CORTEX 103 (2018) 24–43

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without Autism

Dorit Kliemann [a,*], *Hilary Richardson* [b], *Stefano Anzellotti* [b], *Dima Ayyash* [b], *Amanda J. Haskins* [b], *John D.E. Gabrieli* [a] *and Rebecca R. Saxe* [b]

[a] *McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA*
[b] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA*

ARTICLE INFO

Article history:
Received 9 September 2017
Reviewed 29 October 2017
Revised 11 January 2018
Accepted 8 February 2018
Action editor Holger Wiese
Published online 21 February 2018

Keywords:
Social cognition
fMRI
Emotional faces
MVPA
Autism

ABSTRACT

Individuals with Autism Spectrum Disorders (ASD) report difficulties extracting meaningful information from dynamic and complex social cues, like facial expressions. The nature and mechanisms of these difficulties remain unclear. Here we tested whether that difficulty can be traced to the pattern of activity in "social brain" regions, when viewing dynamic facial expressions. In two studies, adult participants (male and female) watched brief videos of a range of positive and negative facial expressions, while undergoing functional magnetic resonance imaging (Study 1: ASD n = 16, control n = 21; Study 2: ASD n = 22, control n = 30). Patterns of hemodynamic activity differentiated among facial emotional expressions in left and right superior temporal sulcus, fusiform gyrus, and parts of medial prefrontal cortex. In both control participants and high-functioning individuals with ASD, we observed (i) similar responses to emotional valence that generalized across facial expressions and animated social events; (ii) similar flexibility of responses to emotional valence, when manipulating the task-relevance of perceived emotions; and (iii) similar responses to a range of emotions within valence. Altogether, the data indicate that there was little or no group difference in cortical responses to isolated dynamic emotional facial expressions, as measured with fMRI. Difficulties with real-world social communication and social interaction in ASD may instead reflect differences in initiating and maintaining contingent interactions, or in integrating social information over time or context.

* *Corresponding author*. McGovern Institute for Brain Research, Massachusetts Institute of Technology, 43 Vassar Street, 46-4021, Cambridge, MA 02139, USA.
E-mail address: dorit@mit.edu (D. Kliemann).
https://doi.org/10.1016/j.cortex.2018.02.006
0010-9452/

1. Introduction

Humans are generally highly sensitive to social cues from other people: we can infer a friend's disappointment in a turn of the head, a shift of the eyes, a purse of the lips. While such sensitivity is seemingly effortless for most people, individuals with Autism Spectrum Disorder struggle disproportionately with social communication and social interaction (Adolphs, 2006; American Psychological Association, 2013). Here, we tested the hypothesis that social impairments in ASD reflect a disruption of neural mechanisms for extracting relevant social information from brief, dynamic, naturalistic facial expressions.

Capturing the scope and sophistication of human social perception within a laboratory task is a serious challenge. Observers can easily distinguish posed exaggerated facial expressions of a few basic emotions (e.g., happy *vs* sad *vs* afraid), but in natural interactions, expressions are dynamic, mixed, and variable, even within valence (Zaki & Ochsner, 2009). To understand others' emotions, observers must be able to attend to and extract the relevant information from highly variable stimuli (Adolphs, 2006), and integrate facial expressions with additional information from the context (Aviezer, Bentin, Dudarev, & Hassin, 2011). In healthy individuals, multiple cortical regions appear implicated in these aspects of emotion understanding. Emotional facial expressions elicit distinct patterns of activity in the fusiform face area (FFA, e.g., Harry, Williams, Davis, & Kim, 2013), in parts of the superior temporal sulcus (STS, e.g., Said, Moore, Norman, Haxby, & Todorov, 2010), and in the medial prefrontal cortex (MPFC, e.g., Chavez & Heatherton, 2015). The FFA is sensitive to configurations of facial features (Liu, Harris, & Kanwisher, 2010), but parts of STS and MPFC also integrate emotional information from body postures, vocal tones (Peelen, Atkinson, & Vuilleumier, 2010) and the surrounding context (Skerry & Saxe, 2014). Responses in all of these regions are modulated by whether emotions are currently task-relevant for the observer (Kliemann, Jacoby, Anzellotti, & Saxe, 2016). Disruption in these functional networks could cause difficulties in social perception and social interaction.

Individuals with Autism Spectrum Disorders often (but not always) show impairments on recognition of emotional facial expressions (Harms, Martin, & Wallace, 2010; Uljarevic and Hamilton, 2013). These impairments appear to be especially marked when facial expressions are dynamic and naturalistic (Pelphrey, Morris, McCarthy, & Labar, 2007), and when combining cues from faces and from the surrounding context (Rosenblau, Kliemann, Heekeren, & Dziobek, 2015). In addition, individuals with ASD may show altered social attention, which could affect their ability to endogenously direct attention to social aspects of stimuli or situations (Wang et al., 2015).

The neural source of these impairments has been difficult to ascertain. Many previous studies have measured the magnitude of hemodynamic responses in the brain, while individuals with ASD viewed emotional or dynamic faces. Unfortunately, the results have been highly heterogeneous, with some studies finding hypo-activation, some finding hyper-activation, and some finding no group difference in the FFA, STS and MPFC (Alaerts et al., 2014; Boelte et al., 2015; Hadjikhani et al., 2014; Rahko et al., 2012; Scherf, Elbich, Minshew, & Behrmann, 2015; Schneider et al., 2013; Weisberg et al., 2014). Explicit instructions to attend to emotional faces also yield conflicting results, with some studies finding no group differences during explicit social tasks (Boelte et al., 2015; Kana, Patriquin, Black, Channell, & Wicker, 2016; Schneider et al., 2013), and other studies finding no difference in the same regions in the absence of a task (e.g., Pantelis, Byrge, Tyszka, Adolphs, & Kennedy, 2015). One possibility is that the average magnitude of response in FFA, STS, and MPFC is not a straightforward measure of social information processing. The magnitude of hemodynamic activity in a region likely reflects a mix of distinct cognitive and neural factors (e.g., effort, prediction error, domain specificity, sparse coding, etc.). Multivariate patterns of activity sometimes offer a more sensitive measure of the information represented in a brain region than univariate analyses (Haxby et al., 2001; Koster-Hale, Saxe, Dungan, & Young, 2013) and therefore could reveal differences between groups (Coutanche, Thompson-Schill, & Schultz, 2011).

Here, in two studies, we measured the extraction of emotion-relevant information from dynamic facial expressions using multivariate pattern analyses. In Study 1, participants viewed naturalistic movie clips of emotional expressions from 96 different individuals. We tested the degree to which cortical regions extracted valence, a fundamental emotionally-relevant dimension (Russel, 1980), from these stimuli; and whether the pattern of activity also generalized to a completely different emotional stimulus (animated events). In Study 2, participants viewed more exaggerated expressions of 10 emotions from 20 actors. Again, we tested whether cortical regions extracted the valence of the expressions (as well as the finer grained structure); and we additionally tested whether these responses were modulated by the participant's endogenous attention. To increase the power of our analyses, where possible we combined data from the two studies (for a total of 89 datasets from 80 participants). In sum, these experiments provide multiple metrics of the scope and flexibility of cortical representations of facial emotional expressions, and therefore should provide a sensitive test of the hypothesis that social impairments in individuals with ASD reflect difficulty extracting socially relevant features from dynamic events.

2. Materials and methods

2.1. Participants

In Study 1, we recruited 18 participants diagnosed with Autism Spectrum Disorder and 21 neurotypically-developed adults (NT) with otherwise no history of neurological or psychiatric disorders. The control participants' data is re-analyzed from Skerry and Saxe (2014). In Study 2, we recruited 24 participants diagnosed with Autism Spectrum Disorder and 32 neurotypically-developed adults with no other history of neurological disorders. Nine participants (1 NT, 8 ASD) participated in both studies. We excluded two ASD participants in Study 1 and two in Study 2 [n = 3 excessive in-scanner head motion, n = 1 low performance in the task (see

Table 1 – Descriptive values of ADOS scales across ASD participants (Study 1 and Study 2).

Scale	Mean	SD	Min	Max
Communication	3.5	1.39	2	6
Social	6.91	2.12	4	12
RRB	1.91	1.75	0	6
Comm + Soc	10.09	3.11	7	18

Abbreviations: Comm, Communication; min, minimum value; max, maximum value; RRB, restricted and repetetive behavior; SD, Standard Deviation; Soc, Social.

Section 3)]. Two NT participants scored above the cut-off value on the Autism Spectrum Quotient (>31, Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) and were excluded. The final data set included 21 NT and 16 ASD in Study 1 and 30 NT and 22 ASD in Study 2. Table 2 shows the demographics of this final sample. In Study 1, all participants were right-handed, whereas Study 2 included left-handed participants (NT: n = 4, 4 males; ASD: n = 4, 2 females, 2 males).

All but one participant with a previous clinical diagnosis of ASD were assessed using the Autism Diagnostic Observation Schedule (ADOS-2, see Table 1 for scales) by research-reliable administrators in order to confirm the diagnosis (Lord et al., 2012). For the second study, we additionally assessed current and prior history of psychiatric conditions, as well as medication status (see Supplementary materials for details).

All participants were verbally fluent with average to above-average intellectual ability, as measured by the Kauffman Brief Intelligence Test, Second Edition (KBIT-2) (Kaufman, 2004). All subjects had normal or corrected-to-normal vision, were paid for participation and gave written informed consent prior to participating, in accordance with the Committee on the Use of Human Experimental Subjects (COUHES) at the Massachusetts Institute of Technology (MIT).

2.2. Social attribution tasks

In both studies, participants made judgments about social attributes in the scanner (see Fig. 1).

2.2.1. Study 1: Emotion attribution task

Participants watched 384 short video clips. Half of the videos included faces expressing a positive (happy/smiling) or negative (sad/frowning) emotion (*expressions condition*), and the other half were brief animations in which a simple geometric character experienced an event that would elicit a positive or negative emotion (*situations condition*). The facial expressions were clips from movies, with the constraint that each clip showed a continuous close-up of one character, expressing an unambiguous emotional expression. Characters were either male or female and showed positive or negative expressions. The situations were animated interactions between characters shown as simple geometric shapes. The shapes had eyes but no other facial features, and did not change configuration in response to the events (i.e., they did not make emotional expressions). A single target character, distinguished by color, acted to achieve a physical goal (climb a hill, retrieve an object) or social goal (inclusion in a group); the character either succeeded or failed in this goal. For further details on the stimuli, please see Skerry and Saxe (2014).

After each video, participants simply indicated the intensity of the target's emotion independent of valence, by pressing 1 of 4 buttons (1–4, neutral to extreme). Participants were not required to classify the emotion.

The experiment consisted of 8 runs (9.43 min/run). Each run contained 12 stimuli in each of the 4 conditions [(positive or negative) × (expressions or situations)], resulting in a total number of 48 stimuli per condition (over runs). Each condition included 24 distinct video clips, each presented twice; the second presentation was horizontally (i.e., left *vs* right) flipped to increase the number of stimuli. Clips were presented at the center of the screen for 4 sec, followed by a 1750 msec response window, and a 250 msec blank screen. The clips were presented in a jittered, event-related design and a central fixation cross was presented between trials with a variable inter-stimulus interval of 0–14 sec (average 8 sec). Stimulus presentation schedules used a first-order counterbalancing constraint such that each condition preceded each other with approximately equal probability across the experiment. Condition assignments to positions within this sequence were randomized across participants. The order of individual

Table 2 – Demographic information per study sample (Study 1, Study 2, Combined) and group with respective test of between group differences.

Sample	Group	KBIT-2			Age (years)			Sex			Handedness		
		Mean (SD)	t	*p*	Mean (SD)	t	*p*	(Female/male)	X^2 (df)	*p*	(Left/right)	X^2 (df)	*p*
Study 1	NT	112.14 (11.88)	1.49 (35)	.146	28.19 (5.74)	.771 (35)	.446	7/14	.302 (1)	.583	0/21	n.a.	n.a.
	ASD	118.06 (12.12)			26.48 (7.33)			4/12			0/16		
Study 2	NT	117.63 (11.51)	−2.22 (50)	**.031**	28.13 (7.22)	.599 (50)	.522	15/15	3.99 (1)	**.046**	4/26	.229 (1)	.632
	ASD	109.91 (13.51)			29.41 (8.06)			5/17			4/18		
Combined	NT	115.17 (11.89)	−.890 (78)	.376	27.25 (7.16)	.614 (78)	.541	22/28	3.47 (1)	.063	4/50	.593 (1)	.441
	ASD	112.52 (14.46)			28.27 (7.17)			7/23			4/30		

Bold values indicate a significance value below $p < .05$.
Abbreviations: ASD, Autism Spectrum Disorder; df, degrees of freedom; KBIT-2, Kaufmann Brief Intelligence Test – Second Edition; n.a., not applicable; NT, neurotypically developed; *p*, significance value; SD, Standard Deviation; t, t-value from an independent samples *t*-test; X^2, chi-square test value for the goodness of fit.

