NeuroImage 245 (2021) 118741

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing

Haemy Lee Masson*, Leyla Isik

Department of Cognitive Science, Johns Hopkins University, MD 21218, United States

ARTICLE INFO

Keywords:
Social interaction perception
Theory of mind
Superior temporal sulcus
Medial prefrontal cortex
Encoding model
Variance partitioning

ABSTRACT

Recognizing others' social interactions is a crucial human ability. Using simple stimuli, previous studies have shown that social interactions are selectively processed in the superior temporal sulcus (STS), but prior work with movies has suggested that social interactions are processed in the medial prefrontal cortex (mPFC), part of the theory of mind network. It remains unknown to what extent social interaction selectivity is observed in real world stimuli when controlling for other covarying perceptual and social information, such as faces, voices, and theory of mind. The current study utilizes a functional magnetic resonance imaging (fMRI) movie paradigm and advanced machine learning methods to uncover the brain mechanisms uniquely underlying naturalistic social interaction perception. We analyzed two publicly available fMRI datasets, collected while both male and female human participants ($n = 17$ and 18) watched two different commercial movies in the MRI scanner. By performing voxel-wise encoding and variance partitioning analyses, we found that broad social-affective features predict neural responses in social brain regions, including the STS and mPFC. However, only the STS showed robust and unique selectivity specifically to social interactions, independent from other covarying features. This selectivity was observed across two separate fMRI datasets. These findings suggest that naturalistic social interaction perception recruits dedicated neural circuity in the STS, separate from the theory of mind network, and is a critical dimension of human social understanding.

1. Introduction

Humans learn a great deal about the social world by observing others' social interactions, defined here as action or communication between two or more people that are directed at and contingent upon each other. This social observation helps us form impressions of interacting people (e.g., whether they are prosocial or antisocial), their social relationships, and relative social status (Quadflieg and Koldewyn, 2017). Our ability to extract social information about interacting people appears to emerge in infancy (Hamlin and Wynn, 2011) and is preserved across species (Cheney and Seyfarth, 1986), underscoring the importance of this observational ability.

Previous studies have reported that the posterior superior temporal sulcus (STS) (Isik et al., 2017; Walbrin et al., 2018; Walbrin and Koldewyn, 2019) is selectively involved in recognizing others' social interactions, in a manner that is functionally distinct from other related social functions, including detecting faces, animacy, other agents' goals and theory of mind. These prior studies, however, have used simple stimuli, such as point-light figures or staged images that are largely devoid of other social information. It thus remains unknown how the human brain processes social interactions in real world scenes, where the interactions are highly confounded with other perceptual and social properties, including motion, faces, and theory of mind. The goal of the current study is to identify the brain mechanisms underlying real-world social interaction perception by adopting a naturalistic movie viewing paradigm.

A growing body of evidence in visual neuroscience suggests that naturalistic neuroimaging paradigms better uncover neural representations of objects, faces, scenes, and actions (Haxby et al., 2020a; Nishimoto et al., 2011; Wen et al., 2018). Natural movies elicit stronger and more reliable brain responses than traditional experiments (Hasson et al., 2010; va der Meer et al., 2020; Sonkusare et al., 2019). In contrast, social neuroscience has not yet fully exploited these methods, although improving ecological validity in (social) cognitive neuroscience is considered an urgent challenge (Nastase et al., 2020; Redcay and Moraczewski, 2020). While several recent social neuroscience studies have adopted this paradigm to investigate theory of mind (Jacoby et al., 2016; Richardson, 2019; Richardson et al., 2018) and affective experience (Chen et al., 2020), only one study has investigated how the human brain processes social interactions during movie viewing (Wagner et al., 2016). This study revealed the involvement of the mPFC in social interaction perception, conjecturing that others' social interactions invite a viewer to infer others' personality and intention. However, it is un-

* Corresponding author.
E-mail address: haemy.leemasson@jhu.edu (H. Lee Masson).

https://doi.org/10.1016/j.neuroimage.2021.118741.
Received 8 July 2021; Received in revised form 15 September 2021; Accepted 16 November 2021
Available online 17 November 2021.

clear how the presence of other social features in the movie, particularly theory of mind, influenced these findings, which contribute to a longstanding debate about the distinct vs. overlapping roles of social perception and mentalization. Moreover, these results are in conflict with prior studies that used animated videos (Lahnakoski et al., 2012) and the above-mentioned studies with controlled stimuli, which identified the STS in social interaction perception. As a result, the neural underpinnings of social interaction perception, particularly in naturalistic contexts, remain unknown.

