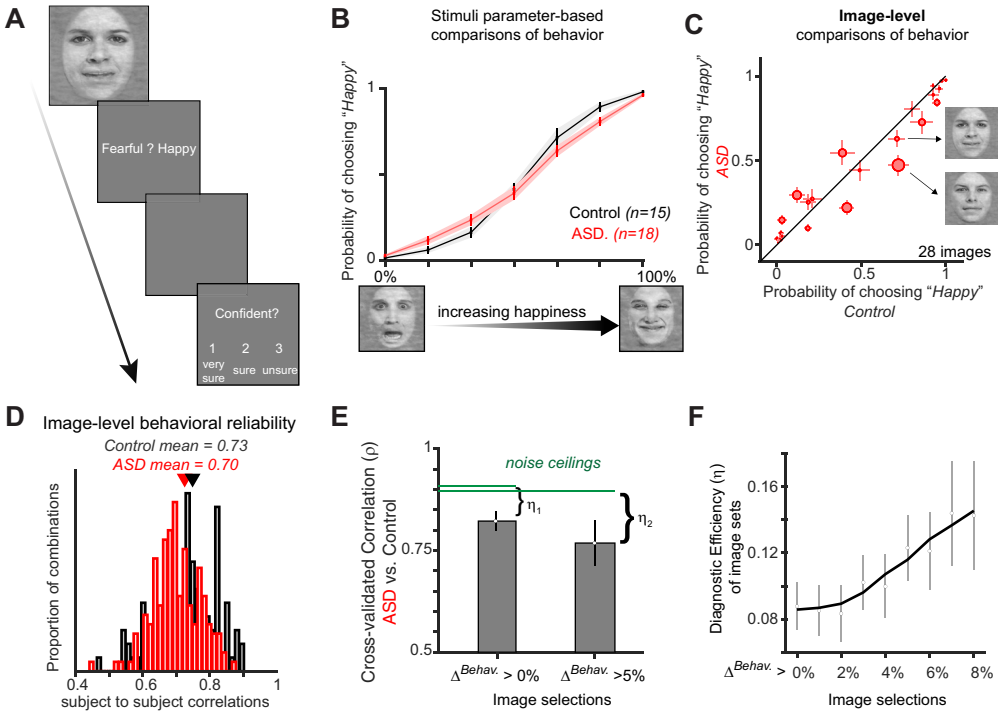


**Figure 1.** Behavioral task and image-level assessment of behavioral markers. ***A***, Subjects, both a neurotypical (*Control*, $n = 15$) population and individuals with autism (ASD, $n = 18$) viewed a face for 1 s in their central ~12°, followed by a question asking them to identify the facial emotion (fearful or happy). After a blank screen of 500 ms, subjects were then asked to indicate their confidence in their decision (1 for very sure, 2 for sure, or 3 for unsure). ***B***, The psychometric curves show the proportion of trials judged as happy as a function of facial emotion morph levels ranging from 0% happy (100% fearful, left) to 100% happy (0% fearful, right). ASD (red curve), on average, showed lower specificity (slope of the psychometric curve) compared with the *Controls* (black curve). The shaded area and error bars denote SEM across participants. ***C***, Image-level differences in behavior between *Controls* versus ASD. Each red dot corresponds to an image. The size of the dot is scaled by the difference in behavior between the *Controls* and ASD. Error bars denote SEM across subjects. Right, Two example images show similar emotional (happiness) judgments by the *Controls* but drive significantly different behaviors in ASD, demonstrating the importance of investigating individual image-level differences. ***D***, The estimated image-by-image happiness judgments were highly reliable as demonstrated by comparisons across individuals (estimated separately for each group). The mean reliability (average of the individual subject to subject correlations) was 0.73 and 0.70 for the *Controls* (black histogram) and ASD (red histogram), respectively. ***E***, Correlation between image-by-image behavioral patterns measured in *Controls* versus ASD, with two different selections of images (cross-validated image selections with held-out subjects). Noise ceilings were calculated based on measured behavioral (split half) reliability across populations within each group (see above, Materials and Methods). The difference between the noise ceiling and the mean raw correlation is referred to as the diagnostic efficiency of the image set ($\eta$). ***F***, Diagnostic efficiency ($\eta$) as a function of image selection criteria. Error bars denote bootstrap confidence intervals.

respond as quickly as possible but only after stimulus offset. No feedback message was displayed, and the order of faces was completely randomized for each participant. Images were presented approximately in the central 12° of visual angle. A subset of the participants (11 participants with autism and 11 *Controls*) also performed confidence ratings after emotion judgment and a 500 ms blank screen, participants were asked to indicate their confidence by pushing the button labeled 1 for very sure, 2 for sure, or 3 for unsure. This question also gave participants 2 s to respond.

*Depth recording in human amygdala*

In this study I have reanalyzed the neural data that was previously collected and used in a study by Wang et al.(Wang et al., 2017). The raw neural dataset was kindly shared via personal communication. Wang et al. (2017) recorded bilaterally from implanted depth electrodes in the amygdala from patients with pharmacologically intractable epilepsy. Target locations in the amygdala were verified using postimplantation structural MRIs. At each site, they recorded from eight 40 $\mu$m microwires inserted into a clinical electrode. Bipolar wideband recordings (0.1–9 kHz), using one of the eight microwires as reference, were sampled at 32 kHz and stored continuously for off-line analysis with a Neuralynx system (Cheetah software). The raw signal was filtered with a zero-phase lag 300–3000 Hz bandpass filter, and spikes were sorted using a semiautomatic template-matching algorithm. Units were carefully isolated, and spike sorting quality were assessed quantitatively. Subjects were presented each image for 1 s (similar to the task description above) to discriminate between two emotions, fear and happiness.

*Experimental design and statistical analysis*

*Estimating image-level behavioral reliability.* To estimate the image-level behavioral reliability (Fig. 1*D*), I first estimated the probability of choosing the happy image per image in each subject (15 *Controls*, 18 ASD), referred to as the $\vec{P}_C$ and the $\vec{P}_{IwA}$ vectors. Then, for each possible combination of selecting two subjects from the subject pools, I estimated the subject-to-subject Kendall rank correlation coefficient. This was done separately for the *Controls* and ASD, leading to the red and black histograms in Figure 1*D*, respectively. These correlation scores are not corrected by the individual subjects' internal reliability (across trials). Therefore, they represent the lower bound of the intersubject correlations.

*Estimating noise ceilings for ASD versus* Control *correlations.* I define the noise ceiling of a correlation as the highest possible value of a correlation expected given the noise measured independently in the two variables that are being tested. To estimate this, first I individually estimate the split-half reliability of the $\vec{P}_C$ and the $\vec{P}_{ASD}$ vectors, for the *Control* and ASD groups, respectively. Each split is constructed with a random sampling of half of the subjects and taking the average across them and doing same for the other half of the subjects. For each iteration, such splits were made, and the correlation between the resulting vectors was computed. This correlation score was corrected by the Spearman–Brown correction procedure to account for the halving of subject numbers. I then computed the average across 100 such iterations, referred to as $\rho_{\vec{P}_{C^1},\vec{P}_{C^2}}$ and $\rho_{\vec{P}_{ASD^1},\vec{P}_{ASD^2}}$ for the *Controls* and ASD, respectively. The noise ceiling was then estimated as follows:

$$\sqrt{\rho_{\vec{P}_{C^1},\vec{P}_{C^2}} * \rho_{\vec{P}_{ASD^1},\vec{P}_{ASD^2}}}.$$

Intuitively, if both groups provided noiseless data, then these reliabilities should be each at one, and therefore the noise ceiling shall also be set at one. Noisy data will lead to <1 values for the individual $\rho_{\vec{P}_{C^1},\vec{P}_{C^2}}$ and $\rho_{\vec{P}_{ASD^1},\vec{P}_{ASD^2}}$ reliabilities, and hence the noise ceiling shall also be <1. Of note, each selection of image with result in a different $\vec{P}$ vector and therefore will result in a slightly different noise ceiling estimate, as demonstrated in Fig. 1*E* (two green lines).