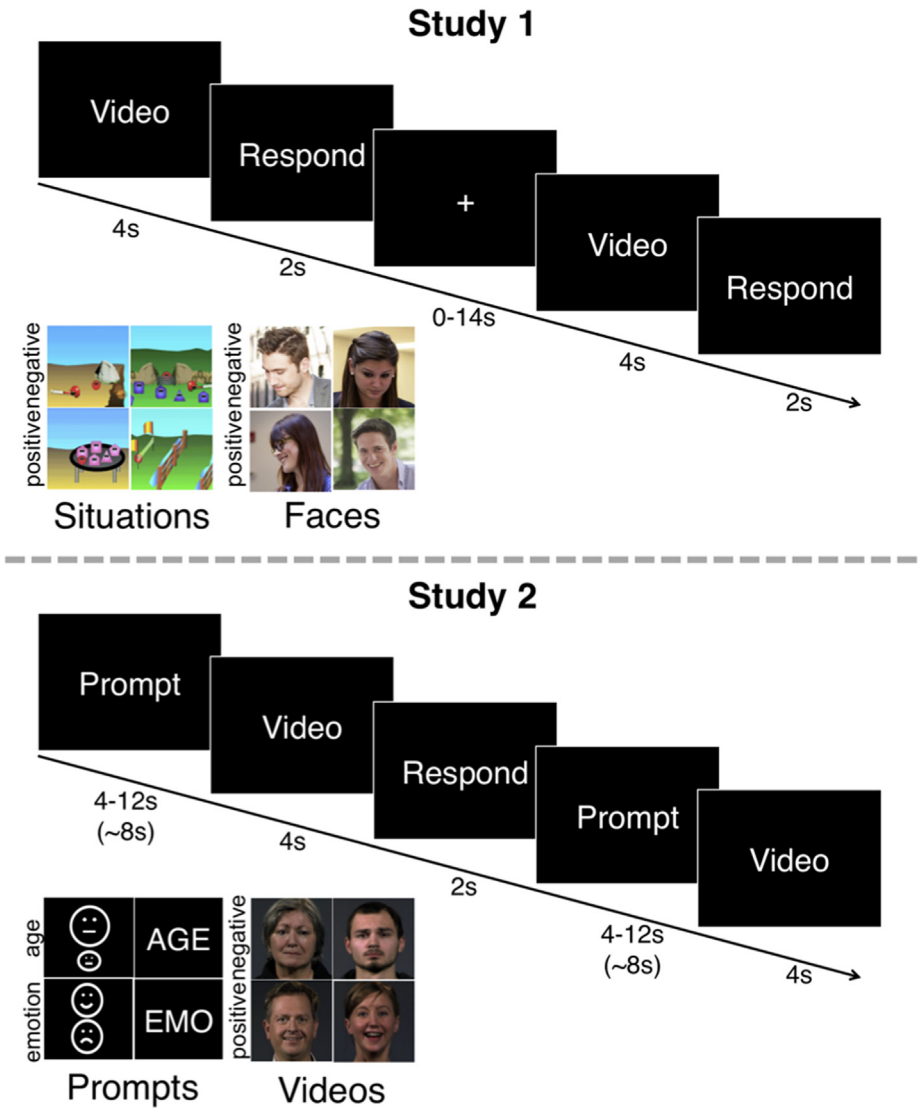

Fig. 1 – Task per study. In the *emotion attribution task* (Study 1, upper) trials started with the display of short movie clips (4s) showing positive or negative dynamic facial expressions (Faces) or animated situations (Situations). Participants were asked to rate the emotional intensity (1–4, neutral to extreme) within the following 2 sec. Durations of inter-trial intervals were jittered between 0 and 14 sec (~8 sec). In the *age/emotion attribution task* (Study 2, lower) trials started with the display of a prompt (either a word or symbol cue), indicating whether participants were asked to make a judgment about the age (older *vs* younger than 30 years) or the emotional valence (positive *vs* negative) on a given trial. After the prompt, participants viewed a short facial expression movie clip (4 sec) and were asked to indicate their response within 2 sec afterwards. Duration of prompts was jittered between 4 and 12 sec (~8 sec).

stimulus clips for a given condition was chosen pseudo-randomly for each participant, with the constraint that repetitions of each stimulus occurred in the same even-odd folds as the first presentation (e.g., an event first presented in run 2 would be repeated in run 6, and an event presented in run 3 would be repeated in run 7). The last trial in each run ended with the presentation of a blank screen for 12 sec, resulting in a total run time of 595 sec.

2.2.2. *Study 2: Emotion/age attribution task*

Participants watched 192 short movie clips of dynamic facial expressions and judged the valence of the emotional expression (emotion task: positive *vs* negative) or, to direct attention away from emotions, judged the individual's age (age task: over *vs* under 30 years old). We chose 20 actors (10 males, 10 females) expressing 10 different emotional states [5 positive (amused, thankful, enthusiastic, happy, confident), 5 negative (disgusted, angry, worried, sad, furious)] from a larger set of stimuli (for details on the stimuli production, see Kliemann, Rosenblau, Bolte, Heekeren, & Dziobek, 2013). For each participant, 192 videos were drawn from the resulting set of 200. Half of the actors in each gender category were 'older adults' and the other half 'younger adults'. An independent MTurk study validated that participants could readily discern

the emotional valence, and the actors' age range ('over 30 years' *vs* 'under 30 years'), and viewed the emotions as equivalently "believable" across age and valence.

The videos were presented over 6 runs, each containing 32 trials per run (16 positive, 16 negative). Each trial started with a task-prompt screen indicating the task (emotion *vs* age) for a given trial, presented for varying durations (4–12 sec, mean = 8 sec). The clips were presented at fixation for 4 sec, followed by 250 msec blank and 1.75 sec response screen. Prompts were presented in two formats: three letters (emotion task: "EMO"; age task: "AGE") or iconic symbols (emotion task: smiling and sad emoticons; age task: small and bigger neutral emoticon; see Fig. 1). Response screens were identical for both task conditions, consisting of a plus and a minus symbol (emotion task: plus = positive, minus = negative; age task: plus = 'over 30', minus = 'under 30'), and their position was randomized across trials. Participants responded by pressing the left or right button. The next trial started immediately after the response screen. Presentation order of the four main conditions (positive/negative expression, older/younger character) within each and over all runs was optimized using Optseq2 (https://surfer.nmr.mgh.harvard.edu/optseq/) with a first-order counterbalancing constraint. The order of items within a scheduled condition was then pseudo-randomized across runs, with the constraint that each movie clip was presented once in each task condition over runs. Ordering of response option arrangement, gender of the face, and task prompt format were balanced within runs (i.e., each run had the same number of females, symbol prompts, etc.). The last trial in each run ended with the presentation of a blank screen for 12 sec, resulting in a total run time of 492 sec.

Participants were trained on the tasks and completed one practice run before the scan, with different clips, to ensure understanding of task and response requirements.

2.3. *Localizer tasks*

In both studies, participants also completed localizer tasks [theory of mind (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011) and face localizer (Hariri, Bookheimer, & Mazziotta, 2000; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011)] reported in detail elsewhere. In addition to the main Regions of Interest (ROI) and voxel selection process (see below), we conducted a secondary analysis in which the localizer data (rather than the facial expressions data) were used for the feature selection process (see Supplementary Material). The results confirm the main MVPA analyses and results reported here.

2.4. *Behavioral tasks*

Participants in both studies performed behavioral tasks outside of the scanner to characterize intellectual functioning level [crystallized (verbal) and fluid (nonverbal) intelligence; Kaufmann Brief Intelligence Test, Second Edition (KBIT-2), Kaufman, 2004)] and social functioning [level of autistic symptoms in typical populations with the Autism Spectrum Quotient (AQ; Baron-Cohen et al., 2001)].

In Study 2, all participants also completed the explicit version of the Face Puzzle tasks (Kliemann et al., 2013). The task measures behavioral facial emotion recognition performance from dynamic videos including positive and negative basic, as well as more complex social emotions. Participants viewed 25 facial emotional expression videos (11 positive, 14 negative) and had to choose the correct emotional label word for the expressed emotion out of four options. The task has previously been shown to be sensitive to social impairments in ASD versus controls and is described in detail elsewhere (see Kliemann et al., 2013). This task allowed for further investigation of whether valence processing was affected at all in the ASD sample by investigating the types of errors made (i.e., an error analysis). Of the four labels, the three distractor words were designed as follows: two were of the same valence, with either close (Error Type 1, ET1) or distant (Error type 2, ET2) intensity, whereas the third distractor was of opposite valence (Error Type 3, ET3). Note that the intensity and valence ratings as the basis for the task constructions were not actual ratings for the stimuli, but ratings of the emotion words (Hepach, Kliemann, Gruneisen, Heekeren, & Dziobek, 2011).

2.5. *fMRI acquisition*

Data were acquired on a 3-T Tim Trio scanner (Siemens; Erlangen, Germany) at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a Siemens 32-channel phased-array head coil. For each subject we collected high-resolution T1-weighted anatomical images (MPRAGE, voxel size = 1 × 1 × 1 mm, TR = 2530 msec, slices = 176, FoV = 256 mm) with whole brain coverage to register functional data to individual and standard anatomy. We then collected functional images acquired with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast (voxel size = 3 × 3 × 3 mm, TR = 2000 msec, TE = 30 msec, flip angle = 90°, FoV = 192 mm). Slices were aligned with the anterior/posterior commissure and provided near whole-brain coverage (excluding the cerebellum).

2.6. *fMRI data analyses*

2.6.1. *Preprocessing*

We used SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) and custom software written in Matlab (www.mathworks.com; Natick, MA, USA) to analyze the MRI data. Each participant's data were first registered to the first image of each run, then all functional runs were co-registered with each other and then with the participant's anatomical scan. All images (functional and anatomical) were normalized to a common (Montreal Neurological Institute, EPI template) brain space. Functional images were smoothed using a Gaussian kernel filter [5 mm FWHM (full-width-half-maximum)]. Smoothing does not substantially affect decoding performance of multi-voxel pattern analyses (Op de Beeck, 2010; Zhang, Meeson, Welchman, & Kourtzi, 2010). Data were high-pass filtered (cut-off 128 sec) to remove low-frequency noise and SPM imaging scaling was applied. Functional data were further corrected for motion artifacts, defined using the ART toolbox (Whitfield-Gabrieli, Nieto-Castanon, & Ghosh, 2011) as timepoints during which motion exceeded 2 mm in any direction relative to the previous timepoint or a change in global signal exceeded a

threshold of three standard deviations from the mean global signal. We additionally included five PCA-based noise regressors created using CompCor (Behzadi, Restom, Liau, & Liu, 2007) within individual subject white-matter masks. Masks were eroded in all directions by two voxels, in order to prevent partial voluming. CompCor regressors were defined using scrubbed data (e.g., artifact timepoints were identified and interpolated over prior to running CompCor).

We first performed whole-brain first level analyses on each participant's functional data by applying a general linear model (GLM) with SPM modeled as a boxcar function using a standard hemodynamic response function (HRF) matching the onset and duration of experiment specific regressors. For Study 1, data were modeled in principle with 8 regressors per run for stimulus (expressions *vs* situations) × valence (positive *vs* negative) × [gender (male *vs* female) for the expressions or story type (social *vs* nonsocial) for the situations condition]. To arrive at a larger set of datapoints per training and test sets for the SVM, we created 16 'pseudo' regressors from the original 8: instead of having 6 trials per 8 conditions we doubled the number of regressors and modeled 3 trials per main condition, i.e., 16. For Study 2, we modeled prompt types (word/symbol × emotion/age task) with onsets at the time of the prompt, and 8 conditions at the time of the videos: task (age *vs* emotion) × expression valence (positive *vs* negative) × character's age (younger *vs* older). Similarly to Study 1, we also doubled the number of regressors for Study 2 by modeling 16 'pseudo-condition' regressors with 2 trials instead of 8 regressors with 4 trials. For both studies, nuisance covariates were added to the model i) for timepoints with head motion artifacts, ii) CompCor regressors, iii) to correct for run effects, and (iv) response screen. To account for variability related to reaction time, we included a parametric mean-centered regressor, with an amplitude on each trial corresponding to the trial's reaction time. If no response was recorded on a single trial, we used the participant's mean reaction time over all trials.

For two participants in Study 1 and three participants in Study 2, fMRI data could not be analyzed for one run due to technical reasons or low overall performance in the task.