Here, we implemented computer vision techniques, machine-learning-based encoding model analyses, and variance partitioning to identify the unique contribution of social interactions to responses in the human brain. To improve ecological validity and replicability, we answered this novel question by applying the same analyses on two different fMRI datasets (Aliko et al., 2020; Chen et al., 2017) collected by different labs showing movies from different genres (crime vs. romance) to groups of participants living in different countries (US vs. UK) on different MRI scanners (3T vs. 1.5T). Our findings reveal that social-affective features, consisting of an agent speaking, social interactions, theory of mind, perceived valence, and arousal, independently contribute to predicting brain responses in the STS and the mPFC. However, we find that the STS, but not mPFC, shows unique selectivity for others' social interactions in particular, independent of other co-varying features.

2. Material and methods

2.1. fMRI data sources

We analyzed two publicly available fMRI datasets from two different studies. In the first study (Chen et al., 2017), 17 participants (male = 10) watched the first episode of the Sherlock BBC TV series (duration ~ 45 min) in the scanner. In the second study (Aliko et al., 2020), 86 participants watched 10 different movies from 10 genres. We selected 20 participants who watched a commercial movie, 500 Days of Summer (duration ~ 90 min). We excluded two participants from the second study as one participant (ID 14 in the original study) was scanned with a different head coil, and another participant (ID 16 in the original study) was offered glasses only after the first run, leaving 18 participants (male = 9) for our second movie analyses. The studies were approved by the Princeton University Institutional Review Board and Ethics Committee of University College London, respectively. All subjects provided their written informed consent before the experiment.

2.2. fMRI data acquisition and preprocessing

In the first study (Sherlock), fMRI data were obtained on a 3T Siemens Skyra scanner with a 20-channel head coil. Whole-brain images were acquired (27 slices; voxel size = $4 \times 3 \times 3$ mm^3) with an echo-planar (EPI) T2∗-weighted sequence with the following acquisition parameters: repetition time (TR) = 1500 ms, echo time (TE) = 28 ms, flip angle (FA) = 64°, field of view (FOV) = 192×192 mm^2. In the second study (Summer), fMRI data were obtained on a 1.5T Siemens MAGNETOM Avanto with a 32-channel head coil. Whole-brain images were acquired (40 slices; voxel size = $3.2 \times 3.2 \times 3.2$ mm^3) with a multiband EPI sequence with the following acquisition parameters: multiband factor = 4, no in-plane acceleration, TR = 1000 ms, TE = 54.8 ms, FA = 75 °

We obtained preprocessed fMRI data from the authors of each original study. In the original Sherlock study, preprocessing steps included slice-timing correction, motion correction, linear detrending, temporal high-pass filtering (140 s cut off), spatial normalization to a Montreal Neurological Institute (MNI) space with a re-sampling size of $3 \times 3 \times 3$ mm^3, and spatial smoothing with a 6-mm full width at half maximum (FWHM) Gaussian kernel. fMRI data were also shifted by 4.5 s (3 TRs) from the stimulus onset to account for the hemodynamic delay. Lastly, timeseries blood oxygenation level dependent (BOLD) signals were z-score standardized.

In the Summer study, authors performed slice-timing correction, despiking, motion correction, spatial normalization to MNI space with a re-sampling size of $3 \times 3 \times 3$ mm^3, and spatial smoothing with a 6-mm FWHM Gaussian kernel. Timeseries BOLD signals were scaled to 0~1 and detrended based on run lengths varying across participants and runs, head-motion parameters, and averaged BOLD signals in white matter and cerebrospinal fluid regions. Timing correction was also applied to align the fMRI timeseries and the movie. Detailed information on how movie was paused and restarted for each run and timing correction related this issue can be found in the original study (Aliko et al., 2020). In addition, similar to the Sherlock study, we shifted Summer fMRI by 4 s (4 TRs) from the stimulus onset to account for the hemodynamic delay.