*Estimating cross-validated diagnostic efficiency of image sets.* Diagnostic efficiency ($\eta$; Fig. 1*E*,*F*) of an image set is defined as the cross-validated estimate of the difference between the noise ceiling and the raw correlation between the $\vec{P}_C$ and the $\vec{P}_{ASD}$ vectors. The cross-validation is achieved by choosing the images based on a specific subset of subjects and then measuring the noise ceiling and the raw correlation on a different held-out set of subjects. For efficient collection of human subject data that could optimally discriminate between the behavior measured in *Controls* and ASD, one must aspire for the highest $\eta$ values for image sets.

*Selection of neurons for analyses.* In the original study, only units with an average firing rate of at least 0.2 Hz (entire task) were considered, and only single units were considered. In addition, in this study I have further restricted the neural dataset to neurons that have a significant visual response (both increase and decrease). To estimate that, I compared the neural firing rates (per image) averaged across two specific time bins, [−1000, 0] and [250, 1250], where 0 is the onset of the image. If the paired Wilcoxon signed-rank test between these two firing rate vectors was significant, the site was considered for further analyses. Thus, I considered 156 total neurons, 99 visually facilitated (VF) neurons and 57 visually suppressed (VS) neurons.

*Decoding facial emotion judgment from neural population activity.* To decode facial emotion judgments from the neural responses per image, I used a linear model that linked the neural responses to the levels of happiness (ground truth from image generation). Building the model essentially involves solving a regression problem estimating the weights ($\vec{w}$) per neuron and a *bias* term. I used a partial least squares regression procedure (MATLAB command, plsregress), using 15 retained components. I also used 10-fold cross-validation. For each fold, the model was trained (i.e., $\vec{w}$ and *bias* were estimated) using the data from the other nine folds (training data), and predictions were generated for the held-out fold (test images). This was repeated for each of the folds, and the entire procedure was repeated 100 times. The predictions of the trained neural model on the held-out test images were used for future correlation analyses. Given the training scheme, every image was assigned as the test-image once per iteration.

*ANN models of primate vision.* The term model in this study always refers to a specific modification of a pretrained ANN. For instance, I have used an ImageNet pretrained deep neural network, AlexNet, to build multiple models. Each model was constructed by deleting all layers succeeding a given layer. For instance, the fifth convolutional (*cnv5*) layer model was built by removing all layers of AlexNet that followed the output of its fifth convolutional layer. The feature activations from the fifth convolutional layer output were then trained with the linear regression procedure (similar to the neural decodes).

*Estimating model facial emotion judgment behavior.* To decode facial emotion judgments from the model responses per image, I used the same linear modeling approach that I used for the neural data (see above), which linked the model feature activations to the level of happiness (ground truth from image generation). The model features, per layer, were extracted using the MATLAB command activations for AlexNet (Krizhevsky et al., 2012), VGGFace (Parkhi et al., 2015), and EmotionNet (https://github.com/mathinking/FaceGenderAgeEmotionDetection) in MATLAB-R 2020b. For the CORnet-S (Kubilius et al., 2019) model, I used the code from https://github.com/dicarlolab/CORnet.

*Estimation of discriminatory index (d′).* The discrimination index was computed to quantify the difference between the match of the behavioral predictions of the ANNs (models per layer) and the behavior measured in *Controls* and ASD (Fig. 2*E*). It was calculated as follows:

$$\frac{\rho^{Control} - \rho^{IwA}}{\sqrt{\frac{1}{2} * (\sigma^2_{\text{Control}} + \sigma^2_{\text{IwA}})}},$$

where $\rho^{Control}$ and $\rho^{IwA}$ was the correlation between ANN predictions and behavior measured in *Controls* and ASD, respectively; and $\sigma_{Control}$ and $\sigma_{IwA}$ was the SD of the bootstrap estimates of the correlations with random subsampling features from the model layers. To make the comparisons fair across all layers, 1000 features were randomly subsampled (without repetition) 100 times to estimate the ANN predictions.

*Estimation of residuals between ANN-IT and behavioral predictions of human amygdala.* I first estimated the cross-validated test predictions ($ANN^{Pred}$) of behavioral patterns from an ANN-IT layer (e.g., AlexNet fc7 model used in the study) using the partial least squares regression method. The ground truth values of image-level facial happiness were used as the dependent variable in this analysis. Next, I used the same algorithm but with the human amygdala neural features (instead of the ANN-IT features) as the predictors to estimate the neurally decoded behavioral patterns ($Amygdala^{Pred}$). I then used a generalized linear regression model (MATLAB glmfit) to estimate the residues while using $ANN^{Pred}$ as the predictor and $Amygdala^{Pred}$ as the dependent variable. The square of the Pearson correlation [percentage of explained variance (%EV)] between this residue vector (one value per image) and the image-level behavioral vector (probability of choosing "Happy" per image) measured in the *Controls* is plotted (see Fig. 4, left, *y*-axis). These %EV values were corrected by the noise estimates in the behavioral data per image selection. In addition, all %EV values were estimated in a cross validated way, wherein the image selections and the final estimates were done based on different groups of subjects.

### In silico *model perturbation and training*

*Generation of activity scaled additive noise values.* To estimate how much noise shall be added to each unit (feature) of the model layer, I used the following procedure. First, I estimated the SD ($\sigma$, across all 28 images) of the activation distribution per unit in a noise-free model. The addition of noise was made proportional to this value. To vary noise levels, a scalar factor (*c*; Fig. 5*D*,*E*, *x*-axis) was multiplied with $\sigma$ per unit. For each unit, the noise added to the original activation was drawn from a normal distribution that had a SD of $c * \sigma$.

*Training the model with and without noise.* To simulate a learning scheme with noise, I modified the model feature activations in the following way. During training of the regression model (i.e., estimating $\vec{w}$ and *bias*), the noisy version of the model was generated by concatenating 1000 randomly drawn features (which were fixed for each iteration of the procedure), with 10 repetitions of the same features but with the added noise on top of it. This procedure was repeated several times to estimate the variance in the model predictions per noise level. For the noise-free model, the same 1000 randomly drawn features were repeated without addition of any noise.

### *Statistical tests*

All correlation scores reported in this study are Kendall rank coefficients (unless otherwise mentioned). For significance tests of correlations (between two variables of interest), I have used a bootstrapped permutation test. To do this, I first constructed a null hypothesis by mixing the two variables and then randomly drew (as many times as the number of elements in the original variable) with replacements two elements from the mixed dataset to create two vectors. These two vectors can be constructed multiple times (typically >100) and correlated. The resulting correlation distribution was considered as the null hypothesis. Then the true raw correlation was compared with this distribution to determine a *p* value of rejecting the null distribution.

### *Data and availability*

All data and code used in this study is available to download and use from https://github.com/kohitij-kar/2021_faceEmotion_ASD.