2.6.2. *Logic of the analyses*

One goal of this research was to test whether, in adults with ASD, the valence of a dynamic facial expression is extracted from variable stimuli by the same cortical regions, to the same degree, as in neurotypically developed control participants. Thus, the key analysis in each experiment tested whether the pattern of activity in each cortical region (in regions of interest and a whole brain searchlight) could be used to classify the valence of the facial expression in the stimulus. For this analysis, we included only the facial expressions condition (in Study 1), and the emotion task (in Study 2). We tested for classification of valence across all participants in each study, and then separately for the ASD and control groups. Then we tested whether there were any significant differences between the two groups, in any region. To foreshadow our results, we did not find significant group differences in any ROI, in either study. We therefore combined the data from the two studies, to increase the power and sensitivity of our analyses, and again tested whether there were any group differences in the classification of valence in facial expressions.

In addition to this shared question across studies, each experiment was designed to test a separate second question. Study 1 depicted valence in both facial expressions and situations, and so was designed to test whether representations of a character's emotional valence generalize across the format of stimulus input. We therefore also tested whether activity could be used to classify emotional valence, when training and testing on distinct stimulus formats (i.e., training on facial expressions and testing on situations, or vice versa; Skerry & Saxe, 2014). We tested whether this classification showed any significant group difference: if adults with ASD construct less efficient or abstract representations of emotion, we would expect to find less robust classification of valence in ASD individuals in this generalization test. Note that the NT data in Study 1 is a re-analysis of the data presented in Skerry and Saxe (2014), however with slightly differing preprocessing and analyses (see Section 2 and Supplementary Material).

Study 2 showed facial expressions of emotions while participants were instructed to attend to emotion, or to attend to a different feature, the character's age. This experiment was designed to test whether representations of others' emotional valence are flexible in response to the participant's own endogenous goals. We therefore also tested whether classification of emotional valence in each brain region was task-dependent, showing better classification of valence when emotion was task-relevant than when it was task-irrelevant. We hypothesized that individuals with ASD might show less flexible social processing, and therefore less change in social representations in response to changes in the task context. If so, the difference in classification of valence between the emotion- versus age-task should be greater in the control group than in the ASD group. Incidentally, this task design also allowed us to test whether each cortical region encoded the orthogonal stimulus feature, the character's age, when this feature was task-relevant.

2.7. *Regions of interest*

Based on prior studies (Peelen et al., 2010; Skerry & Saxe, 2014), we tested five *a priori* regions of interest (ROIs) previously reported to contain information about the emotion/valence of facial expressions: left posterior superior temporal cortex (lpSTC), right middle STS (rmSTS), right fusiform face area (rFFA), as well as dorsal and middle medial prefrontal cortex (d/mMPFC).

Following Skerry and Saxe (2014), to define individual ROIs we first used group-level spatial constraints derived from previous studies. We defined the search space for dMPFC and mMPFC from a theory of mind task (Dufour et al., 2013), for rmSTS and rFFA from a face task (Julian, Fedorenko, Webster, & Kanwisher, 2012) and for lpSTC from Skerry and Saxe (2014). Within these five group-level constraints, we selected features for analysis in each individual subject.

2.8. *Feature selection*

fMRI data is high-dimensional, and due to its limited temporal resolution, experimental designs provide relatively small

numbers of training data examples. In this context, feature selection can be useful to remove high-variance, noisy voxels from the feature set (Di Martino et al., 2009; Mitchell et al., 2004; Pereira, Mitchell, & Botvinick, 2009). We thus selected voxels (i.e., features) within each ROI using a univariate selection procedure. Within each participant and each ROI hypothesis space, we selected the 80 most active voxels based on the contrast of all movie clip stimuli > rest. This criterion is orthogonal to all between-condition classifications. Selecting a fixed number of voxels eliminates differences in the number of voxels analyzed across regions, participants or groups; the choice of this fixed number (n = 80) was determined by pilot testing in an independent set of participants (as described in Skerry & Saxe, 2014).

2.9. Multi-voxel pattern analyses (MVPA)

We conducted multi-voxel pattern analyses (MVPA) using in-house code developed in Matlab (www.mathworks.com; Natick, MA, USA) and the publicly available Library for Support Vector Machines (LIBSVM, https://www.csie.ntu.edu.tw/~cjlin/libsvm/, Chang & Lin, 2011).

2.9.1. ROI-based MVPA

Preprocessing and modeling the fMRI data with SPM (http://fil.ion.ucl.ac.uk/spm/) provided a voxel-wise summary (beta values) for each condition. An SVM model was used to test classification of each contrast, in each ROI and in each participant. Binary classifications were conducted with a linear kernel, using a fixed regularization parameter (C = 1) to control for training error and margin size. We used a linear kernel to perform binary classifications assuming that a linearly decodable signal represents a plausible readout mechanism for downstream neurons (Hung, Kreiman, Poggio, & DiCarlo, 2005; Seung & Sompolinsky, 1993; Shamir & Sompolinsky, 2006). Linear separability within a population can be considered a conservative yet reasonable estimate of information that is available for explicit readout (DiCarlo & Cox, 2007).

Each participant's data were partitioned into cross-validation folds by run of data acquisition. The classifier was trained iteratively on all runs but one (leave-one-run-out), and tested on the remaining run. Thus, training and testing were always performed on independent subsets of the data. Classification accuracy was then averaged across runs, to obtain a single classification accuracy for each participant in each ROI.

For the multivariate analysis of the neural data we tested for above chance classification accuracy in a given region (chance = .5) by performing a one-sample *t*-test (one-tailed) across participants. To correct for multiple comparisons we adjusted the significance threshold by dividing by the number of regions tested (Bonferroni correction: five social brain ROIs gives a corrected α value of .01). To test for between group differences in valence encoding, we conducted repeated measures ANOVAs with the between-subjects factor group (ASD *vs* NT) and condition specific within-subjects factors [e.g., ROI (dMPFC *vs* mMPFC *vs* rFFA *vs* rmSTS *vs* lpSTC); Study 2: task (age *vs* emotion)] while adding IQ and motion as covariates. In addition to the ANOVA, we ran a linear mixed effects model with R (R Core Team, 2012; lme4, Bates, Machler, Bolker, & Walker, 2015) to test whether classification of valence was influenced by the between-subjects factor group (ASD *vs* NT) or interactions between group and any other factor [study (Study 1 *vs* Study 2), handedness (left *vs* right), sex (female *vs* male) and the within-subjects factors ROI (dMPFC *vs* mMPFC *vs* rFFA, *vs* rmSTS *vs* lpSTC), IQ, and head motion]. The model additionally included Subject ID as a random effect. For further details on the linear mixed effects model, please refer to the results section (Section 3) and Table 4.

In addition to standard null hypothesis testing, we included a Bayesian procedure to explicitly compare the evidence *in favor* or *against* the null hypothesis (no group difference). To this end, we used publicly available code (http://cognitivegenetic.rutgers.edu/ptn/; Gallistel, 2009) to test the likelihood that two datasets come from the same distribution (null hypothesis) or from different distributions (alternative hypothesis). In short, we estimated the Bayes Factor (odds) against and for a group difference in classification accuracy of valence, along with the weight (log base 10 of the Bayes Factor) of evidence. Given the natural bounds of classification accuracy in the context of our study (theoretically ranging from 0 to 100 as possible values), we used 0 and 100 as upper and lower analytical limits. Using a two-tailed hypothesis, we assumed that the mean of the experimental (i.e., ASD) data lies within the interval of the true mean of the control (i.e., NT) data ± 50, hence limiting our incremental prior to −50 and + 50.

2.9.2. Searchlight-based MVPA

To investigate classification accuracy in the rest of the brain other than the predefined regions of interest we performed a searchlight analysis. This procedure was identical to the ROI-based approach, except that we applied the classifier iteratively to spheres tiling the whole brain, rather than to *a priori* defined ROIs. For each voxel in a gray matter mask (Harvard/Oxford atlas, Eickhoff et al., 2005, >30% probability grey matter), we defined a sphere containing all voxels within a 3-voxel radius (123 voxels) of the center voxel. The searchlight size was set *a priori* following the procedure as described in Skerry and Saxe (2014). Within each individual sphere, we conducted a *t*-test (all movie clips *vs* rest) to select the 80 most active voxels in the sphere. Classification was then performed on each cross-validation fold, and the classification accuracy for that sphere was assigned to the central voxel. This procedure resulted in whole-brain images for each cross-validation fold, which were then averaged together to generate a single accuracy image for each participant, for a given classification. We then conducted a one-sample *t*-test over subjects' accuracy maps. Resulting maps of t-statistics were corrected for multiple comparisons with $p < .05$, family-wise error (FWE) correction based on Gaussian random fields, similar to Skerry and Saxe (2014).

2.10. Representational Dissimilarity Matrices

In an exploratory analysis we investigated the (dis-)similarity of neural responses to five negative emotions (disgusted, angry, worried, sad, furious) and five positive emotions (amused, thankful, enthusiastic, happy, confident). We computed dissimilarity (Euclidian distance) between the average voxel patterns for each emotion across all runs for

Table 3 – In-scanner head motion per study sample (Study 1, Study 2, Combined) and group with respective test of between group differences.

Sample	Group	Artifact timepoints			Mean translation			Mean rotation			Mean distance		
		Total	t	p	Mean (SD)	t	p	Mean (SD)	t	p	Mean (SD)	t	p
Study 1	NT	52.62 (47.39)	−2.55 (27.3)	**.017**[1]	.069 (.020)	−.019 (35)	.985	.028 (.012)	−1.41 (35)	.168	.156 (.044)	−.061 (35)	.952
	ASD	23.81 (18.43)			.069 (.032)			.023 (.011)			.155 (.071)		
Study 2	NT	18.07 (15.79)	2.52 (23.02)	**.019**[a]	.064 (.032)	2.84 (33.9)	**.008**[a]	.023 (.011)	2.43 (23.8)	**.023**[a]	.144 (.071)	2.82 (34.6)	**.008**[a]
	ASD	52.05 (61.89)			.097 (.048)			.041 (.034)			.217 (.105)		
Combined[b]	NT	32.82 (36.74)	1.26 (45.35)	.213[a]	.065 (.027)	2.68 (42.2)	**.011**[a]	.025 (.012)	1.91 (34.6)	.064[a]	.148 (.062)	2.66 (43.5)	**.011**[a]
	ASD	46.87 (53.88)			.089 (.045)			.036 (.030)			.20 (.096)		

Bold values indicate a significance value below $p < .05$.
Abbreviations: ASD, Autism Spectrum Disorder; df, degrees of freedom; NT, neurotypically developed; p, significance value; SD, Standard Deviation; t, t-value from an independent samples t-test.
[a] Levene's Test revealed inequality of variances between groups. Degrees of freedom, t and p-values are thus reported corrected (equal variances not assumed).
[b] Measures of participants that took part in both studies were averaged. Note that between-groups effects would not change when taking all data points from each participant [all $p < .05$, except mean rotation remaining not significant ($p = .094$)].

Table 4 – Detailed statistics of the linear mixed effects analysis on the classification accuracy for valence of facial expressions across both studies.

Effects	df	t-value	p-value
Group	74	1.046370	.2988
ROI: lpSTC	349	−.337625	.7358
ROI: MMPFC	349	−.337625	.7358
ROI: rFFA	349	−1.641783	.1015
ROI: rmSTS	349	−.278044	.7811
Study	349	.609595	.5425
Handedness	74	1.451886	.5425
Sex	74	.628434	.5317
Motion	349	−1.612323	.1078
Age	349	.253365	.8001
IQ	349	−.713990	.4757
Group*lpSTC	349	−1.129107	.2596
Group*MMPFC	349	−1.182353	.2379
Group*rFFA	349	.797765	.4255
Group*rmSTS	349	−.110424	.9121
Group*study	349	.190986	.8486
Group*handedness	74	−.956207	.3421
Group*sex	74	−.882208	.6123
Group*motion	349	.882208	.3783
Group*age	349	−1.817723	.3019
Group*IQ	349	1.817723	.0700

Note that the factor steps of ROI are reported separately compared to the reference ROI factor step DMPFC.
Abbreviations: df, degrees of freedom.

each ROI (using the same features as in the classification analysis). We normalized each RDM by subtracting its minimum value and dividing by the range, resulting in distances ranging between 0 and 1. All participant RDMs were then averaged to generate one RDM per group (NT, ASD) per ROI. To test whether there was a significant group difference in the overall RDMs, we used a permutation test on the absolute magnitude of the difference between the group average RDMs in each region. To compute null distributions of this difference, we permuted the group labels on the RDMs 5000 times, randomly assigning 30 RDMs to group 1 and 22 RDMs to group 2, and calculated the absolute sum of differences for each permutation. If the observed group difference was larger than 95% of the observed null distribution in any region, this would constitute evidence for a significant group difference in the overall RDM, and would require follow-up testing to investigate the source of the group difference.