2.3. Movie analysis and annotations

To examine how BOLD responses relate to each stimulus feature, we fit a linear regression model where voxel-wise responses are predicted based on a linear combination of stimulus features (Fig. 1A). To create a stimulus feature space, we annotated the movies with a mix of fully automatized approaches and human labeling. We first split the full-length Sherlock episode and the Summer movie into 1.5 and 3 s segments, respectively. We excluded the introductory video clip, a short-animated movie shown before the Sherlock episode in the original study, from the analysis. For the Summer movie, the opening and ending credits were excluded from the analysis. fMRI volumes matching these scenes were truncated and excluded from further analysis. A total of 1921 and 1722 video segments were generated from the Sherlock and Summer stimuli, respectively.

Using MATLAB (R2020a, The Mathworks, Natick, MA) built-in functions, we extracted low-level visual features, namely hue, saturation, pixel values (HSV) and motion energy, and auditory features, namely amplitude and pitch. Specifically, we computed HSV ('rgb2hsv') and motion energy ('opticalFlowFarneback') in each pixel for all frames (640 × 360 pixels and 38 frames for Sherlock, 720 × 576 pixels and 75 frames for Summer) and averaged these values over pixels and frames. In both movies, the majority of motion comes from humans moving so we believe the motion energy to also be a good proxy for biological motion. We also extracted high-level visual features from a deep convolutional neural network model (DNN, see below section). For audio amplitude ('audioread') and pitch ('pitch'), we averaged values over the audio samples and channels (66,150 samples and two channels for Sherlock and 132,300 samples and two channels for Summer). This computation allowed us to obtain one value per feature for each video segment (the duration of 1 TR).

The authors of the Summer study (Aliko et al., 2020) used the 'Amazon Rekognition' service (https://aws.amazon.com/rekognition/) to obtain faces annotations. Based on their annotations, we extracted the presence or absence of faces in each segment. We implemented the same analysis pipeline ('start_face_detection' and 'get_face_detection' functions from the Amazon Rekognition) to the Sherlock video segments. We used the binary label (presence of a face(s) = 1, absence of a face = 0) for each video segment with the confidence level 99% as a threshold.

We also included some of the publicly available annotations, made by human raters in the original Sherlock study (Chen et al., 2017) – whether the location of the scene is indoor or outdoor (indoor = 1, outdoor = 0), whether or not music plays in the background (presence of music = 1, absence of music = 0), and whether or not there are written words on the screen (presence of written words = 1, absence of written words = 0). In the current study, two human raters made the same annotations for the Summer video segments.

Two human raters indicated whether or not a video segment contains social interactions (defined as actions or communication between two or more individuals that are directed at and contingent upon each