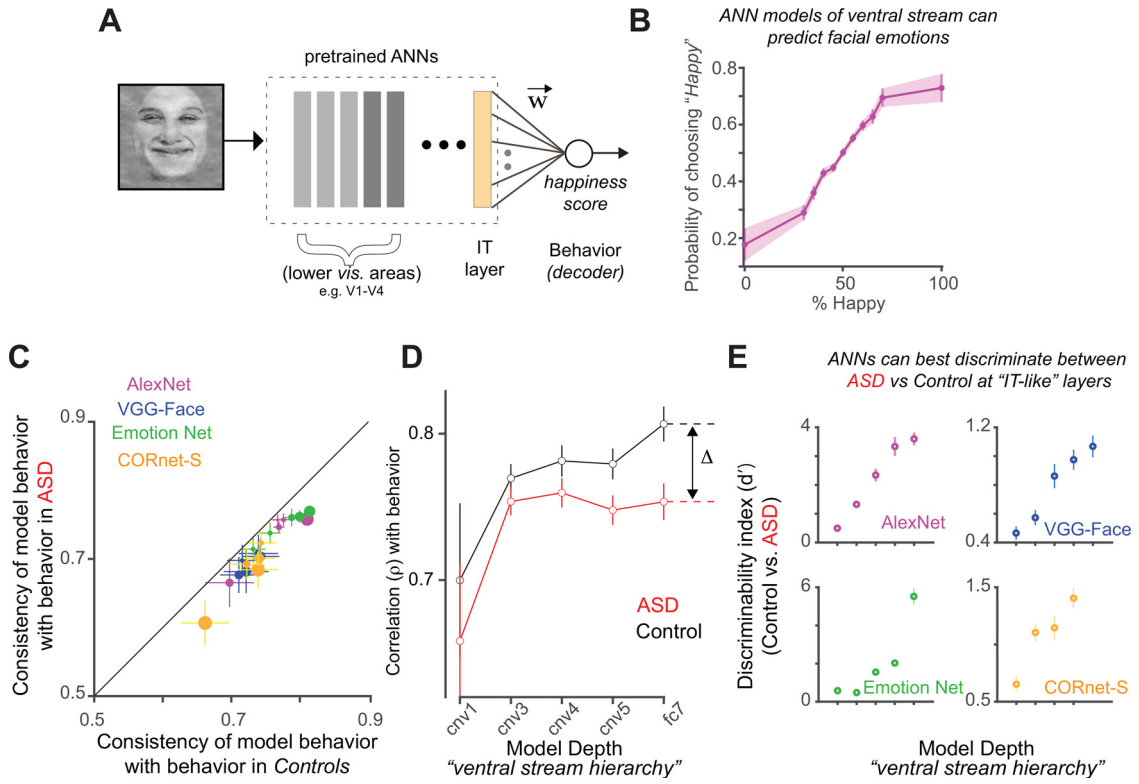


**Figure 2.** Testing ANN models on facial emotion recognition tasks. ***A***, ANN models of the primate ventral stream (typically comprising V1, V2, V4, and IT-like layers) can be trained to predict human facial emotion judgments. This involves building a regression model, that is, determining the weights $\vec{w}$ based on the model layer activations (as the predictor) to predict the image ground truth (level of happiness) on a set of training images and then testing the predictions of this model on held-out images. ***B***, The predicted psychometric curves of an ANN model (e.g., AlexNet, shown here) show the proportion of trials judged as happy as a function of facial emotion morph levels ranging from 0% happy (100% fearful, left) to 100% happy (0% fearful, right). This curve demonstrates that activations of ANN layers (layer fc7, which corresponds to the model-IT layer) can be successfully trained to predict facial emotions. ***C***, Comparison of the image-level behavioral patterns of the ANN with the behavior measured in *Controls* (*x*-axis) and ASD (*y*-axis). Four ANNs (with 5 models each generated from different layers of the ANNs are shown here in different colors). ANN predictions better match the behavior measured in the *Controls* compared with ASD. The correlation values (*x*-axis and *y*-axis) were corrected by the noise estimates per human population so that the differences are not because of differences in noise levels in measurements across the ASD and *Control* subject pools. The dot size refers to the degree of discrepancy between ANN predictivity of *Controls* versus ASD. ***D***, A comparison of the ANN predictivity (results from AlexNet shown here) of behavior measured in ASD versus *Controls* as a function of model layers [cnv layers 1, 3, 4, and 5 and the fully connected layer 7 (fc7), which approximately corresponds to the ventral stream cortical hierarchy]. The difference between the predictivity of behavior of the ANN in ASD and *Controls* increases with depth and is referred to as Δ. ***E***, Discriminability index (d′; ability to discriminate between image-level behavioral patterns measured in ASD vs *Controls*; see above, Materials and Methods) as a function of model layers (all 4 tested models shown separately). The difference in ANN predictivity between *Controls* and ASD was largest at the deeper (more IT-like) layers of the models instead of earlier (more V1, V2, and V4-like) layers. Error bars denote bootstrap confidence intervals.

## Results

As outlined above, I reasoned that the ability to predict the image-level differences in facial emotion judgments between individuals with ASD and neurotypical adults (*Controls*) allow us to (1) design more efficient experiments to study the atypical facial processing observed in ASD and (2) efficiently probe the underlying neural correlates. In this study, I first took a data-driven approach to discover such image-level differences in behavior across *Controls* and ASD in a facial emotion discrimination task (Wang and Adolphs, 2017). I then used brain-mapped computational models of primate vision to probe the underlying neural mechanisms that could drive such differences.

The behavioral and neural measurements analyzed in this study were performed by Wang et al. (2017) and Wang and Adolphs (2017). During the task, participants were shown images of individual faces with specific levels of morphed emotions (for 1 s) and asked to discriminate between two emotions, fear and happiness (Fig. 1*A*; see above, Materials and Methods). The authors observed a reduced specificity in facial emotion judgment among individuals with ASD compared with neurotypical *Controls* (Fig. 1*B*). Notably, the study controlled for low-level image confounds, and eye movement patterns across the two groups did not explain the reported behavioral differences. Therefore, the behavioral results significantly narrowed the space of neural hypotheses to sensory and affect-processing circuits.

### Image-level differences can be leveraged to produce stronger behavioral markers of atypical facial emotion judgments in autism

Wang and Adolphs (2017) primarily investigated the differences in behavior of ASD and *Controls* across parametric variations of facial emotion levels (e.g., levels of happiness and fear). Here, I first examined whether the image-by-image behavioral patterns (regardless of their facial identity or emotion levels) across the ASD and *Control* groups could be reliably estimated. Therefore, I computed the individual subject-to-subject correlations in image-level behavior (Fig. 1*D*), which show that both of the groups exhibit highly reliable image-level behavior. The internal reliability (see above, Materials and Methods) for *Control* and ASD groups are 0.73 and 0.70, respectively. A visual inspection of the comparison of behavioral patterns across the two groups (Fig. 1*C*) shows there are pairs of images (two such examples are

shown in Fig. 1*C*) for which the *Control* group exhibited very similar behavior, but the ASD made very different behavioral responses. This further confirms that diagnostic image-level variations in behavior could be further used to gain more insight into the mechanisms that drive the atypical facial emotion responses in ASD. Next, I quantified how stimuli selection based on high image-level differences can be leveraged to design more efficient behavioral experiments. To do this, I selected images based on the difference in behavior between the two groups ($\Delta^{Behav}$, using data from four randomly selected individual subjects from each group) and tested the resulting correlation between the behavior of the two groups (using the held-out subject population). This was repeated several times to get a mean measure of the cross-validated raw correlation (Fig. 1*E*, *y*-axis). Subjects from either group can exhibit different levels of behavioral variability in their responses for each group of selected images. It is critical to account for these varying noise levels while comparing across image groups. Therefore, noise ceilings were estimated for each image set selection based on image-level internal reliability of the held-out test population (see above, Materials and Methods). The difference between the noise ceiling and the raw correlation is referred to as the diagnostic efficiency $\eta$ of the image set, which is a measure of how efficient the image set is in discriminating between the ASD and *Control* behavior. The subject-level cross-validation was performed to estimate how $\eta$ might vary based on the specific cohorts of subjects chosen while determining the specific images. Figure 1*F* shows how $\eta$ varies across a more and more efficient selection of image sets (based on higher differences in image-level behavior with *Controls* and ASD). These results suggest that one reasonable goal of the field should be to find more efficient ways to predict which images will produce the highest $\eta$ values. Focusing human behavioral testing on such images is likely going to yield stronger inferences and lead to a better understanding of the behavioral and neural markers driving the difference in behavior.

### ANN models of primate vision trained on varied objectives can perform facial emotion judgment tasks

To investigate how one can predict the image-level facial emotion judgments, I first tested how accurately current ANN models of primate vision can be trained to perform such tasks. One advantage of using these ANNs is that there are significant correspondences between their architectural components and the areas in the primate ventral visual cortex (Yamins et al., 2014; Bashivan et al., 2019; Cadena et al., 2019; Fig. 2*A*, schematic). Also, there is a significant match in the predicted behavioral patterns of such models with primate behavior (including face-related tasks) measured during multiple object recognition tasks (Rajalingham et al., 2015, 2018). Together, these models are great candidates for generating testable hypotheses regarding both neural and behavioral markers of specific visual tasks. I selected four different ANNs to test their behavioral predictions with respect to the facial emotional judgment task. These ANNs were pretrained to perform image classification (AlexNet, Krizhevsky et al., 2012; CORnet-S, Kubilius et al., 2019), face recognition (VGGFace, Parkhi et al., 2015), and emotion recognition (Emotion Net, https://github.com/mathinking/FaceGenderAgeEmotionDetection). I observed that a 10-fold cross-validated partial least squares regression model (see above, Materials and Methods) could be used to train each model to perform the task. The variation of the behavioral responses of the model with parametric changes in the level of happiness in the faces qualitatively matched the patterns observed in the human data (Fig. 2*B*).