2.11. *Participants overlap across studies ($n = 9$)*

For the analyses combining data from Study 1 and Study 2, we dealt with the participants that took part in both studies in two ways: For the one sample t-test (across both groups), the between-group ANOVA and the whole brain analyses, we combined the two datasets from the same participant into a single estimate by, e.g., averaging their respective classification accuracies. For the linear mixed effects model, we used all available data from all participants and accounted for repeated measures by including subject identity as a random effect in the model.

3. Results

3.1. Participants

In Study 1, groups did not differ significantly in age, sex, or intellectual functioning as measured with the KBIT-2 (crystalline and fluid intelligence, in the following referred to as IQ). In Study 2, groups were matched on age but the ASD group showed significantly lower IQ values and had significantly fewer female participants than in the NT group. Details can be found in Table 2. In both studies, the AQ test confirmed low levels of autistic symptoms in the NT group [Study 1, NT: mean = 18.33 (SD = 4.61), ASD: mean = 32.13 (SD = 9.47), $t(35) = 5.84$, $p = 6.2 \times 10^{-13}$; Study 2, NT: mean = 13.67 (SD = 5.73), ASD: mean = 32.46 (SD = 8.39), $t(50) = 9.60$, $p = 1 \times 10^{-6}$]. When combining Study 1 and Study 2, groups did not differ significantly in age, IQ, or number of females.

3.2. Behavioral results

We investigated participants' behavioral performance in the scanner tasks (Study 1: intensity ratings; Study 2: accuracy; in both studies: reaction times) with separate 2 × 2 repeated measures ANOVAs [within-subjects factors condition (Study1: expressions *vs* situations; Study 2: emotion *vs* age task) and valence (positive *vs* negative), between-subjects factor group (NT *vs* ASD), with IQ as covariate]. Reaction times were analyzed for correct trials only in Study 2.

3.2.1. Study 1: Emotion attribution task

Mean intensity ratings across all conditions was 2.44 (range = 1.9–3.3, SD = .32). Overall the intensity of the situation movie clips tended to be slightly higher [mean = 2.61 (SD = .55)] than for the expressions [mean = 2.28 (SD = .24); main effect of stimulus type, $F(1,34) = 3.90$, $p = .056$, $\eta_p^2 = .103$]. Individuals with ASD did not perform differently than control participants [main effect of group, $F(1,34) = 1.74$, $p = .196$, $\eta_p^2 = .049$]; stimulus type*group [$F(1,34) = 3.08$, $p = .088$, $\eta_p^2 = .08$], valence*group [$F(1,34) = .138$, $p = .712$, $\eta_p^2 = .004$], stimulus type *valence*group [$F(1,34) = .22$, $p = .641$, $\eta_p^2 = .006$].

We only recorded reaction times during the 2 sec response screen, hence reaction times longer than 2 sec could not be analyzed. Numbers of trials missed or with reactions times longer than 2 sec did not differ between the groups [ASD: mean = 21%, NT 16%, $t(35) = .89$, $p = .375$]. On average participants responded in 750 msec (range = 550–1100, SD = 108 msec). There were no significant main effects or interactions of group, valence or task (all $p > .14$).

3.2.2. Study 2: Emotion/age attribution task

Across all participants and conditions, the mean accuracy was high [mean = 91.8% correct (SD = 4.8)]. Nevertheless, we excluded the first run for three participants because their performance in the respective run was very low (<75% of trials had correct responses, $p < .05$, binomial sign test). To account for individual subjectivity in the age ratings of certain characters across the task, we adjusted accuracy for the age ratings prior to the imaging analysis. In other words, if a participant consistently (sign test $p < .05$) rated a character in a specific age category (older/young than 30) over the course of the task, we acknowledged these responses as correct, even when they did not conform with the 'ground truth' of the experimental design.

The ANOVA revealed that individuals in the ASD group were significantly less accurate than the NT group [$F(1,49) = 11.57$, $p = .001$, $\eta_p^2 = .19$; ASD: mean = 89.1% (SD = 5.1), NT: mean = 93.6% (SD = 3.3)], regardless of whether they were asked to perform the emotion or age judgment. Individuals with lower IQ scores were also less accurate in their judgments, as shown by a main effect of the IQ covariate [$F(1,49) = 10.65$, $p = .002$, $\eta_p^2 = .17$]. There were no other main effects or interactions (all $p > .1$).

Participants in the NT group responded significantly faster [mean = 631 msec (SD = 89 msec)] than those in the ASD group [mean = 686 msec (SD = 93 msec)], as revealed by a main effect of group [$F(1,49) = 4.64$, $p = .036$, $\eta_p^2 = .087$]. In addition, reaction times were faster for positive than for negative emotions across both groups [$F(1,49) = 5.85$, $p = .019$, $\eta_p^2 = .107$]. The ANOVA further revealed an interaction of the factors valence and IQ [$F(1,49) = 4.48$ $p = .037$, $\eta_p^2 = .085$]. There were no other main effects or interactions, suggesting that participants with ASD were not disproportionately slow on the emotion task. An exploratory analysis of the valence and IQ interactions showed that the difference between faster RTs for positive > negative items was smaller for individuals with higher intellectual functioning [partial correction of the difference between RT for positive > negative trials with IQ while controlling for group: ($r_{(49)} = .292$, $p = .037$)].

3.2.3. Study 2: Face Puzzle behavioral task

Participants in the ASD group showed overall fewer correct responses when labeling the dynamic facial expressions with emotion words than the NT group [ASD: mean = 55.27% (SD = 18.83), NT: mean = 69.07% (SD = 11.69); chance = 25%]. A repeated measures ANOVA with the within-subjects factor valence (positive *vs* negative), between-subjects factor group (ASD *vs* NT) and IQ as covariate yielded a significant main effect of group [$F(1,49) = 6.03$, $p = .018$, $\eta_p^2 = .11$], no interaction or main effect of valence (all $p > .8$). Individuals with ASD showed an overall impairment in recognizing others' emotions from facial expressions.

Three types of 'errors' were possible on the Face Puzzle task (same-valence similar-intensity same-valence dissimilar-intensity, opposite valence). In an ANOVA (error type × group, IQ as a covariate), we found main effects of error type [$F(1,49) = 3.56$, $p = .032$, $\eta_p^2 = .7$] and group [$F(1,49) = 591$, $p = .019$, $\eta_p^2 = .11$], and a trend of an interaction of these two factors [$F(1,49) = 2.96$, $p = .056$, $\eta_p^2 = .06$]. Both groups made most errors by choosing a different label of the same valence. Individuals with ASD made more errors of all types, including more errors of the opposing valence (see Fig. 2).

3.3. fMRI results

3.3.1. In-scanner head motion

Given the importance of accounting for in-scanner head motion when analyzing MRI data (Van Dijk, Sabuncu, & Buckner, 2012), especially in studies of Autism (Deen & Pelphrey, 2012; Yendiki, Koldewyn, Kakunoori, Kanwisher, & Fischl, 2014),

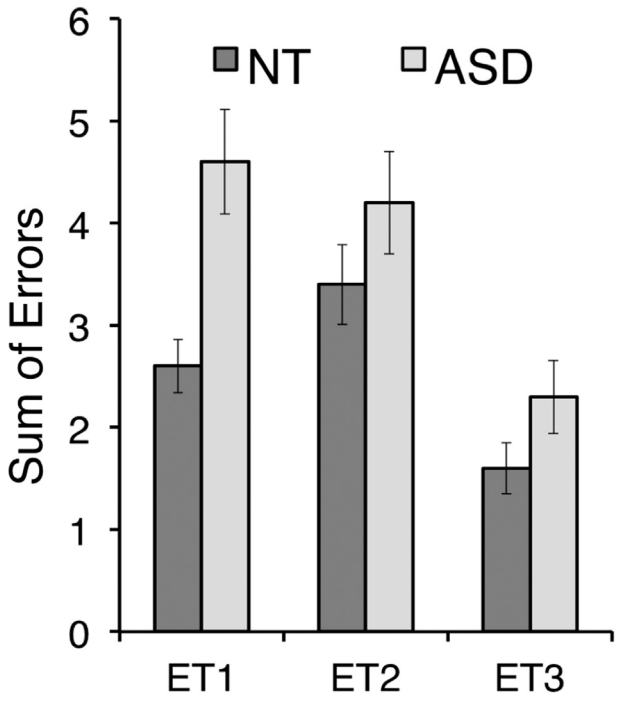

Fig. 2 – Types of errors made in the Face Puzzle task in Study 2 plotted per group. In both groups, most errors were made by choosing a different label of the same valence (ET1 and ET2). Individuals with ASD made more errors of all types, including more errors of the opposing valence (ET3). Error bars represent Standard Error of the Mean (SEM). *Abbreviations*: ET1, error type 1; ET2, error type 2; ET3, error type 3.

we calculated several standard measures to quantify motion in individuals and groups: total number of artifacts (see Section 2.6.1 for identification and removal of motion artifact timepoints), mean translation, mean rotation and mean distance. In Study 1, the NT group had significantly more motion artifacts than the ASD group. This effect was reversed in Study 2 (see Table 3 for details). After the removal of motion artifact timepoints, measures of head motion showed equal amounts of motion in the remaining, analyzed timepoints between groups in Study 1 but higher levels of motion in the ASD group in Study 2.

In the analysis combining data from both studies, there was no difference in the number of total motion artifacts between groups, but ASD participants had more motion than NT participants in the remaining scrubbed (analyzed) data. To additionally control for the effect of head motion on the between group results we include mean translation (as a representative variable of individual levels of head motion) as a covariate/nuisance variable when testing for group effects. Additionally, as described in the methods, five PCA-based regressors defined in white matter were included in the model in order to capture changes in signal intensity driven by noise.

3.3.2. *Valence encoding from facial expressions*

First we asked whether valence of dynamic facial expressions is encoded in social brain regions in ASD as in NT. We therefore quantified the information about emotional valence, in dynamic facial expressions, across cortical regions (see Fig. 3).

3.3.2.1. Study 1. Across all participants, we found significant encoding of valence averaged over all ROIs [mean = 54.95 (SD = 4.19), t(36) = 7.18, $p = 1 \times 10^{-8}$] and in each of the five ROIs separately [dMPFC: mean = 56.69 (SD = 7.33), t(36) = 5.56, $p = 1.2 \times 10^{-6}$; mMPFC: mean = 53.60 (SD = 7.77), t(36) = 2.82, $p = .004$; rFFA: mean = 53.11 (SD = 6.87), t(36) = 2.76, $p = .005$; rmSTS: mean = 55.19 (SD = 7.76), t(36) = 4.07, $p = 1.2 \times 10^{-4}$; lpSTC: mean = 56.13 (SD = 8.69), t(36) = 4.29, $p = 6.4 \times 10^{-5}$].

In the control group, averaged over all ROIs, responses of multivoxel patterns classified valence facial expressions significantly above chance [mean = 55.03 (SD = 3.91), t(20) = 5.88, $p = 4.7 \times 10^{-6}$]. When investigating each ROI separately, the classifier significantly decoded stimulus valence in the medial prefrontal cortex ROIs [dMPFC: mean = 56.10 (SD = 7.09), t(20) = 3.94, $p = 4.1 \times 10^{-4}$; mMPFC: mean = 55.43 (SD = 8.42), t(20) = 2.96, $p = .0039$] and superior temporal cortices [rmSTS: mean = 54.17 (SD = 7.30), t(20) = 2.62, $p = .008$; lpSTC: mean = 57.22 (SD = .05), t(20) = 3.66, $p = 7.8 \times 10^{-4}$], but failed multiple comparison correction for rFFA [rFFA: mean = 52.23 (SD = 5.89), t(20) = 1.73, $p = .049$]. As mentioned above, the NT data presented here are re-analyzed from Skerry and Saxe (2014).

Similarly, the ASD group showed above chance classification of valence when averaging across all ROIs [mean = 54.84 (SD = 4.65), t(15) = 4.16, $p = 4.2 \times 10^{-4}$]. Testing each ROI separately revealed robust decoding of valence from voxel responses in dMPFC [mean = 57.48 (SD = 7.79), t(15) = 3.84, $p = 8\ 1 \times 10^{-4}$] and rmSTS [mean = 56.45 (SD = 8.02), t(15) = 3.13, $p = .003$] but not mMPFC [facial expressions: mean = 51.19 (SD = 6.27), t(15) = .765, $p = .23$]. Accuracies in rFFA and lpSTC did not pass multiple comparison correction [rFFA: mean = 54.27 (SD = 8.02), t(15) = 2.13, $p = .025$; lpSTC: mean = 54.70 (SD = 8.25), t(15) = 2.28, $p = .019$].

A repeated measures ANOVA testing directly for group differences revealed no significant main effect of group [$F(1,33) = 1.09$, $p = .743$, $\eta_p^2 = .003$] or interaction of group by ROI [$F(1,33) = 1.44$, $p = .231$, $\eta_p^2 = .043$].