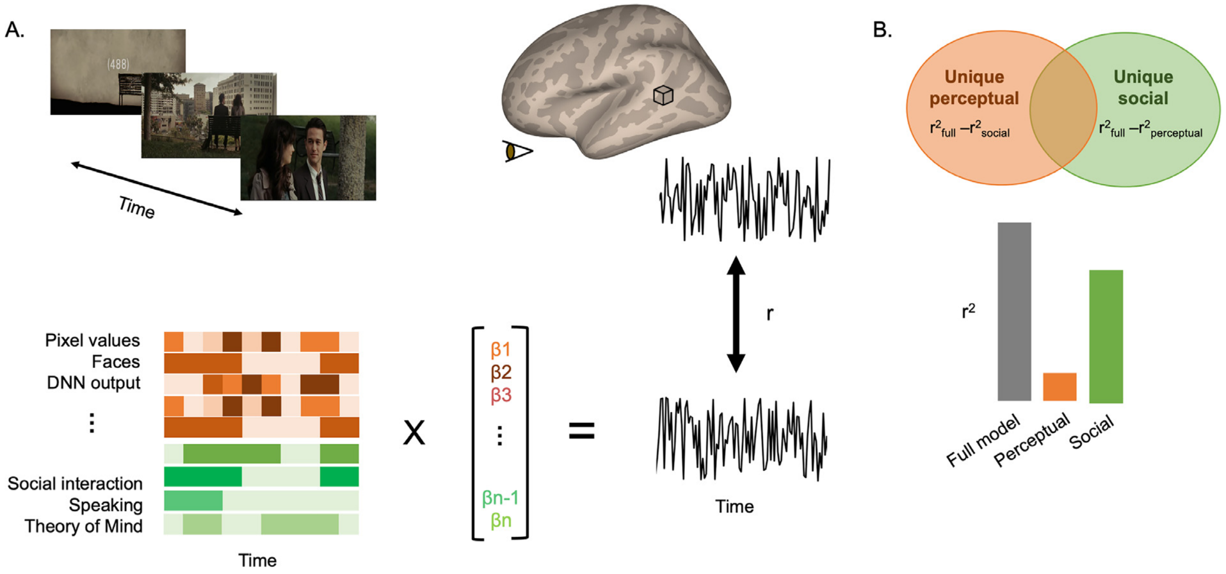


Fig. 1. A. Encoding model overview. We labeled perceptual and social-affective features from a movie that participants viewed in the MRI scanner. The features include hue, saturation, value for each pixel, motion energy, the presence/absence of written words, indoor vs. outdoor scenes, the presence/absence of faces, audio amplitude, pitch, the presence/absence of music, the presence/absence of social interactions, the presence/absence of an agent speaking, the presence/absence of agent talking about others' mental states (theory of mind), perceived valence, and arousal of the scene for both movies, and additionally the social touch feature for the Summer movie. During model training, we learned a set of beta weights linking the features to fMRI BOLD responses over time. We then predicted the response to held-out movie data by multiplying the movie feature vectors by their corresponding beta weights, and correlated these predictions with the actual responses extracted while participants viewed the movie. As a result, a prediction performance score (r) of the model is assigned to each voxel. B. Variance partitioning analysis overview. Prediction performance r values from the full, perceptual, and social-affective models are used to calculate unique variance explained by the perceptual or social-affective model for each voxel.

Table 1
Relative frequency of feature occurrence in each movie. Only binary features are included. Note that multiple annotators used binary scores to rate these features, and their ratings were averaged, yielding values ranging from 0 to 1. In the case of decimals, we round them to the nearest whole number (0 or 1) when counting the occurrence.

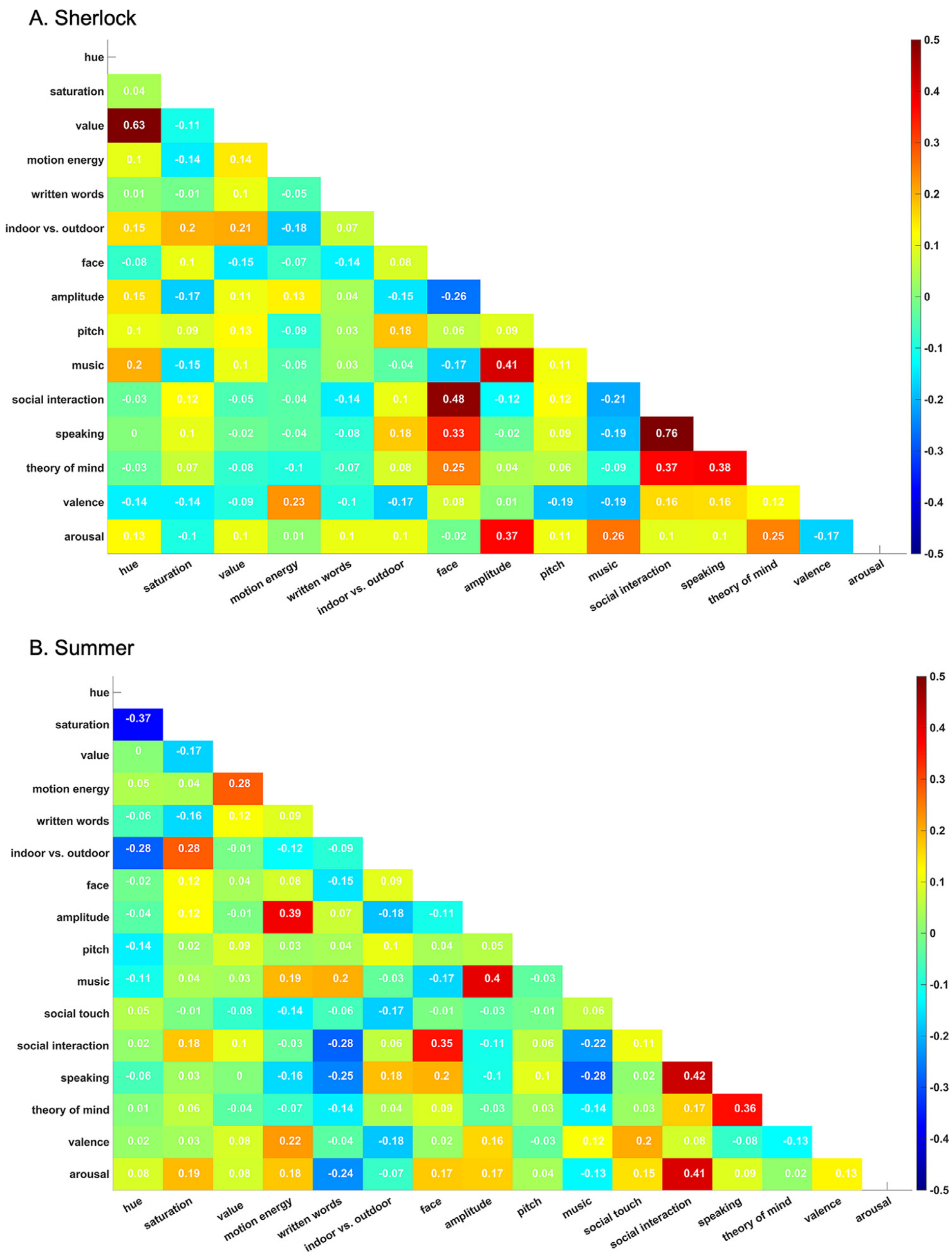
	Written Text	Music	Indoor	Face	Speaking	Social Interaction	Theory of Mind	Touch
Sherlock	4%	63%	71%	85%	65%	75%	32%	not included
Summer	19%	66%	66%	92%	58%	77%	20%	11%