### ANN model predictions better match the behavioral patterns measured in neurotypical adults compared with individuals with autism

Next, I quantified how well the ANNs can predict the human image-level behavioral responses (across both *Controls* and ASD). Interestingly, ANN models significantly better predicted the image-level behavior measured in *Control* compared with the behavior measured in ASD (Fig. 2*C*; 20 models tested; paired *t* test, $p < 00{,}001$, $t_{(19)} = 10.99$). To account for the difference in variability in the behavior across ASD and *Control* groups, the correlation between model predictions and human data (Fig. 2*C*; referred to as Consistency) was normalized by the trial split-half reliability (i.e., noise ceiling estimates) of each group independently. To dissect which layer of the ANN best discriminated between the behavior of *Controls* and ASD, I compared individual models constructed from different layers of the same pretrained ANN architectures. This revealed two critical points. First, the correlation between model behavior and the *Control* group behavior increased as a function of model depth (Fig. 2*D*, black line, AlexNet), which corresponds to the ventral visual hierarchy as reported in many studies (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). Second, the difference in the predictivity of the model of behavior measured in *Controls* versus ASD across layers is also highest at deeper layers, which corresponds to primate IT (Fig. 2*D*, comparison of the black and the red line for AlexNet). This overall qualitative observation was consistent across all four tested models (Fig. 2*E*). Given the high discriminability index (see above, Materials and Methods), established mappings between the layers and primate brain, as well as wide usage among researchers, I have used AlexNet for the subsequent analysis presented in this study. Therefore, these results suggest that population neural activity in primate IT could play a significant role in the atypical facial emotion processing in people with autism, and the image-level differences in sensory representations in IT might explain the difference in behavior observed across the images. However, such a role has been previously attributed to the human amygdala responses (Wang et al., 2017). Therefore, I next tested whether the human amygdala responses can predict the image-level behavior and how well this predictivity could be explained by the ANN-IT representations.

### Two distinct neural population coding schemes in the human amygdala

Wang et al. (2017) recorded bilaterally from implanted depth electrodes in the human amygdala (Fig. 3*A*, schematic) from patients with pharmacologically intractable epilepsy. Subjects were presented each image for 1 s, same as the task description above (Wang and Adolphs, 2017), to discriminate between two emotions, fear and happiness. Similar to previous reports (Wang et al., 2017), I observed two distinct population of neurons in the human amygdala. These two populations were marked by significant response suppression, VS (57 neurons; Fig. 3*B*, right) and facilitation VF (99 neurons; Fig. 3*B*, left), respectively, after the onset of the facial image stimulus. I first tested how well the population-level activity (250–1500 ms postimage onset) of three specific subsamples of the amygdala neurons (VS only, VF only, and VS plus VS neurons) predicted the behavioral patterns measured in human subjects. I observed that each of these populations of VF, VS, and mixed (equal number of VS and VF neurons) could significantly ($p < 0.0001$; permutation test for significance of correlation) predict the image-level facial emotion judgments measured in *Controls*. Figure 3*C* shows how these three populations predict the image-level behavior measured in

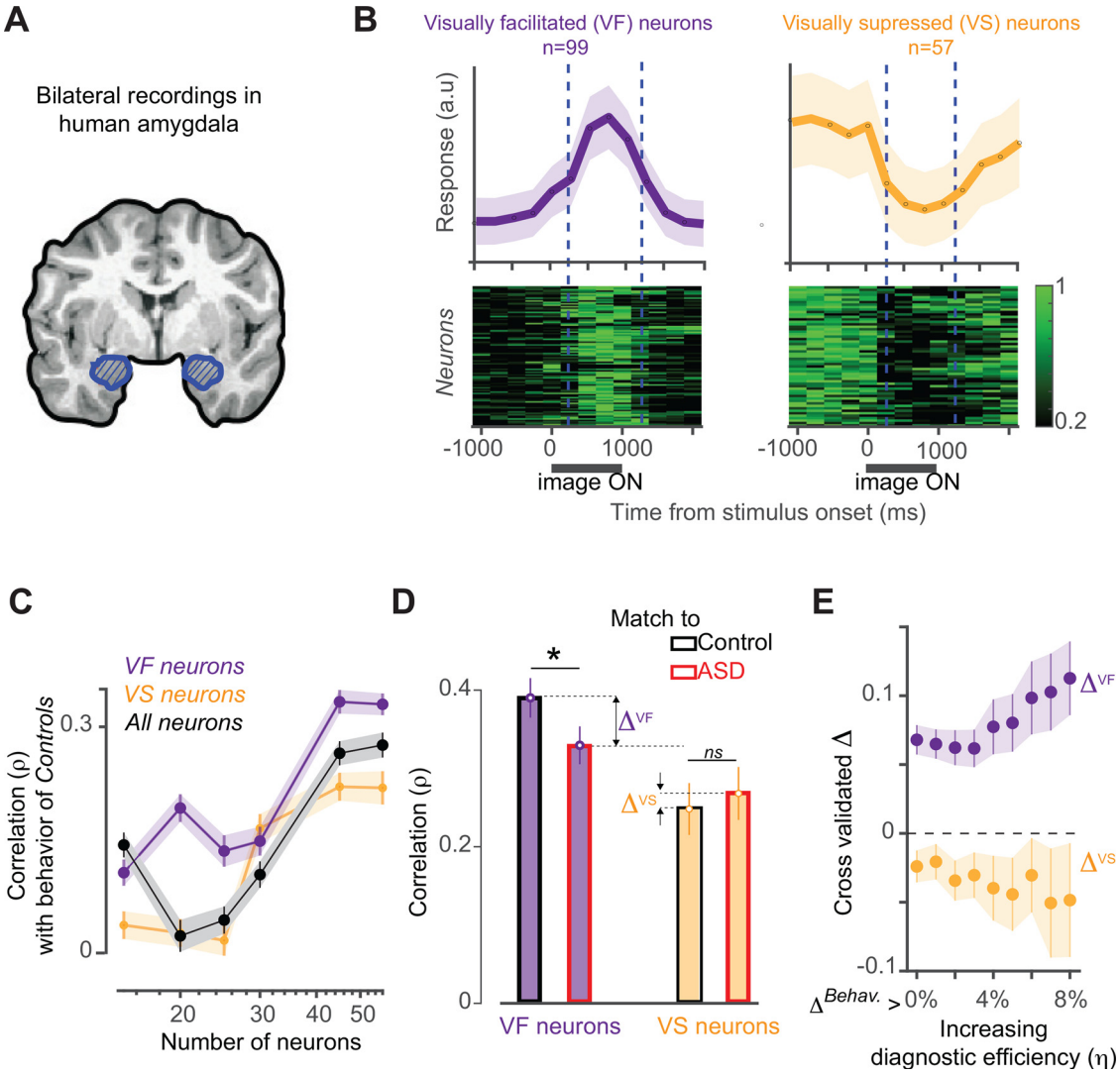


**Figure 3.** Facial emotion representation in the population neural activity of human amygdala. ***A***, Schematic of bilateral amygdala (blue patch) recordings performed by Wang et al. (2017). ***B***, Two distinct populations of neurons observed in the human amygdala. Top left, The VF (purple) neurons ($n = 99$) increased their responses after the onset of the face stimuli (averaged normalized spike rate across time; 250 ms time bins). Bottom left, The normalized firing rate across time for each VF neuron. Top right, The VS (yellow) neurons ($n = 57$) decreased their responses after the onset of the face stimuli (averaged normalized spike rate across time; 250 ms time bins). Bottom right, The normalized firing rates across time for each VS neuron. Error bars denote SEM across neurons. ***C***, An estimate (correlation) of how three subsamples of neural populations, VS (yellow), VF (purple), and VS plus VF (All, black) predict the image-level behavior measured in *Controls* as a function of the number of neurons sampled to build the neural decoders. Error bars denote bootstrapped CI. ***D***, Comparison of how well the VS (yellow bars) and VF (purple bars) neurons predict the behavior measured in *Controls* versus ASD. The red and black edges denote the predictivity of ASD and *Controls*, respectively. $\Delta^{VF}$ and $\Delta^{VS}$ are the differences in the human amygdala (neural decode) predictivity of facial emotion judgments measured in *Controls* and ASD from the VF and VS neurons, respectively. Error bars denote bootstrap CI. ***E***, $\Delta^{VF}$ and $\Delta^{VS}$ as function of image selection (which is proportional to the diagnostic efficiency $\eta$ estimated per image set). Cross-validation was done at the level of subjects for each image selection. Error bars denote bootstrap CI. *denotes a statistically significant difference ($p < 0.05$) and ns denotes no statistically significant difference.