3.3.2.2. Study 2. We found significant encoding of valence across all participants averaged over all ROIs [mean = 53.34 (SD = 5.25), t(51) = 5.88, $p = 1.6 \times 10^{-7}$] and in all ROIs separately [dMPFC: mean = 54.02 (SD = 8.48), t(51) = 3.42, $p = 6.2 \times 10^{-4}$; mMPFC: mean = 53.29 (SD = 8.67), t(36) = 2.74, $p = .004$.; rFFA: mean = 53.61 (SD = 9.08), t(51) = 2.87, $p = .003$; rmSTS: mean = 54.17 (SD = 9.44), t(51) = 3.18, $p = .001$], except in lpSTC [mean = 51.59 (SD = 8.), t(51) = 1.40, $p = .08$].

In the control group, averaged over all ROIs, classification accuracy of valence was significantly above chance [mean = 53.76 (SD = 5.56), t(29) = 3.71, $p = 4.4 \times 10^{-4}$]. Decoding in dMPFC [mean = 54.38 (SD = 8.38), t(29) = 2.86, $p = .004$], mMPFC [mean = 53.96 (SD = 8.41), t(29) = 2.58, $p = .008$] and rmSTS [mean = 55.00 (SD = 9.09), t(29) = 3.01, $p = 003$] were significant when correcting for multiple comparisons, whereas accuracy in rFFA [mean = 52.78 (SD = 9.54), t(29) = 1.59, $p = .061$] and lpSTC [mean = 52.71 (SD = 9.51), t(29) = 1.56, $p = .065$] showed only trends towards significance.

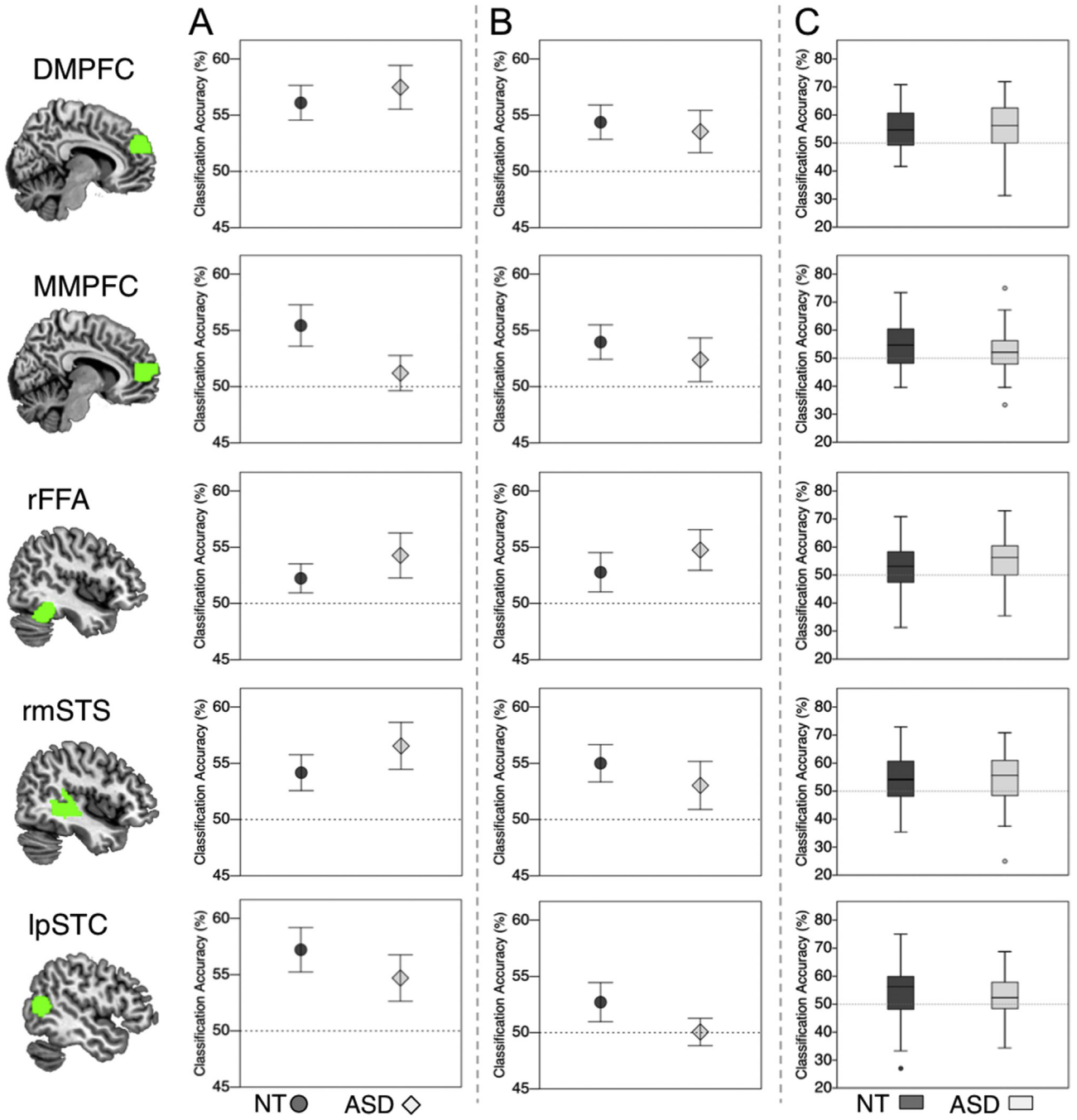

Fig. 3 – Valence encoding from facial expressions. Mean of classification accuracy per group and ROI (rows) for Study 1 (column A), Study 2 (column B), as well as boxplots as a measure of variance across both studies (column C). *Abbreviations*: ASD, Autism Spectrum Disorder; DMPFC, dorso-medial prefrontal cortex; rFFA, right fusiform gyrus; MMPFC, middle-medial prefrontal cortex; NT, neurotypically-developed; rmSTS, right middle superior temporal sulcus; ROI, region of interest; lpSTC, left posterior superior temporal cortex.

In the ASD sample, average classification accuracy over all ROIs was significantly above chance [mean = 52.75 (SD = 5.56), t(21) = 2.66, p = .007]. Investigating each ROI separately showed that only rFFA passed multiple comparison correction [mean = 54.75 (SD = 8.49), t(21) = 2.63, p = .008]. DMPFC [mean = 53.54 (SD = 8.79), t(21) = 1.89, p = .036] did not pass the multiple comparison threshold, rmSTS [mean = 53.03 (SD = 9.99), t(21) = 1.42, p = .085] showed a trend towards significance and mMPFC [mean = 52.39 (SD = 9.12), t(21) = 1.23, p = .12] and lpSTC [mean = 50.06 (SD = 5.72), t(21) = .047, p = .48] did not classify valence.

A repeated measures ANOVA, testing directly for group differences revealed no significant main effect of group [F(1,48) = .077, p = .783, η_p^2 = .002] or interaction of group by ROI [F(1,48) = .541, p = .706, η_p^2 = .011].

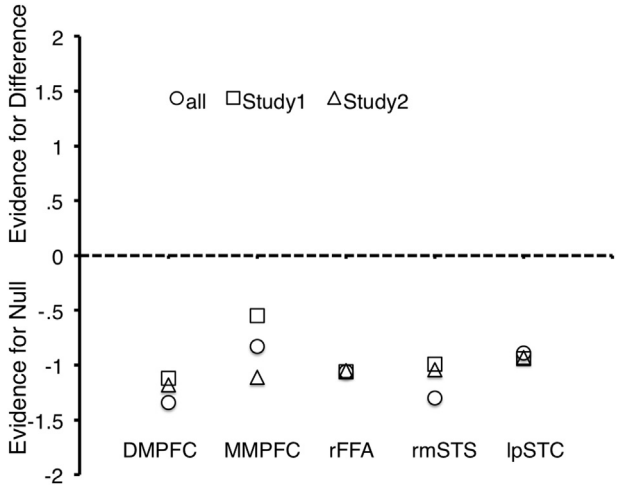

Fig. 4 – Weights of evidence. Plotted *against* (above dotted line) and *for* (below dotted line) evidence in favor of the null hypothesis (no group differences) for the classification accuracy of valence in facial expressions per ROI and samples (Study 1: rectangle, Study 2: triangle, combined: circle). Value on the y-axis represent the weight of the evidence (log_{10} of Bayes Factor) for (positive) or against (negative) the null hypothesis. *Abbreviations*: ASD, Autism Spectrum Disorder; DMPFC, dorso-medial prefrontal cortex; rFFA, right fusiform gyrus; MMPFC, middle-medial prefrontal cortex; NT, neurotypically-developed; rmSTS, right middle superior temporal sulcus; ROI, region of interest; lpSTC, left posterior superior temporal cortex.

Table 5 – Bayes Factor [odds by which the null hypothesis (no group difference) is favored] for the social brain ROIs combined (all) and per study (Study 1, Study 2).

ROI	All	Study 1	Study 2
DMPFC	21.7	13.1	14.9
MMPFC	6.7	3.6	12.7
rFFA	11.6	11.5	11.2
rmSTS	20.1	9.7	10.8
lpSTC	7.8	8.8	8.6

Abbreviations: ROI, region of interest; DMPFC, dorso-medial prefrontal cortex; MMPFC, middle-medial prefrontal cortex; rFFA, right fusiform gyrus; right middle superior temporal sulcus; lpSTC, left posterior superios temporal cortex.

3.3.2.3. Combining study 1 and study 2

3.3.2.3.1. ROI based MVPA analyses. Study 1 and Study 2 both tested perception of valence in brief dynamic facial expressions, using slightly different stimuli and tasks. We combined data from these studies into one large sample to increase power to detect potential group differences.

We first tested for overall classification of valence. Over all participants, we found significant classification accuracies in all social brain ROIs [DMPFC, mean = 54.99 (SD = 8.03), $t(79) = 5.56$, $p = 3.5 \times 10^{-7}$; MMPFC, mean = 53.64 (SD = 8.02), $t(79) = 4.06$, $p = 1.2 \times 10^{-4}$; rFFA, mean = 53.36 (SD = 8.00), $t(79) = 3.76$, $p = 3.3 \times 10^{-4}$; rmSTS, mean = 54.13 (SD = 8.37), $t(79) = 4.42$, $p = 3 \times 10^{-5}$; lpSTC, mean = 53.21 (SD = 8.64), $t(79) = 3.32$, $p = .001$].

A repeated measures ANOVA testing directly for group differences again revealed no significant main effect of group [$F(1,76) = .077$, $p = .782$, $\eta_p^2 = .001$] or interaction of group by ROI [$F(1,76) = .1.31$, $p = .265$, $\eta_p^2 = .017$].

To test the effect of group on valence classification in all collected data across studies, and to test for interactions of group with other potentially relevant factors (i.e., sex, handedness, IQ, motion, study; see Section 2 for model specification), we conducted a linear mixed effects regression analysis. The results of this model again showed no significant main effects or interactions with group (all $p > .5$ see Table 4 for details on the model statistics).

Weighing the evidence for and against a group difference between the ASD and NT group with a Bayes analysis approach resulted in moderate evidence in favor of the null hypothesis for all five ROIs tested for each study separately, and stronger evidence when combining all data (see Fig. 4 plotting the weights and Table 5). Using all data from both studies (n = 89) the odds in favor of the null hypothesis in each ROI range between 6:1 (MMPFC) and 21:1 (DMPFC). Thus, the current results suggest reasonable confidence in accepting the null hypothesis that there are no group differences in valence extraction from the social brain regions tested.

In sum, these analyses suggest that the valence of dynamic facial expressions is represented in response patterns of social brain regions, in both adults diagnosed with ASD and typically developing control participants. Although supporting the null hypothesis with regards to group differences, these are not "null results": on the contrary, significantly above-chance classification of valence was obtained across ROIs for both groups.

Univariate analyses similarly revealed no group difference in average magnitude of response in any region (see Supplementary material for details).

3.3.2.3.2. Whole brain searchlight-based MVPA. To test whether group differences might exist outside our *a priori* ROIs we conducted a searchlight procedure across the whole brain in the combined data from Study 1 and Study 2. Across all participants, the whole brain searchlight revealed five significant clusters where patterns of activity could classify valence ($p < .05$, $k > 9$, FWE correction for multiple comparison): dorsal medial prefrontal cortex, superior and middle temporal gyrus, left postcentral gyrus and middle occipital gyrus (see Table 6 and Fig. 5). However, no clusters showed differential classification across groups (NT > ASD, or NT > ASD).

3.3.3. *Generalization of valence information across stimulus formats (Study 1)*

Abstract information about a character's emotional valence would generalize across different formats of stimulus (facial expressions and animated situations). We trained a valence classifier on patterns of activity in one stimulus condition (e.g., expressions) and tested the classification accuracy in the other stimulus condition (e.g., situations). As reported in Skerry and Saxe (2014), in control participants, classification of valence when generalizing across stimulus type was

Table 6 – Whole brain searchlight results, Combined samples ($p < .05$, $k > 9$, FWE corrected).