other, presence of social interactions = 1, absence = 0). These labels were highly consistent across our two raters ($r = 0.92$ and $r = 0.86$ for Sherlock and Summer, respectively). The raters also rated additional social features, including whether or not a person speaks in the scene (yes = 1, no = 0), and whether or not a person infers mental states of others (yes = 1, no = 0). The annotation of the theory of mind feature is based on whether a movie character is inferring other characters' thoughts and emotions in each scene (an example in Sherlock – Ms. Patterson is seated at a table at a press conference, reading her statement saying: "He loved his family and his work"; an example in Summer – The main character seen sitting next to other colleagues at a meeting and a narrator says: "The boy, Tom Hansen of Margate of New Jersey, grew up believing that he'd never truly be happy until the day he met the one."). We selected this criterion to be as objective as possible, and to avoid raters guessing about whether the subjects were engaged in mentalization. This type of second-order theory of mind task activates the theory of mind network (Tholen et al., 2020). As in prior studies (Saxe and Powell, 2006), the description of a character's appearance (e.g., she is tall and thin) or bodily sensation (e.g., she had been sick for three days) were not considered theory of mind. For Summer, we additionally included a touch feature – whether or not a person makes physical contact with another person (social touch = 1), him(her)self or an object (nonsocial touch = –1), or not (absence of touch = 0). We were unable to include the touch feature in the Sherlock data as there were less than 10 video segments displaying social touch.

Considering that the time-scale of high-level cognitive events is longer than a low-level perceptual event (Baldassano et al., 2017), we merged each pair of consecutive Sherlock 1.5 s length segments into one for annotations of social features, resulting in 3 s length segments for both studies. For binary features, the relative frequency of feature occurrence is described in Table 1.

Lastly, we added valence and arousal ratings, offered by the authors of another Sherlock fMRI study (Kim et al., 2020). In their study, 4.5 s video segments, parsed from the full-length Sherlock episode, were rated by 113 participants with a 1–9 Likert scale. 35 participants rated all video segments, and the remaining participants rated only a quarter, yielding about 55 ratings per video segment. Group mean valence and arousal ratings were used in the current study. For Summer segments, four human raters judged how pleasant and arousing the scene is using a 1–5 Likert scale. Despite the small number of raters, inter-rater consistencies are relatively high (valence Spearman r (r_s) = 0.62, arousal $r_s = 0.35$), compared to that (valence $r = 0.30$) reported in the Sherlock study (Kim et al., 2020).