*Controls* as a function of the number of neurons sampled to build the neural population decoders. Given that all these groups exhibit an increase in behavioral predictivity with the number of neurons, it is difficult to reject any of these decoding models (with the current neural dataset). Therefore, in the following analyses I have examined the VF and VS units separately. Next, I estimated how well the VS and VF population predicted the behavioral patterns measured in the *Controls* and ASD, respectively. Interestingly, I observed that similar to the ANN-IT behavior, neural decodes of the VF neurons in the human amygdala better match the *Control* group behavior compared with the ones measured in ASD (Fig. 3*C*; $\Delta^{VF}$ is significantly $>0$; permutation test of correlation, $p < 0.05$). However, the VS neurons did not show this trend (Fig. 3*D*; $\Delta^{VS}$ is not significantly different from 0; permutation test of correlation, $p > 0.05$). Figure 3*E* shows how VF (and not VS) neurons become more discriminatory of the ASD versus *Control* behavior (i.e., $\Delta^{VF}$ increases) as we choose image sets with higher diagnostic efficiencies ($\eta$). Consistent with prior work, these results provide evidence that neural responses in the human amygdala are implicated in atypical facial processing in people with autism. However, the results presented here also critically identify the VF neurons as stronger candidate neural marker of the differences in facial emotion processing observed in ASD.

### ANN-IT features can explain a significant fraction of the image-level behavioral predictivity of the human amygdala population

Given the significant predictivity of facial emotion judgments observed in the ANN-IT layers and the presence of strong

anatomic connections between primate IT and amygdala (Webster et al., 1991), I further asked how much of the image-level predictivity estimated from the amygdala activity is likely driven by input projections from the IT cortex. To test this, I first asked (with a linear regression analysis; see above, Materials and Methods) how well the image-by-image behavioral predictions from the ANN-IT models (AlexNet-fc7 tested here) can explain the image-by-image neural decoding patterns estimated from the amygdala neurons (separately for VS and VF neurons). The residue of this analyses (see above, Materials and Methods) contained the variance in the amygdala decodes that was not explained by the predictions of the ANN-IT models. Therefore, the amount of variance in the measured behavioral patterns explained by this residue provides an estimate of how much of the behavior is purely driven by the amygdala responses independent of the image-driven sensory re presentations. Assuming a feedforward hierarchical circuit whereby the IT cortex drives the human amygdala and not the other way around, a lower %EV obtained after such an analysis should indicate that the source of the signal in amygdala is at least partially coming from the IT cortex. Interestingly, this analysis revealed that the behavioral predictivity (%EV) of the human amygdala is significantly reduced once I regressed out the variance that is driven by the ANN-IT responses. For instance, when considering all images (i.e., very low diagnostic efficiency of the image set), I observed that VS and VF neurons could explain ~17.24 and 17.39% (a lower bound of the %EV as neural noise has not been accounted for) of the behavioral variance (Fig. 4*A*,*B*, left). However, once the ANN-IT-driven variance was regressed out, these values significantly dropped to 0.06 and 0.2%, respectively (Fig. 4*A*,*B*, right). Overall, VF neural residuals (after regressing out ANN-IT predictions) explained significantly less variance at all tested $\eta$ levels. VS neural residuals explained significantly less variance only at lower $\eta$ levels ($\Delta^{Behav} < 2.5\%$). Given that VS neurons showed a drop in %EV for higher $\eta$ levels, it is not surprising that I did not observe any differences with the residual predictivity at those levels. Interestingly, there was no significant change in %EV across the image selections when VS activity was regressed out of VF activity (and vice versa; Fig. 4*A*,*B*, middle), providing further evidence that they largely support a complementary coding scheme for facial emotions within the amygdala. In sum, these results suggest that input projections from the IT cortex into the amygdala (Webster et al., 1991) might be the primary career of the facial-emotion-related signals. Furthermore, the results also suggest a likely difference in how VS and VF neurons are affected in ASD, with VF neurons being more diagnostic of the atypical behavior observed in ASD.

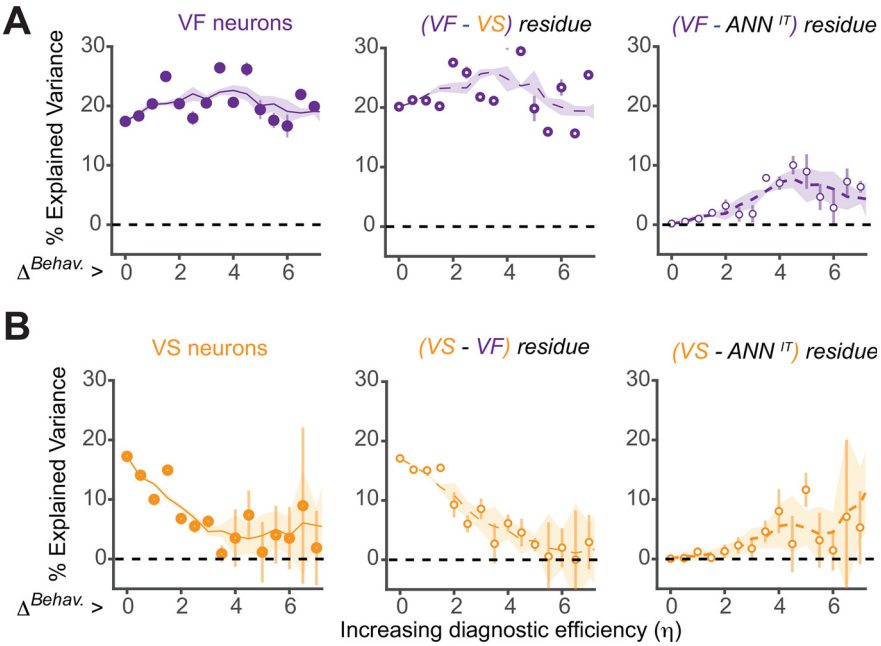


**Figure 4.** Amount of behavioral variance (measured in *Controls*) explained by different neural markers. ***A***, Left, Percentage of behavioral variance explained by the human amygdala (VF) neural activity as a function of the overall differences in image-level behavior between ASD and *Controls*. As demonstrated in Figure 1*F*, the *x*-axis is proportional to the diagnostic efficiency ($\eta$). Middle, Percentage of variance explained by the residual (VS-based predictions regressed out of the predictions from VF-based neural decodes). There was no significant change in %EV across the image selections when VS was regressed out, suggesting a complementary coding scheme. Right, Percentage of behavioral variance explained by the residual (ANN-IT predictions regressed out of the predictions from VF-based neural decodes). There was a significant difference (reduction in %EV) between the two cases for all levels of tested $\eta$. ***B***, Left, Percentage of behavioral variance explained by the human amygdala (VS) neural activity as a function of the overall differences in image-level behavior between ASD and *Controls*. Middle, Percentage of variance explained by the residual (VF-based predictions regressed out of the predictions from VS-based neural decodes). There was no significant change in %EV across the image selections when VF was regressed out, suggesting a complementary coding scheme. Right, Percentage of variance explained by the residual (ANN-IT predictions regressed out of the predictions from VS-based neural decodes). There was a significant difference (reduction in %EV) between the two cases, whereas $\Delta^{Behav}$ was <2. All %EV values were estimated in a cross-validated way, wherein the image selections and the final estimates were done based on different groups of subjects. Error bars denote bootstrapped CI.