Cluster	Region	Number of voxels	Peak t	x	y	z
1	Medial Frontal Gyrus	39	6.75	4	48	32
2	Postcentral Gyrus (left)	13	6.67	−48	−34	50
3	Middle Occipital Gyrus (left)	12	6.42	−32	−92	0
4	Superior Temporal Gyrus (left)	15	6.42	−54	−52	10
5	Middle Temporal Gyrus (left)	11	6.08	−54	−40	0

Abbreviations: t, t-value.

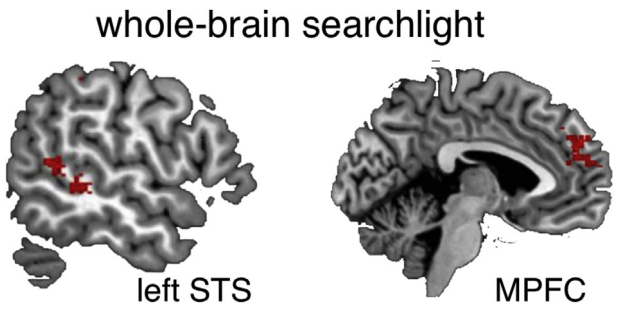

Fig. 5 – Whole brain results for valence encoding from facial expressions across both studies. The searchlight approach across the whole brain revealed significant clusters of activation ($p < .05$, FWE corrected) in medial prefrontal cortex, as well as left posterior and middle superior temporal sulcus across both studies and across both groups. For additional regions and details see Table 4.

observed in dMPFC [mean = 54.87 (SD = 6.32), t(20) = 3.52, $p = .001$] and mMPFC [mean = 51.97 (SD = 3.69), t(20) = 2.45, $p = .012$, see Fig. 6]. In the ASD group, generalization was also observed in both dMPFC [mean = 53.52 (SD = 6.24), t(15) = 2.52, $p = .012$], and mMPFC [mean = 53.81 (SD = 4.33), t(15) = 3.52 $p = .002$]. A repeated measures ANOVA testing directly for group differences again revealed no significant main effect of group and no interaction of group by ROI [ROI*group (F(1,33) = 1.18, $p = .725$, $\eta_p^2 = .034$), group (F(1,33) = 2.48, $p = .125$, $\eta_p^2 = .07$)].

3.3.4. *Flexible representation of valence (Study 2)*

A flexible representation of valence would render valence information more robustly classifiable when valence is task-relevant. Thus classification accuracy of an expression's valence should be higher when participants attend to emotion, in the emotion attribution task, than when they attend to a different stimulus feature, in the age attribution task.

In the NT group, the ANOVA yielded a significant main effect of task [F(1,29) = 5.97, $p = .021$, $\eta_p^2 = .17$], suggesting flexibility of valence representation in social brain regions (see Fig. 8). There were no effects or interactions by ROI.

In the ASD group, there was a trend towards an effect of task [F(1,21) = 4.18, $p = .054$, $\eta_p^2 = .17$; see Fig. 8] on classification accuracy. Note that the size of the task effect is similar in both groups.

When using a repeated measures ANOVA testing directly for group differences we again found no significant main effect of group (i.e., groups did not differ), and no interactions with the factor group (all $p > .4$).

However, there was a significant interaction of task and head motion [F(1,48) = 5.90, $p = .019$, $\eta_p^2 = .11$]. To further quantify the contribution of head motion to the flexible valence representation across all participants (irrespective of group), we conducted a correlation between head motion and classification accuracy (averaged across all ROIs) for the emotion as well as for the age condition. We found a negative correlation for head motion and valence representation in the emotion condition [Pearson's r(52) = −.34, $p = .013$] but not the age condition [Pearson's r(52) = −.06, $p = .68$]. Correlations differed significantly between conditions [Fisher's $z = .04$ (2-tailed)]. In other words, in the emotion task condition only, there was information about the expressions' valence, which was degraded by increasing head motion. We believe that this effect reflects the overall vulnerability of multiple-voxel pattern analyses to in-scanner head motion.

3.3.4.1. Representations of finer-grained emotional expressions (Study 2).

In addition to valence, the responses in these cortical regions may differentiate finer-grained emotions. The face movies in Study 2 depicted five negative emotions (disgusted, angry, worried, sad, furious) and five positive emotions (amused, thankful, enthusiastic, happy, confident). However, there was no evidence of a group difference between the average RDMs for NT and ASD participants in any ROI (permutation test, all observed differences within 95% of the null distribution, see Fig. 7).

3.3.5. *Encoding of character's age in the videos (Study 2)*

Although not the focus of this study, Study 2 allowed us to test for encoding of the character's age in neural patterns of cortical (social) brain regions, when this feature was task-relevant (i.e., during the age attribution task condition) versus when it was not task-relevant (i.e., during the emotion attribution task). This can be considered a different measure of the flexibility of social information processing.

We found a significant main effect of task on age classification in the NT group [F(1,29) = 7.59, $p = .010$, $\eta_p^2 = .21$]: the character's age could be classified from the pattern of response during the age attribution task [mean = 54.17 (SD = 5.14)] but not in the emotion attribution task [mean = 51.32 (SD = 4.25)].

There were no significant main effects or interactions in the data from the ASD group (all $p > .4$). A repeated measures ANOVA testing directly for group differences revealed a significant main effect of group [F(1,48) = 7.34, $p = .009$, $\eta_p^2 = .13$] and a trend towards an interaction of task and group

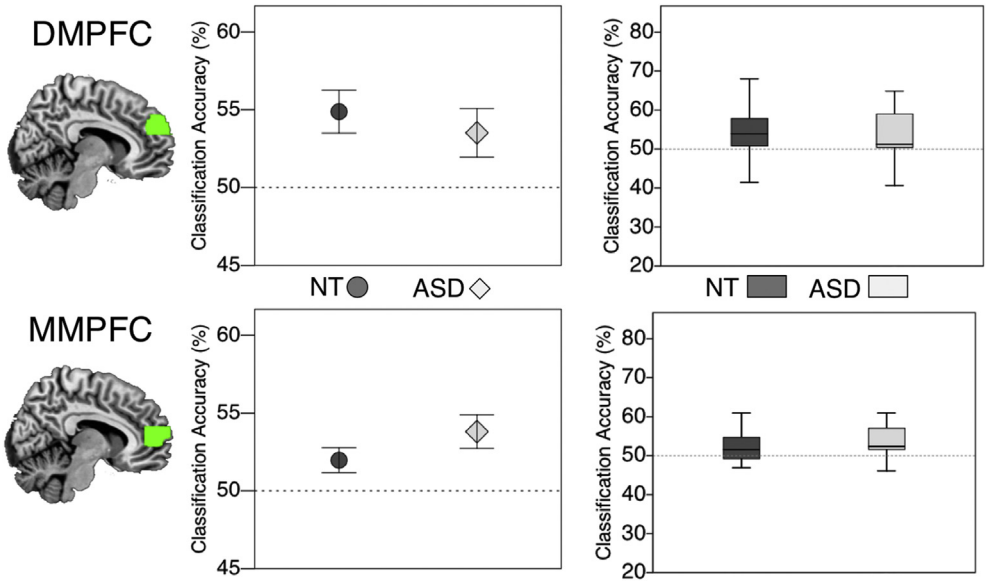

Fig. 6 – Stimulus-type independent encoding of valence. In Study 2, DMPFC (upper row) and MMPFC (lower row) showed above chance classification accuracy for both groups (middle column) and no differences in variance between groups (right columns), as visualized by boxplots. *Abbreviations*: ASD, Autism Spectrum Disorder; DMPFC, dorso-medial prefrontal cortex; MMPFC, middle-medial prefrontal cortex; NT, neurotypically-developed.

[$F(1,48) = 3.48$, $p = .068$, $\eta_p^2 = .07$]. Thus, unexpectedly, individuals with ASD appeared to be less likely than controls to represent a character's age (but not emotion) in their social brain regions, when this stimulus aspect was task-relevant. Since the character's age was not manipulated in Study 1, this unexpected finding could not be replicated in our data, but could be a potential target for future experiments.

3.3.6. *Variance analyses*

One re-occurring argument in studies reporting the absence of activation differences between Autism versus neurotypical groups is that levels of variance may differ between the groups in the measure of interest. We sought to test this hypothesis in our data and conducted Levene tests for homogeneity (equality) of variances (Levene, 1960). This test is less sensitive to potential violations of the normality assumption [than, e.g., the Bartlett test (Snedecor & Cochran, 1989)] thus it represents a good strategy to apply the same test across a variety of measures. In particular, we tested for equality of variances in classification accuracies between groups per ROI averaged across studies for facial expressions and per study (i.e., Study 1: across stimulus conditions, Study 2: for the age task condition). In short, there were no significant differences (all $p > .1$) in variances between the groups in any neural measure tested (see Tables 7 and 8 for details on the statistics). For visualizations of variance per group see Figs. 3, 4, 6 and 8.

4. Discussion

We measured the cortical activity underlying sensitive and flexible social perception, using MVPA. The pattern of activity in an observer's "social brain regions", including medial prefrontal cortex and superior temporal sulcus, could be used to decode the valence of an observed facial expression from naturalistic dynamic videos. Patterns of activity in MPFC generalized from facial expressions to other cues to emotion, like events (e.g., social exclusion). Across many regions, patterns of activity were sensitive to the observer's endogenous goals, making valence more decodable when emotions are task-relevant. These patterns of activity reflect the flexible extraction of emotionally-relevant information, from brief dynamic stimuli, that generalizes to new people and new expressions. Thus, this task seems to require the kind of social processing that individuals with ASD describe as unusually challenging. Surprisingly, individuals with ASD showed (i) similar decoding of emotional expressions, in the same cortical regions, (ii) similar responses to emotional valence that generalize across facial expressions and animated social events; and (iii) similar flexibility in response to task-relevance.

Historically, cognitive neuroscientists have been reluctant to make inferences based on findings of no group difference, or "null results", for good reasons. True group differences could easily be missed by typically under-powered studies with small samples of heterogeneous groups; and standard statistical analyses are designed only to assess evidence against the null hypothesis, not evidence for it. However, the current data are not just "null results" in this familiar sense. The 'null result' of no group difference is composed of a series of *positive results* in the ASD sample: including replicable, reliable classification of emotional valence from dynamic videos across multiple cortical regions. The observed absence of a group difference is evidentially supported and theoretically important.