In summary, perceptual features consist of components from the DNN fifth layer (see below), HSV, motion energy, faces, indoor/outdoor, written words, amplitude, pitch, and music. Social-affective features consist of social interactions, speaking, theory of mind, valence, and arousal for both experiments, and additionally the touch feature for the Summer data. Fig 2 illustrates correlations across one-dimensional features. Although these feature spaces are moderately correlated, we be-


A. Sherlock

B. Summer

Fig. 2. The pairwise rank correlations between features in Sherlock (A) and 500 days of Summer (B) movie. Cells in the matrix show the correlation coefficients between features, indicated by Spearman r values and color coding (red: positive correlation; blue: negative correlation). DNN features are not included in this correlational analysis as they are high-dimensional, unlike other features (1 vector per feature) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

lieve them to capture important non-overlapping information. We ran principal component analysis (PCA) on the five and six social-affective features labeled in the Sherlock and Summer movies, respectively. According to this analysis five (Sherlock) and six (Summer) principal components are needed to account for 99% of the feature variance, suggesting that the dimensionality of these feature spaces cannot be reduced without losing explanatory power.

2.4. Deep convolutional neural network

We used a deep convolutional neural network to extract visual features from both stimuli. We included DNN features as they have recently been shown to explain a great deal of variance throughout both early and later stages of visual cortex (Wen et al., 2018). Specifically, PyTorch (version 1.4.0) implementation of AlexNet (Krizhevsky et al., 2012), pre-trained on the ImageNet dataset (Russakovsky et al., 2015), was used. AlexNet consists of eight layers with five convolutional layers and three fully connected layers. AlexNet in PyTorch adopted 64, 192, 384, 256, and 256 kernels from the first through fifth layers instead of 96, 256, 384, 384, and 256 reported in the original study (Krizhevsky et al., 2012). The size of the kernels is the same as original AlexNet – 11 × 11, 5 × 5, 3 × 3, 3 × 3, 3 × 3 from the 1st to 5th layers, respectively.

The first frame of each video segment was, cropped, normalized, and fed into the first layer of AlexNet, with the image input size 3 × 224 × 224. The first frame was chosen as it matches the onset of each TR and visual scenes do not change dramatically every 1.5 or 3 s in the movies. The output of the previous layer served as the input of the following layer. Including all DNN layers in the encoding model generates more features than samples, which risks overfitting, and is computationally expensive. We captured visual features of each video segment by taking the activations of all units from the fifth layer right before the rectified linear activation function and the max-pooling layer during the forward pass. As activations from the fifth layer are highly correlated to those of early ($r > 0.75$ with second layer) and late layers ($r > 0.5$ seventh layer), we assume that the fifth layer captures low, mid, and high-level visual information shown in the movies without risking overfitting. The output size of the fifth layer is 256 × 13 × 13, which was flattened to an array with 43,264 elements. Unlike other features (1 × the number of video segments), the features from the DNN layer are high dimensional (i.e., 43,264 × the number of video segments). We reduced the dimensionality without losing crucial information by applying PCA to the DNN activations. Components first through Nth were selected until the amount of explained variance reached 70%. 147 and 150 components were added to the feature matrix of the fifth layer for Sherlock and Summer segments, respectively.

2.5. Inter-subject brain correlation

Inter-subject correlation (ISC) analysis was implemented to create a brain mask containing voxels showing shared stimuli-evoked responses across participants. ISC is a well-validated fMRI method that identifies voxels with reliable neural responses across time while rejecting idiosyncratic and noisy voxels (Nastase et al., 2019). Using the brain imaging analysis kit's (BrainIAK, version 0.1.0) built-in functions ('isc', 'permutation_isc'), we measured ISC with a leave-one-subject-out approach. In other words, for every voxel, time-series BOLD responses for all but one subject were averaged and correlated with that of the remaining subject. We repeated this process for every participant. Fisher-Z transformed correlation values were averaged across folds. Encoding model analysis was masked to include only voxels with ISC value > 0.25 in the case of the Sherlock dataset, based on the previous permutation results ($P_{\text{the false discovery rate (FDR)}} < 0.001$) (Baldassano et al., 2017). For the Summer dataset, we ran a sign permutation test (5000 iterations) across participants, as implemented in BrainIAK, and created a mask consisting of voxels passing the statistical threshold, $P_{\text{FDR}} < 0.005$ after the multiple comparison correction with the FDR. In other words, subjects' ISC values were randomly multiplied with +1 or −1 for 5000 times, which resulted in empirical null distribution of ISC values for each voxel. From this distribution, two-tailed P values were calculated and adjusted with FDR correction. We used a less stringent threshold ($P_{\text{FDR}} < 0.005$ as opposed to $P_{\text{FDR}} < 0.001$) on Summer ISC to have a mask with the similar number of voxels with that of Sherlock (25,468 and 22,044 voxels, respectively). Voxels outside of the mask or whose BOLD responses did not change over time were excluded from the further encoding analyze.