### *In silico* perturbations with additional noise in ANN-IT layers improves the match of the model with the behavior of individuals with autism

To further probe how IT representations might be different in ASD compared with *Controls* (Fig. 5*A*), I compared ANNs independently trained to predict the behavior of *Controls* and ASD. I directly compared the learned weights, that is, the synaptic strengths between the model-IT layer and the behavioral output node in the two cases. I observed that models trained on the behavior measured in ASD yielded weaker synaptic strengths for both excitatory (positively weighted) and inhibitory (negatively weighted) connections (Fig. 5*B*), compared with models trained to reproduce the behavior measured in *Controls*. I further explored how this modest difference in the models could be simulated so that an ANN trained on ground truth labels of human facial emotions could be transformed into behaving more like what we observe in ASD. Based on previous studies (MacDonald et al., 2006; Haigh et al., 2015), I hypothesized that increased noise (scaled according to overall responsiveness of the model units) in the sensory representations during learning could potentially yield weaker synaptic strengths between the model-IT layer and the trained behavioral output node. Of note, although a noisy representation likely yields a reduced specificity in behavioral performance, an addition of specific amounts of noise does not necessarily guarantee a stronger or weaker

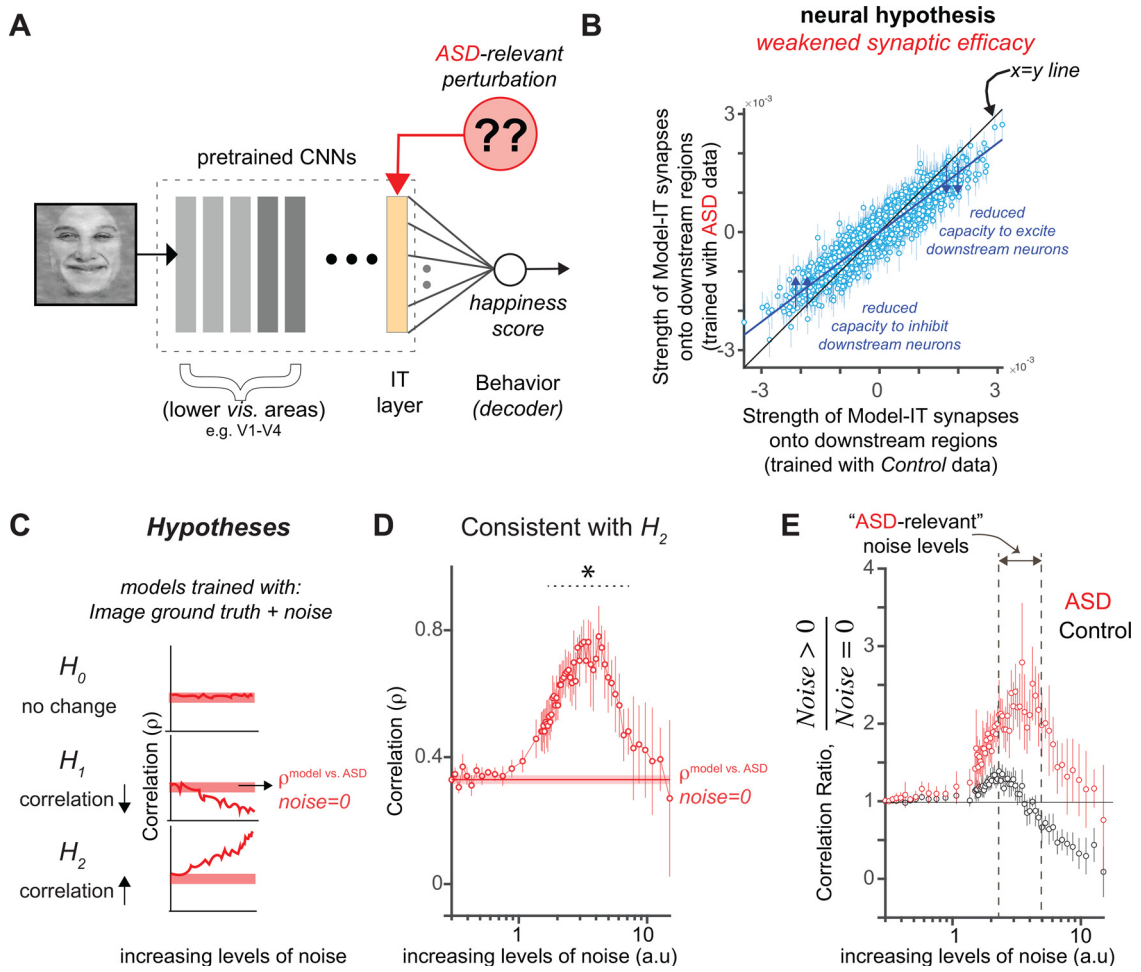


**Figure 5.** *In silico* experiments on ANNs to probe neural mechanisms underlying atypical facial emotion judgments in individuals with ASD. ***A***, What changes can one induce in the model-IT layer to simulate the behavioral patterns measured in ASD? ***B***, Comparison of synaptic strengths (weights) between ANN-IT and the behavioral node when models are independently trained with the behavior measured in ASD versus *Controls*. ANN fits to the behavior of ASD yielded weaker synaptic strengths for both excitatory (positively weighted) and inhibitory (negatively weighted) connections. Each blue dot refers to the weights in the connection between an individual model unit in the IT layer and the decision (level of happiness) node. ***C***, Hypotheses and corresponding predictions; $H_0$, addition of noise could lead to no differences in how it affects the match of the model to behavior measured in ASD; $H_1$, addition of noise could reduce the match of the models to behavior measured in ASD compared with the noise-free model, and $H_2$, addition of noise could improve the match of the models to the behavior measured in ASD compared with the noise-free model.$H_2$ supports the high IT variability in autism hypotheses. ***D***, Correlation of ANN behavior with ASD as a function of levels of added noise. The results show that at specific noise regimes ANNs are significantly more predictive of the behavior measured in ASD compared with the noiseless model. Error bars denote bootstrapped CI. ***E***, Ratio of ANN behavioral predictivity of noisy versus noise-free ANNs. At specific levels of noise, referred to as the ASD-relevant noise levels, the ANNs trained with noise show much higher predictivity for behavior measured in ASD while suffering a reduction in predictivity of the *Controls*. Error bars denote bootstrapped CI. Facial images shown in this figure are morphed and processed versions of the original face images. These images have full reuse permission. *denotes a statistically significant difference ($p < 0.05$).

correlation with the image-level behavioral patterns observed in ASD. Therefore, such *in silico* perturbations could produce three primary outcomes. First, adding noise might produce no effects in the behavioral match of the model with the behavior of ASD (Fig. 5*C*, top, $H_0$). Second, the added noise might weaken the correlation achieved by a noiseless model (Fig. 5*C*, middle, $H_1$). Third, and consistent with an ASD-relevant mechanism, addition of noise could improve the correlation with the image-level behavior measured in ASD (Fig. 5*C*, bottom, $H_2$). I observed that at specific levels of added noise (Fig. 5*D*; dashed black line) during the model training (transfer learning), the behavioral match of the model with ASD significantly improved (assessed by permutation test of correlation) beyond the levels noted with a noise-free model (Fig. 5*D*). In addition, this increase in the predictivity of ASD behavior with the addition of noise is significantly higher than that observed when compared with the predictivity of the model of the behavior measured in the *Controls* (Fig. 5*E*). Within the dashed black lines (Fig. 5*E*), noise added to each model unit was drawn from a normal distribution with a zero mean and SD equal to two to five times the width of the response distribution of that unit across all tested images. Together, this strongly suggests that additional noise in sensory representations is a very likely candidate mechanism implicated in atypical facial emotion processing in adults with autism.