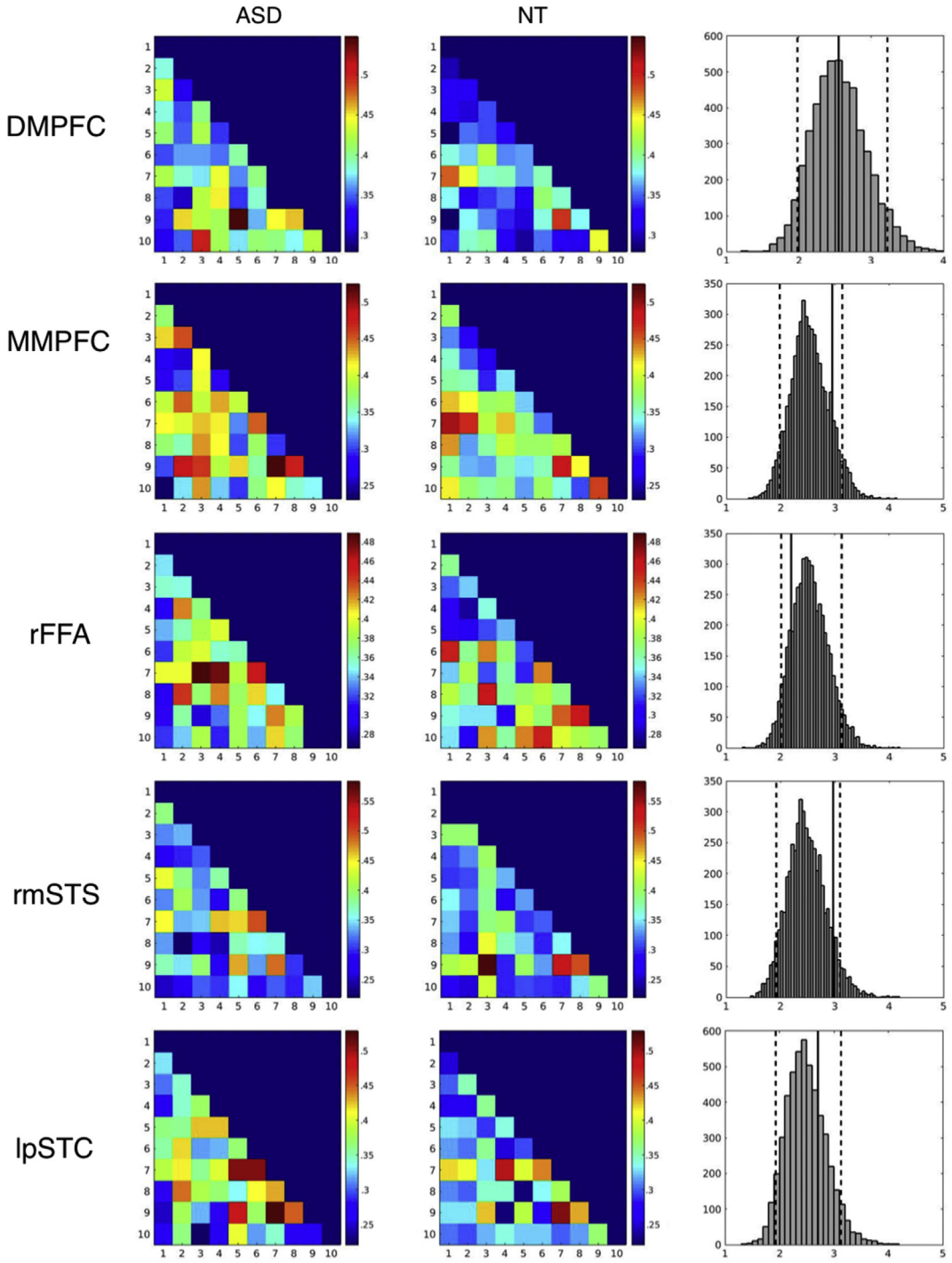

Fig. 7 – Representations of finer-grained emotional expressions per group (Study 2). Representational Dissimilarity Matrices of neural responses to the 10 emotional categories per ROI and averaged for the ASD (left column) and NT group (middle column). Only values for the lower diagonal fields of the RDM are displayed. The first five numbers represent negative emotions (1 disgusted, 2 angry, 3 worried, 4 sad, 5 furious) the last five positive (6 amused, 7 thankful, 8 enthusiastic, 9 happy, 10 confident). Histograms of permuted sum of absolute differences per RDM fields per region (right column). The observed sum (between the actual ASD and NT group assignment) is plotted as the solid vertical line. The 5th (left) and the 95th (right) quantile of the permuted sums are displayed as dotted lines.

Statistically, the current data provide strong evidence against a diagnostically relevant difference between groups in the neural measures. To increase power, we combined data from two studies (resulting in 89 datasets from 80 participants). In a Bayesian analysis, we found reasonably strong support for the conclusion that ASD and control datasets were drawn from the same underlying distribution (Gallistel, 2009; Kass & Raftery, 1995; Rouder, Speckman, Sun,

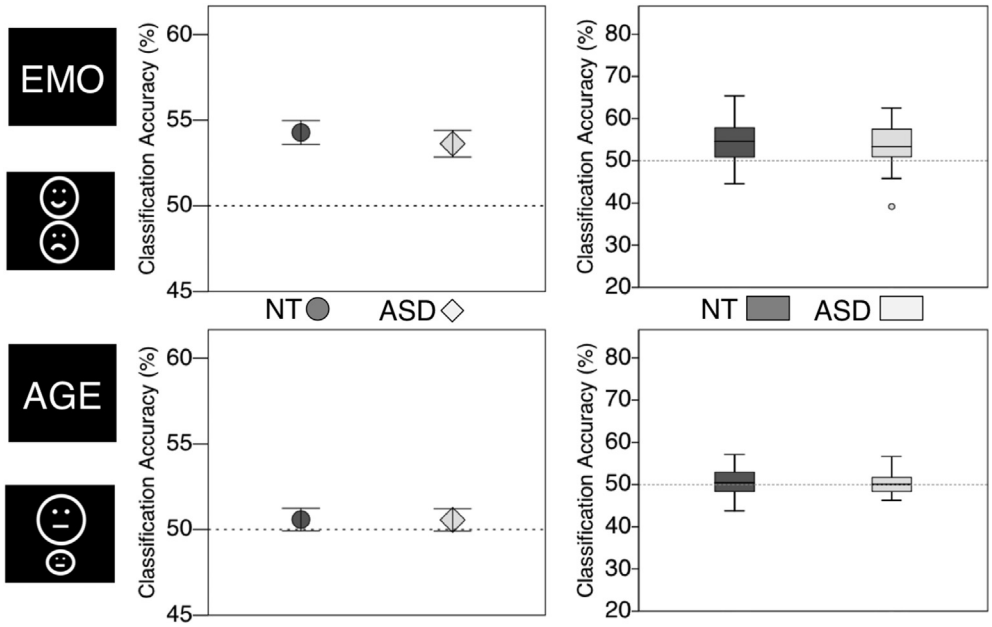

Fig. 8 – Task-dependent encoding of valence. In Study 1, both the NT and the ASD groups showed above chance classification (left column) of valence and when endogenously directed to focus their attention on emotion (upper row), but not when focusing on the character's age (lower row). Variance between groups is displayed in boxplots (right column). *Abbreviations*: ASD, Autism Spectrum Disorder; NT, neurotypically-developed.

Table 7 – Test for inequality of variances in classification accuracies for encoding valence from facial expressions.

ROI	Study	F(df)	*p*
DMPFC	Study 1	.001 (35)	.973
	Study 2	.022 (50)	.882
	Combined	.004 (78)	.952
MMPFC	Study 1	.758 (35)	.390
	Study 2	.144 (50)	.705
	Combined	1.473 (78)	.29
rFFA	Study 1	.144 (50)	.705
	Study 2	2.23 (35)	.144
	Combined	.014 (78)	.907
rmSTS	Study 1	.519 (35)	.476
	Study 2	.002 (50)	.967
	Combined	.140 (78)	.710
lpSTC	Study 1	.352 (35)	.557
	Study 2	3.983	.051
	Combined	2.707	.104

Statistics of the Levene test. *Abbreviations*: F, F-statistic; df, degrees of freedom; *p*, Significance value.

Table 8 – Test for inequality of variances in classification accuracies for valence independent from stimulus-type (Study1, faces *vs* situations) and during age attribution (Study 2).

ROI	Valence independent of stimulus-type (Study1)		Valence during age attribution (Study 2)	
	F(df)	*p*	F(df)	*p*
DMPFC	.023 (35)	.881	1.391 (50)	.244
MMPFC	.671 (35)	.418	1.605 (50)	.211
rFFA	1.317 (35)	.259	.314 (50)	.578
rmSTS	.277 (35)	.602	2.106 (50)	.153
lpSTC	.476 (35)	.495	.579 (50)	.450

Statistics of the Levene test. *Abbreviations*: F, F-statistic; df, degrees of freedom; *p*, Significance value.

Morey, & Iverson, 2009) (Fig. 4, Table 3). There was also no greater variability in neural measures for the ASD group.

Theoretically, the absence of robust group differences is important because the experiment was a sensitive test of a widespread assumption about the mechanism of social impairment in ASD in general, while addressing valence processing prototypically (Ben Shalom et al., 2006; Celani, Battacchi, & Arcidiacono, 1999; Kuusikko et al., 2009; Tseng et al., 2014). First, all of the stimuli were dynamic videos, and covered basic as well as more complex/social emotions (e.g., angry, surprised, enthusiastic, furious), wide ranges of expressivity (from more subtle to overstated expressions), and a diverse set of individuals (Study 1: 96 characters, Study 2: 20 characters), spanning a wide age range and both genders. Stimuli in Study 1 were subtle and less controlled (with respect to lighting and viewpoint), hence reflecting more naturalistic perceptual content. Stimuli in Study 2 were exaggerated and more controlled (i.e., all with overhead lighting and frontal view). Yet for both sets of stimuli, brain responses to emotional valence were not different in individuals with or without ASD.

Second, the multivariate measure of brain region responses used here is more sensitive than traditional univariate measures (although note that we observed no group difference in average magnitude of response in any region either, see Supplementary material for details). Multivariate

analyses can detect both differences in the average response over whole brain regions, like traditional univariate measures, but are also sensitive to relative differences in the spatial patterns of response. In typically developing adults, many cortical regions show similar overall magnitude, but distinct spatial patterns, of activity in response to positive versus negative facial expressions (Tseng et al., 2016). In the few previous studies using MVPA in participants with ASD, MVPA revealed differences in social information processing in individuals with ASD that were not detected with univariate methods (Coutanche et al., 2011; Gilbert, Meuwese, Towgood, Frith, & Burgess, 2009; Koster-Hale et al., 2013). Whereas those previous studies used MVPA opportunistically on existing data, the current studies were designed to be sensitive to social information processing in each individual participant and across groups.

Third, in addition to finding no difference in basic cortical responses to the valence of facial expressions, each study tested an additional hypothesis about social information processing deficits in ASD, beyond a basic representation of valence. In Study 1, we tested whether in individuals with ASD, responses to emotional valence are relatively cue-specific. That is, individuals with ASD might struggle to see how the same emotional valence could be perceived in facial configurations (down-turned eyes) and event contexts (being left out of a group). Contrary to our hypothesis, individuals with ASD showed the same abstract response to emotional valence, generalizing across faces and events, in the same regions of MPFC.

In Study 2, we tested whether in individuals with ASD, responses to emotional valence are relatively stimulus-driven. That is, individuals with ASD might struggle to flexibly change their response to social stimuli when the task, or relevant feature, shifts. Again contrary to our hypothesis, individuals with ASD showed equally flexible encoding: the expression's valence could be classified from responses in social brain regions when it was task-relevant (i.e., judging emotional intensity in Study 1 or valence in Study 2), but not when emotion was irrelevant to the task (i.e., judging target's age in Study 2). Unexpectedly, we did find a group difference in the flexibility of the control dimension. In an exploratory analysis, social brain regions of control participants, but not those with ASD, could be used to classify the age of the target, only when age was the task-relevant dimension. This finding is intriguing but preliminary, and could be pursued in future confirmatory research. In addition to valence, in Study 2 we also used representational dissimilarity analyses to measure the differentiation of 10 fine-grained emotional expressions, and again found no group differences.

In sum, the results reported here suggest little or no group difference in the extraction of emotion-relevant information from dynamic facial expressions in the social brain regions of adults with ASD. This conclusion seems counterintuitive. How can these results be integrated with the existing literature on social cognitive neuroscience of ASD, and more importantly, with the experience of ASD reported by affected individuals and their families?

Perhaps social impairments in ASD do not reflect difficulty extracting meaning from dynamic faces, considered in isolation. Real world social interaction additionally requires integration of information across modalities and timescales (Rosenblau et al., 2015). Social brain regions may be especially sensitive to the meaning of social cues in larger contexts (Byrge, Dubois, Tyszka, Adolphs, & Kennedy, 2015; Pantelis et al., 2015). For example, TPJ and MPFC responses to movies are strongly disrupted by temporal scrambling (Hasson, Furman, Clark, Dudai, & Davachi, 2008), and robustly enhanced by switching from movies to live interpersonal interactions (Redcay et al., 2010). Individuals with ASD might be disproportionately impaired at the interpretation of social cues in such contexts (Pantelis et al., 2015; Redcay et al., 2010).

Second, fMRI may not be the ideal tool to measure clinically relevant differences in information processing, because of its poor temporal resolution. The synchrony and efficiency of neural processing may be dramatically affected by small differences in phase or frequency of neural activity that are masked by the temporal resolution of the hemodynamic lag. If so, measurements with much higher temporal resolution, such as EEG and MEG may be more appropriate.

Third, disproportionate impairments in social information processing might characterize only a subset of individuals diagnosed with ASD. Aggregated results from hundreds of people suggest that only about 20% of individuals diagnosed with ASD perform below the typical range on one widely used test of social perception, the Reading the Mind in the Eye Test (RMET; Lombardo et al., 2016). Similarly, a recent study compared spontaneous looking to dynamic social and non-social videos in hundreds of toddlers; only 20% of toddlers just diagnosed with Autism showed lower-than-normal-range preferential looking to dynamic social stimuli (Geo-Pref Test; Pierce et al., 2016). Relatedly, Byrge et al. (2015) reported that a small subset of individuals with ASD showed robustly atypical neural responses to a social movie stimulus. The heterogeneity of social deficits in ASD is generally compounded by overall heterogeneity in verbal fluency and IQ, but the two dimensions seem to be distinct – at least in the high-functioning samples tested in most research studies. For example, IQ did not account for the differences in RMET, Geo-Pref or neural responses to social movies described above. In the current study, all participants were verbally fluent, with normal or elevated IQ levels ("high-functioning"), but approximately 20% of the participants performed substantially below the typical range on the Face Puzzle task. Future fMRI studies may need to test hundreds, rather than dozens, of individuals in order to identify the neural correlates of these behavioral differences.