2.6. Voxel-wise encoding modeling

We used an encoding model approach to predict the voxel-wise BOLD responses evoked during natural movie viewing. For each voxel, BOLD responses were modeled as a linear combination of various feature spaces (Fig 1A). Each feature was normalized over the course of the movie. Specifically, banded ridge regression was implemented to estimate the beta weights of stimulus features in the nested cross-validation scheme. Unlike classical L2-regularized ridge regression, banded ridge regression does not assume all feature spaces require the same level of regularization and allows more than one ridge penalty in the prediction model (Nunez-Elizalde et al., 2019). Thus, banded ridge regression has advantages over ordinary least squares regression and classical ridge regression. It minimizes overfitting to noise during training, and it is a preferred method when the feature spaces are high-dimensional or suffer from (super) collinearity, ultimately improving the prediction performance (Nunez-Elizalde et al., 2019). In particular, we used two different ridge penalties: one for the high-dimensional DNN features and a second for all other single-dimensional (per video segment) features. We did this to avoid an overweighting of the DNN features that may occur with one shared ridge penalty.

Data were first split into 10 folds. Among them, nine folds were used to estimate beta weights and optimize the regularization parameter (range 0.1–10,000) λ_1 and λ_2. Optimal λ values were selected per voxel via inner loop 5-fold cross-validation. In other words, training data from nine folds were again split into five folds. Four folds were used for the model estimation with various λ values, and unseen data from the remaining fold in the inner loop were used for selecting the λ values that, on average, yielded the smallest mismatch between predicted and actual BOLD responses. After the model estimation and regularization parameter optimization, unseen data from the remaining tenth fold in the outer loop were used for evaluating the performance of the model. We measured model performance by computing the correlation between predicted and actual BOLD responses. For each voxel separately, this process was repeated 10 times, and the performance of the model was averaged over the repetitions.

Specifically, our encoding model was built as follows. The first step is to normalize the dependent (Y) and independent (X) variables by centering and scaling them to have mean 0 and standard deviation 1. Let Y be an array of size T_r (number of total fMRI volumes from the training set) consisting of the zero-centered BOLD signal amplitudes after the normalization. Let X_1 and X_2 be a $T_r \times F_1$ (number of DNN features) and $T_r \times F_2$ (number of additional features) matrix, respectively. Y in the banded ridge regression is then:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{1}$$

where β_1 and β_2 are a F_1 or F_2 sized array containing the beta weights for each feature.

$$\hat{\beta}_{\text{banded ridge}} = \begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \left(\begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} + \begin{bmatrix} \lambda_1^2 I_{F1} & 0 \\ 0 & \lambda_2^2 I_{F2} \end{bmatrix}\right)^{-1} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} Y \tag{2}$$

Note that when $\lambda_1 = \lambda_2$, a regression model becomes classical ridge regression, and when λ_1, λ_2 are 0, a model becomes ordinary least squares regression. Having an option of $\lambda_1 = \lambda_2$ or $\lambda_1 = \lambda_2 = 0$, banded ridge regression performs at least as good as classical ridge or ordinary least squares regression. The size of the diagonal matrix containing λ_1 and λ_2 is $F \times F$, where $F = [\ F_1 \quad F_2\ ]$. The diagonal entries consist of repeats of λ_1, as many as the size of array F_1, and repeats of λ_2 for the remaining entries. As described earlier, the optimal λ values were selected in the inner loop.

Using estimated beta weights, $\hat{\beta}_{\text{banded ridge}}$, unseen BOLD responses were predicted as:

$$Y_{\text{predicted}} = X\hat{\beta}_{banded\ ridge} + \varepsilon, \tag{3}$$

where $Y_{\text{predicted}}$ is an array of size T_e (number of total fMRI volumes from the testing set) consisting of estimated BOLD signals. Let $X = [\ X_1 \quad X_2\ ]$, where X_1 and X_2 are a $T_e \times F_1$ and $T_e \times F_2$ matrix, respectively. The last step is to calculate correlation coefficient value between the predicted and actual BOLD signal to evaluate the model performance. A high r value can be interpreted as high prediction performance.