## Discussion

The overall goal of this study was to identify candidate neural and behavioral markers of atypical facial emotion judgments observed in individuals with autism. Based on discovering reliable image-by-image differences between the behavior of *Controls* and ASD that could not be explained by categorical ambiguity in the stimuli, I reasoned that such image-level variance could be leveraged to probe the neural mechanisms of behavioral differences observed in ASD. Therefore, I used

image-computable, brain-tissue-mapped artificial neural network models of primate vision to further probe the issue. By using computational models (that have established brain tissue correlates) to explain experimental data, I hereby demonstrate how such an approach could be used to probe the neural mechanisms that underlie the differences in facial emotion processing observed in individuals with autism. Below, I discuss the findings with their relevance to future experiments and candidate mechanisms implicated in atypical facial emotion recognition in ASD.

### ANN-based predictions can be used to efficiently screen images and provide neural hypotheses for more powerful experiments

A family of ANN models can currently predict a significant amount of variance measured in various object recognition related behaviors and neural circuits (Schrimpf et al., 2018). Given that the results presented here demonstrate the ability of such ANNs to discriminate between the behavior measured in *Controls* and ASD, we can further leverage the ANNs to screen facial image stimuli and select images where the predicted behavioral differences are maximum. Further, such models can be reverse engineered (Bashivan et al., 2019; Xiao and Kreiman, 2020) to synthesize images that could achieve maximum differences to optimize behavioral testing and diagnosis. Such deep image synthesis methods could also modify the facial images such that the differences in the observed behavior between the *Controls* and ASD are minimized. Although clearly at an early stage, such methods have significant potential to improve future cognitive therapies. Unlike many machine-learning approaches that are not closely tied to the computation and architecture of the primate brain, the ANNs used in this study have established homologies with the primate brain and behavior (Schrimpf et al., 2018). As demonstrated in this study, these links allow us to relate the ANN predictions to distinct brain areas directly. Specifically, the ANN results presented here suggest that population activity patterns in areas like the human and macaque inferior temporal cortex are vital candidates for neural markers of atypical facial processing in autism. The modeling results provide further insights into the most affected aspects of the population responses, implicating noisier sensory representations (see below) as a source of the differences in sensory representation, learning, and subsequent decision-making. In addition to the specific hypotheses generated in this study, it is essential to note that ANN models of primate vision is an active area of research, and we are witnessing the gradual emergence of better brain-matched models (Nayebi et al., 2018; Kubilius et al., 2019; Lee et al., 2020; Zhuang et al., 2021). Therefore, this study establishes a critical link between atypical face processing in autism and how to leverage ANNs to study this.

### Modeling results imply the need for more fine-grain neural measurements in the primate IT cortex and amygdala

The ANN-based computational analyses in this study provide specific neural hypotheses that can be tested using macaque electrophysiology and human fMRI experiments. First, I observed that the ANN-IT layers could best discriminate between the behavior of *Controls* versus ASD. Therefore, such signals are likely also measurable in the primate IT cortex and are key candidates for neural markers of atypical facial emotion processing in autism. Given that most ANN models are feedforward only or have minimal dynamics, it will be critical to test how the different temporal components of IT population responses carry the facial emotion signal. Similar to predictions of ANN-IT layers, I observed that population activity in the human amygdala also better matches behavior measured in the *Controls* than ASD. There can be multiple reasons for the observed differences in behavioral predictivity. First, it is possible that because of the atypical development of the human amygdala in ASDs, the behavior they exhibit does not match well with the neural decodes out of the neurotypical amygdala. Second, the lack of predictivity might be carried forward from responses in the IT cortex, as predicted by the ANNs. The current study attempted to disambiguate these two factors. I asked how well ANN-IT predictions can account for the behavioral patterns of the amygdala activity. Indeed, the image-level predictivity of facial emotion judgments observed in the population activity of the human amygdala (both VF and VS neurons) was significantly explained away by the ANN-IT features (Fig. 4*A*,*B*, left). This result is consistent with the hypothesis that the higher level visual cortices (like IT) primarily drive the facial affect signal observed in the human amygdala. Simultaneous neural recordings in IT and amygdala or finer grain causal perturbation experiments need to be conducted to test this hypothesis more directly. Notably, the behavioral mismatch (neural decodes vs *Control*/ASD behavior) was specific to the decodes constructed from the VF neurons (and not VS neurons). Therefore, future experimental investigations should dissect the role of IT cortex and how it functionally influences the VF and VS neurons, which are likely part of a complementary coding scheme. Furthermore, it will be essential to examine how the IT cortical activity is driven by feedback projections from the amygdala, given that evidence for the importance of such connections from ventrolateral PFC has been demonstrated for object recognition (Kar and DiCarlo, 2021).

### High variability in sensory representation can lead to weaker efferent synaptic strengths during learning and development

In a psychophysical discrimination task, the typical consequence of having a noisy detector is a reduction in the sensitivity of performance, which manifests as a reduced estimated slope of the psychometric function. This is consistent with what Wang and Adolphs (2017) had observed. Given that the idea of higher sensory variability in autism is also consistent with previous findings (Haigh et al., 2015), I considered this as a potential neural mechanism that could explain the image-level differences I have observed in the facial emotion discrimination behavior in ASD. Therefore, I tested the increased sensory noise hypothesis to test whether such a perturbation could simulate the weaker efferent synaptic connections from IT-like layers as revealed by the ANN-based analyses (Fig. 5*B*). Indeed, addition of noise during learning made the ANN behavior more matched with that observed in ASD. First, this could suggest that perhaps the behavior measured in ASD results from additional noise in the sensory representations that affects the subjects' behavior during the task. However, this could also be the result of executing an inference engine (in the brain) that learned its representations under high sensory noise during development (as a child). An estimate of noise levels (sensory cortical signal variability) in children with autism and a quantitative probe into how that could potentially interact with learning new tasks is essential to test this hypothesis. As demonstrated in this study, the ANN models provide a very efficient framework to generate more diagnostic image sets for these future studies given that we can simulate any level (and type) of noise under different learning regimes and make predictions on effect sizes. Such model-driven

hypotheses are likely to play a vital role in guiding future experimental efforts and inferences.

### High variability in sensory representation can qualitatively explain other ASD-specific behavioral reports

Addition of noise during the transfer learning procedure of the ANN models made the behavioral output of the model more consistent with the behavior measured in ASD (Fig. 5*D*). Such a mechanism can indeed qualitatively explain other previous behavioral observations made in individuals with autism. For example, Behrmann et al. (2006b) observed that reaction times measured during object discrimination tasks in adults with autism were significantly higher than the those of *Control* subjects. This difference was especially high during more fine-grained discrimination tasks. Such a behavioral phenomenon can be explained by an increase in sensory noise in ASD that leads to longer time requirements during integration of information (Ratcliff et al., 2016) and weaker performances on finer discrimination tasks. The ANN-based approach demonstrated in this study, however, provides guidance beyond the qualitative predictions of overall effect types. Specific image-level predictions provided by ANNs will help researchers to design more diagnostic behavioral experiments and make measurements that can efficiently discriminate among competing models of brain mechanisms.

### Potential underlying mechanisms for atypical IT responses

In a psychophysical discrimination task, the typical consequence of having a noisy detector is a reduction in the sensitivity of performance, which manifests as a reduced estimated slope of the psychometric function. This is consistent with what Wang and Adolphs (2017) had observed. Given that the idea of higher sensory variability in autism is also consistent with previous findings (Haigh et al., 2015), I speculate that this as a potential neural mechanism that could explain the image-level differences I have observed in the facial emotion discrimination behavior in ASD.