The striking heterogeneity of social and cognitive functioning among individuals diagnosed with ASD constitutes a challenge for comparing results across studies with different samples of individuals. Each group of a few dozen participants may differ from other groups on many untested dimensions. For example, we did not test for co-morbidity of alexithymia, a condition associated with impaired social cognition independent of Autism (Bird & Cook, 2013; Shah, Catmur, & Bird, 2016).

In sum, multiple cortical regions extract generalizable and flexible representations of dynamic emotional facial expressions. However, group differences in the functions of these regions are smaller than cognitive theories of ASD predicted in this domain.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Amy Skerry for design and data collection in Study 1 and Ralph Adolphs for comments on earlier versions of this manuscript. D.K. was in part supported by a Feodor-Lynen Postdoctoral Fellowship of the Alexander von Humboldt society. This work was supported by the NIH grant 4-R01-MH096914-05 and by a grant (project number 6925173) from the Simons Foundation to the Simons Center for the Social Brain at MIT.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.cortex.2018.02.006.

REFERENCES

Adolphs, R. (2006). Perception and emotion: How we recognize facial expressions. *Current Directions in Psychological Science, 15*(5), 222–226. https://doi.org/10.1111/J.1467-8721.2006.00440.X.

Alaerts, K., Woolley, D. G., Steyaert, J., Di Martino, A., Swinnen, S. P., & Wenderoth, N. (2014). Underconnectivity of the superior temporal sulcus predicts emotion recognition deficits in autism. *Social Cognitive and Affective Neuroscience, 9*(10), 1589–1600. https://doi.org/10.1093/scan/nst156.

Association, American Psychological. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.) Washington, DC.

Aviezer, H., Bentin, S., Dudarev, V., & Hassin, R. R. (2011). The automaticity of emotional face-context integration. *Emotion, 11*(6), 1406–1414. https://doi.org/10.1037/a0023578.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17.

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Ben Shalom, D., Mostofsky, S. H., Hazlett, R. L., Goldberg, M. C., Landa, R. J., Faran, Y., et al. (2006). Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism. *Journal of Autism and Developmental Disorders, 36*(3), 395–400. https://doi.org/10.1007/s10803-006-0077-2.

Bird, G., & Cook, R. (2013). Mixed emotions: The contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry, 3*. https://doi.org/10.1038/tp.2013.61. Artn E285.

Boelte, S., Ciaramidaro, A., Schlitt, S., Hainz, D., Kliemann, D., Beyer, A., et al. (2015). Training-induced plasticity of the social brain in autism spectrum disorder. *British Journal of Psychiatry, 207*(2), 149–157. https://doi.org/10.1192/bjp.bp.113.143784.

Byrge, L., Dubois, J., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *Journal of Neuroscience, 35*(14), 5837–5850. https://doi.org/10.1523/JNEUROSCI.5182-14.2015.

Celani, G., Battacchi, M. W., & Arcidiacono, L. (1999). The understanding of the emotional meaning of facial expressions in people with autism. *Journal of Autism and Developmental Disorders, 29*(1), 57–66. https://doi.org/10.1023/A:1025970600181.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1–27.

Chavez, R. S., & Heatherton, T. F. (2015). Representational similarity of social and valence information in the medial pFC. *Journal of Cognitive Neuroscience, 27*(1), 73–82. https://doi.org/10.1162/jocn_a_00697.

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage, 57*(1), 113–123. https://doi.org/10.1016/j.neuroimage.2011.04.016.

Deen, B., & Pelphrey, K. A. (2012). Brain scans need a rethink. *Nature, 491*(7422), S20.

Di Martino, A., Ross, K., Uddin, L. Q., Sklar, A. B., Castellanos, F. X., & Milham, M. P. (2009). Functional brain correlates of social and nonsocial processes in autism spectrum disorders: An activation likelihood estimation meta-analysis. *Biological Psychiatry, 65*(1), 63–74. https://doi.org/10.1016/j.biopsych.2008.09.022.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341. https://doi.org/10.1016/J.Tics.2007.06.010.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040.

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., et al. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS One, 8*(9). https://doi.org/10.1371/journal.pone.0075468. e75468.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage, 25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439–453. https://doi.org/10.1037/a0015251.

Gilbert, S. J., Meuwese, J. D. I., Towgood, K. J., Frith, C. D., & Burgess, P. W. (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: A multi-voxel similarity analysis. *Brain, 132*, 869–878. https://doi.org/10.1093/brain/awn365.

Hadjikhani, N., Zurcher, N. R., Rogier, O., Hippolyte, L., Lemonnier, E., Ruest, T., et al. (2014). Emotional contagion for pain is intact in autism spectrum disorders. *Translational Psychiatry, 4*, e343. https://doi.org/10.1038/tp.2013.113.

Hariri, A. R., Bookheimer, S. Y., & Mazziotta, J. C. (2000). Modulating emotional responses: Effects of a neocortical network on the limbic system. *NeuroReport, 11*(1), 43–48. https://doi.org/10.1097/00001756-200001170-00009.

Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychology Review, 20*(3), 290–322.

Harry, B., Williams, M. A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience, 7*, 692. https://doi.org/10.3389/fnhum.2013.00692.

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron, 57*(3), 452–462. https://doi.org/10.1016/j.neuron.2007.12.009.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736.

Hepach, R., Kliemann, D., Gruneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency - implications for social-cognitive tests and training tools. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00266. Artn 266.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*(5749), 863–866. https://doi.org/10.1126/science.1117593.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage, 60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055.

Kana, R. K., Patriquin, M. A., Black, B. S., Channell, M. M., & Wicker, B. (2016). Altered medial frontal and superior temporal response to implicit processing of emotions in autism. *Autism Research, 9*(1), 55–66. https://doi.org/10.1002/aur.1496.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Kaufman, A. (2004). *Manual for the Kaufman Brief Intelligence Test* (2nd ed.).

Kliemann, D., Jacoby, N., Anzellotti, S., & Saxe, R. R. (2016). Decoding task and stimulus representations in face-responsive cortex. *Cognitive Neuropsychology, 33*(7–8), 362–377. https://doi.org/10.1080/02643294.2016.1256873.

Kliemann, D., Rosenblau, G., Bolte, S., Heekeren, H. R., & Dziobek, I. (2013). Face puzzle-two new video-based tasks for measuring explicit and implicit aspects of facial emotion recognition. *Frontiers in Psychology, 4*, 376. https://doi.org/10.3389/fpsyg.2013.00376.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America, 110*(14), 5648–5653. https://doi.org/10.1073/pnas.1207992110.

Kuusikko, S., Haapsamo, H., Jansson-Verkasalo, E., Hurtig, T., Mattila, M. L., Ebeling, H., et al. (2009). Emotion recognition in children and adolescents with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(6), 938–945. https://doi.org/10.1007/s10803-009-0700-0.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and Statistics: Essays in honor of harold hotelling* (pp. 278–292). Palo Alto: Stanford University Press.

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: An FMRI study. *Journal of Cognitive Neuroscience, 22*(1), 203–211. https://doi.org/10.1162/jocn.2009.21203.

Lombardo, M. V., Lai, M. C., Auyeung, B., Holt, R. J., Allison, C., Smith, P., et al. (2016). Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports, 6*, 35333. https://doi.org/10.1038/srep35333.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule* (2nd ed.). Torrance, CA: Western Psychological Services.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X. R., Just, M., et al. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 57*(1–2), 145–175. https://doi.org/10.1023/B: Mach.0000035475.85309.1b.

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage, 49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047.

Pantelis, P. C., Byrge, L., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. *Social Cognitive and Affective Neuroscience, 10*(10), 1348–1356. https://doi.org/10.1093/scan/nsv021.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience, 30*(30), 10127–10134. https://doi.org/10.1523/JNEUROSCI.2161-10.2010.

Pelphrey, K. A., Morris, J. P., McCarthy, G., & Labar, K. S. (2007). Perception of dynamic changes in facial affect and identity in autism. *Social Cognitive and Affective Neuroscience, 2*(2), 140–149. https://doi.org/10.1093/scan/nsm010.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage, 45*(1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007.

Pierce, K., Marinero, S., Hazin, R., McKenna, B., Barnes, C. C., & Malige, A. (2016). Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. *Biological Psychiatry, 79*(8), 657–666. https://doi.org/10.1016/j.biopsych.2015.03.032.

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage, 56*(4), 2356–2363. https://doi.org/10.1016/j.neuroimage.2011.03.067.

Rahko, J. S., Paakki, J. J., Starck, T. H., Nikkinen, J., Pauls, D. L., Katsyri, J. V., et al. (2012). Valence scaling of dynamic facial expressions is altered in high-functioning subjects with autism spectrum disorders: An fMRI study. *Journal of Autism and Developmental Disorders, 42*(6), 1011–1024. https://doi.org/10.1007/s10803-011-1332-8.

Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D., et al. (2010). Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. *NeuroImage, 50*(4), 1639–1647. https://doi.org/10.1016/j.neuroimage.2010.01.052.

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *Journal of Autism and Developmental Disorders, 45*(4), 953–965. https://doi.org/10.1007/s10803-014-2249-9.

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Russel, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V., & Todorov, A. (2010). Graded representations of emotional expressions in the left superior temporal sulcus. *Frontiers in Systems Neuroscience, 4*, 6. https://doi.org/10.3389/fnsys.2010.00006.

Scherf, K. S., Elbich, D., Minshew, N., & Behrmann, M. (2015). Individual differences in symptom severity and behavior

predict neural activation during face processing in adolescents with autism. *NeuroImage Clinical, 7*, 53–67. https://doi.org/10.1016/j.nicl.2014.11.003.

Schneider, K., Regenbogen, C., Pauly, K. D., Gossen, A., Schneider, D. A., Mevissen, L., et al. (2013). Evidence for gender-specific endophenotypes in high-functioning autism spectrum disorder during empathy. *Autism Research, 6*(6), 506–521. https://doi.org/10.1002/aur.1310.

Seung, H. S., & Sompolinsky, H. (1993). Simple-models for reading neuronal population codes. *Proceedings of the National Academy of Sciences of the United States of America, 90*(22), 10749–10753. https://doi.org/10.1073/Pnas.90.22.10749.

Shah, P., Catmur, C., & Bird, G. (2016). Emotional decision-making in autism spectrum disorder: The roles of interoception and alexithymia. *Molecular Autism, 7*. https://doi.org/10.1186/s13229-016-0104-x. Artn 43.

Shamir, M., & Sompolinsky, H. (2006). Implications of neuronal diversity on population coding. *Neural Computation, 18*(8), 1951–1986. https://doi.org/10.1162/Neco.2006.18.8.1951.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *Journal of Neuroscience, 34*(48), 15997–16008. https://doi.org/10.1523/JNEUROSCI.1676-14.2014.

Snedecor, G., & Cochran, W. (1989). *Statistical methods (Vol. 8th revision)*. Iowa State University Press.

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., et al. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 44*(6), 1332–1346. https://doi.org/10.1007/s10803-013-1993-6.

Tseng, A., Wang, Z. S., Huo, Y. K., Goh, S., Russell, J. A., & Peterson, B. S. (2016). Differences in neural activity when processing emotional arousal and valence in autism spectrum disorders. *Human Brain Mapping, 37*(2), 443–461. https://doi.org/10.1002/hbm.23041.

Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: a formal meta-analysis. *Journal of Autism and Developmental Disorders, 43*(7), 1517–1526.

Van Dijk, K. R., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage, 59*(1), 431–438. https://doi.org/10.1016/j.neuroimage.2011.07.044.

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., et al. (2015). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron, 88*(3), 604–616. https://doi.org/10.1016/j.neuron.2015.09.042.

Weisberg, J., Milleville, S. C., Kenworthy, L., Wallace, G. L., Gotts, S. J., Beauchamp, M. S., et al. (2014). Social perception in autism spectrum disorders: Impaired category selectivity for dynamic but not static images in ventral temporal cortex. *Cerebral Cortex, 24*(1), 37–48. https://doi.org/10.1093/cercor/bhs276.

Whitfield-Gabrieli, S., Nieto-Castanon, A., & Ghosh, S. (2011). *Artifact Detection Tools (ART)* (Vol. 7, p. 11). Cambridge, MA.

Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., & Fischl, B. (2014). Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage, 88*, 79–90. https://doi.org/10.1016/j.neuroimage.2013.11.027.

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences, 1167*, 16–30. https://doi.org/10.1111/j.1749-6632.2009.04601.x.

Zhang, J., Meeson, A., Welchman, A. E., & Kourtzi, Z. (2010). Learning alters the tuning of functional magnetic resonance imaging patterns for visual forms. *Journal of Neuroscience, 30*(42), 14127–14133. https://doi.org/10.1523/JNEUROSCI.2204-10.2010.