Addition of noise during the learning procedure of the ANN models could be used to simulate this hypothesis in the models. Such a mechanism could indeed also qualitatively explain other previous behavioral observations made in individuals with autism. For example, Behrmann et al. (2006b) observed that reaction times measured during object discrimination tasks in adults with autism were significantly higher than those of the *Control* subjects. This difference was especially high during more fine-grained discrimination tasks. Such a behavioral phenomenon could be putatively explained by an increase in sensory noise in ASD that leads to longer time requirements during integration of information (Rubenstein and Merzenich, 2003) and weaker performances on finer discrimination tasks. The ANN-based approach demonstrated in this study, however, provides guidance beyond the qualitative predictions of overall effect types. Specific image-level predictions provided by ANNs will help researchers to design more diagnostic behavioral experiments and make measurements that can efficiently discriminate among competing models of brain mechanisms. An imbalance in the ratio of the excitatory and inhibitory (E/I) processes in cortical circuits has been proposed as an underlying mechanism for various atypical behaviors observed in autism (Rubenstein and Merzenich, 2003). I speculate that such an E/I imbalance could arise because of lower inhibition in the cortical networks. This could lead to larger neural variability and a subsequent noisier, less efficient sensory processing. Therefore, the proposed *in silico* experiments might lead to biologically plausible mechanistic hypotheses. In fact, genetic mutations that have an impact on the generation and function of interneurons have been previously linked with autism (Chao et al., 2010; Sohal and Rubenstein, 2019). Therefore, cell-type-specific causal perturbation approaches are necessary to test whether a decreased inhibition in the visuocortical pathway (especially in the primate IT cortex) leads to noisier sensory representations and can reproduce the specific image-level differences in facial emotion processing reported in this study. The image-level behavioral measurements and ANN predictions reported here will enable such stronger forms of hypothesis testing during the interpretation of such experimental results.

### Limitations of the current study

The current study explores the mechanism of facial emotion recognition differences across a limited set and a specific cohort of ASD and control subjects. Furthermore, the stimuli set is also limited to 28 face images. Given the heterogeneity observed in ASD, the inferences drawn from this study should motivate others to design large-scale studies of the same nature, where further predictions from the ANN models can be tested across a more diverse subject pool and a broader set of images (and possibly videos).

## References


Adolphs R (2008) Fear, faces, and the human amygdala. Curr Opin Neurobiol 18:166–172.

Adolphs R, Tranel D, Damasio H, Damasio A (1994) Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 372:669–672.

Adolphs R, Tranel D, Hamann S, Young AW, Calder AJ, Phelps EA, Anderson A, Lee GP, Damasio AR (1999) Recognition of facial emotion in nine individuals with bilateral amygdala damage. Neuropsychologia 37:1111–1117.

Adolphs R, Sears L, Piven J (2001) Abnormal processing of social information from faces in autism. J Cogn Neurosci 13:232–240.

Bashivan P, Kar K, DiCarlo JJ (2019) Neural population control via deep image synthesis. Science 364:eaav9436.

Behrmann M, Avidan G, Leonard GL, Kimchi R, Luna B, Humphreys K, Minshew N (2006a) Configural processing in autism and its relationship to face processing. Neuropsychologia 44:110–129.

Behrmann M, Thomas C, Humphreys K (2006b) Seeing it differently: visual processing in autism. Trends Cogn Sci 10:258–264.

Broks P, Young AW, Maratos EJ, Coffey PJ, Calder AJ, Isaac CL, Mayes AR, Hodges JR, Montaldi D, Cezayirli E, Roberts N, Hadley D (1998) Face processing impairments after encephalitis: amygdala damage and recognition of fear. Neuropsychologia 36:59–70.

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput Biol 15:e1006897.

Chao HT, Chen H, Samaco RC, Xue M, Chahrour M, Yoo J, Neul JL, Gong S, Lu HC, Heintz N, Ekker M, Rubenstein JL, Noebels JL, Rosenmund C, Zoghbi HY (2010) Dysfunction in GABA signalling mediates autism-like stereotypies and Rett syndrome phenotypes. Nature 468:263–269.

Ekman P, Keltner D (1997) Universal facial expressions of emotion. In: Nonverbal communication: where nature meets culture (Segerstra³le UCO, Molnár P, eds), pp 27–46. Mahwah, NJ: Erlbaum.

Freiwald WA, Tsao DY, Livingstone MS (2009) A face feature space in the macaque temporal lobe. Nat Neurosci 12:1187–1196.

Golarai G, Grill-Spector K, Reiss AL (2006) Autism and the development of face processing. Clin Neurosci Res 6:145–160.

Haigh SM, Heeger DJ, Dinstein I, Minshew N, Behrmann M (2015) Cortical variability in the sensory-evoked response in autism. J Autism Dev Disord 45:1176–1190.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

Kar K, DiCarlo JJ (2021) Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. Neuron 109:164–176.

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci 22:974–983.

Kennedy DP, Adolphs R (2012) Perception of emotions from facial expressions in high-functioning adults with autism. Neuropsychologia 50:3313–3319.

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915.

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Communications of the ACM 60:84–90.

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj N, Issa E, Bashivan P, Prescott-Roy J, Schmidt K (2019) Brain-like object recognition with high-performing shallow recurrent anns. arXiv. Advance online publication. Retrieved May 23, 2022. Available at https://arxiv.org/abs/1909.06161.

Lee H, Margalit E, Jozwik KM, Cohen MA, Kanwisher N, Yamins DL, DiCarlo JJ (2020) Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. bioRxiv. Advance online publication. Retrieved May 23, 2022. Available at https://doi.org/10.1101/2020.07.09.185116.

Lozier LM, Vanmeter JW, Marsh AA (2014) Impairments in facial affect recognition associated with autism spectrum disorders: a meta-analysis. Dev Psychopathol 26:933–945.

MacDonald SW, Nyberg L, Bäckman L (2006) Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. Trends Neurosci 29:474–480.

Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, DiCarlo JJ, Yamins DL (2018) Task-driven convolutional recurrent models of the visual system. arXiv. Advance online publication. Retrieved May 23, 2022. Available at https://arxiv.org/abs/1807.00053.

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. Paper presented at British Machine Vision Conference, Swansea, UK, September.

Rajalingham R, Schmidt K, DiCarlo JJ (2015) Comparison of object recognition behavior in human and monkey. J Neurosci 35:12127–12136.

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J Neurosci 38:7255–7269.

Ratcliff R, Smith PL, Brown SD, McKoon G (2016) Diffusion decision model: current issues and history. Trends Cogn Sci 20:260–281.

Robertson CE, Baron-Cohen S (2017) Sensory perception in autism. Nat Rev Neurosci 18:671–684.

Roy S, Roy C, Fortin I, Ethier-Majcher C, Belin P, Gosselin F (2007) A dynamic facial expression database. Journal of Vision 7:944.

Rubenstein JL, Merzenich MM (2003) Model of autism: increased ratio of excitation/inhibition in key neural systems. Genes Brain Behav 2:255–267.

Rutishauser U, Mamelak AN, Adolphs R (2015) The primate amygdala in social perception—insights from electrophysiological recordings and stimulation. Trends Neurosci 38:295–306.

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Schmidt K (2018) Brain-score: which artificial neural network for object recognition is most brain-like? BioRxiv. Advance online publication. Retrieved May 23, 2022. Available at https://www.biorxiv.org/content/10.1101/407007v1.

Sohal VS, Rubenstein JLR (2019) Excitation-inhibition balance as a framework for investigating mechanisms in neuropsychiatric disorders. Mol Psychiatry 24:1248–1257.

Tsao DY, Livingstone MS (2008) Mechanisms of face perception. Annu Rev Neurosci 31:411–437.

Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RB (2003) Faces and objects in macaque cerebral cortex. Nat Neurosci 6:989–995.

Tsao DY, Moeller S, Freiwald WA (2008) Comparing face patch systems in macaques and humans. Proc Natl Acad Sci U S A 105:19514–19519.

Uljarevic M, Hamilton A (2013) Recognition of emotions in autism: a formal meta-analysis. J Autism Dev Disord 43:1517–1526.

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

Wang S, Yu R, Tyszka JM, Zhen S, Kovach C, Sun S, Huang Y, Hurlemann R, Ross IB, Chung JM, Mamelak AN, Adolphs R, Rutishauser U (2017) The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. Nat Commun 8:14821.

Webster MJ, Ungerleider LG, Bachevalier J (1991) Connections of inferior temporal areas TE and TEO with medial temporal-lobe structures in infant and adult monkeys. J Neurosci 11:1095–1116.

Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE toolbox. Behav Res Methods 42:671–684.

Xiao W, Kreiman G (2020) XDream: finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. PLoS Comput Biol 16:e1007973.

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624.

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, Yamins DLK (2021) Unsupervised neural network models of the ventral visual stream. Proc Natl Acad Sci U S A 118:e2014196118.