Three separate encoding models were built to evaluate which voxel could be accurately predicted by the linear combination of all, perceptual, or social-affective features.

(1) A full model includes all features listed in the sub-section (movie analysis and annotations). In this model, X_1 is composed of high-dimensional DNN components and X_2 is composed of the rest. Beta weights estimated from this model were used in preference mapping analysis described below.
(2) A perceptual model includes all visual and auditory features. X_1 is composed of DNN components and X_2 is composed of the rest of visual and auditory features.
(3) A social-affective model includes all social-affective features. In the social-affective model, the feature space is low-dimensional, so classic ridge regression (where $\lambda_1 = \lambda_2$) was used. Thus, X is composed of all social-affective features.

We calculated the prediction performance map across participants for every model. Random-effect group-level analyses were conducted with a nonparametric permutation test to identify voxels showing significantly above chance prediction performance. Before the permutation test, we removed voxels that were not shared across participants, which resulted in the final number of voxels being slightly less than the total number of voxels contained in the mask (21,649 voxels for Sherlock; 24,477 voxels for Summer). Similar to previous studies (Cichy et al., 2017; Hebart et al., 2018), and the above-mentioned ISC analysis, we conducted a sign permutation test (5000 iterations). From the empirical null distribution of a prediction performance, one-tailed P values were calculated and adjusted with FDR correction. Group averaged prediction performance maps of each model were thresholded at $P_{FDR} < 0.05$ and plotted on the cortical surface. Note that we report original r values as no noise ceiling (upper limit of model performance) normalization was performed.

Two additional encoding models were built to evaluate the unique effects of two main social features of interest, the presence of a social interaction and mentalization, on the BOLD response prediction.

(1) One model includes all features except the presence of a social interaction
(2) The final model includes all features except the mentalization feature.

Prediction performances obtained from these five models were used in variance partitioning analyses.

2.7. Preference mapping

To gain a comprehensive understanding of the relative contribution of different features to activity in each voxel, we performed preference mapping analysis to visualize the stimulus features that were best and the next best at explaining each voxel's activation. The primary advantage of this method is that it allows us to consider all features in a single analysis. Beta weights generated from a full model were used to predict the withheld BOLD responses for each voxel via a cross-validation as described above. During the testing session, we used the beta weight(s) of a feature of interest (e.g., the amplitude of audio) to estimate unseen BOLD responses while assigning 0 values to beta weights of other features. This method is superior to having a separate model for each feature when evaluating the relative contribution of each individual feature in predicting voxel activations (Nunez-Elizalde et al., 2019). Moreover, this method is suitable for examining the prediction performance for high-dimensional feature spaces, such as DNN units, which produces hundreds of beta weights. We repeated this procedure for every feature and participant. In the end, the winning feature that yielded the highest group averaged prediction performance and the next highest were selected for each voxel. We colored every voxel to reflect which feature predicts BOLD responses of that voxel the best and the second best. The first and second preference maps were plotted on the cortical surface.

2.8. Variance partitioning

While a preference mapping analysis measures the relative contribution of each stimulus feature in predicting held out BOLD responses, a variance partitioning analysis determines the unique contribution of each feature or group of features to this prediction. Unlike preference mapping it is not feasible to compare all features individually in this analysis, thus we used this to better understand the unique contributions of two broad groupings of features (perceptual and social-affective) and two important social features (the presence of a social interaction and theory of mind). To this end, we compared the amount of variance explained by a perceptual or social-affective model with that of a full model consisting of all features (Fig. 1B), as well as the unique contribution of the presence of a social interaction and mentalization, in predicting the BOLD responses. The amount of unique variance explained by a model/feature of interest was calculated as:

$$U_{voi} = r_x^2 - r_{x-voi}^2$$

X reflects all features listed in the sub-section (2.3 movie analysis and annotations). VOI reflects a variable of interest (e.g., all social-affective features or a mentalization feature). r^2 is the squared prediction performance value obtained from the encoding model, which can be interpreted as the variance explained by a model consisting of either all features X or all features except the variable of interest X – VOI. U_{voi} is the amount of unique variance explained by VOI. We computed U_{voi} for every participant and voxel. Thus, U_{voi} is an N × V sized matrix where N reflects the total number of participants and V reflects the total number of voxels included in the encoding model analysis for each dataset (i.e., N = 17 and Voxels = 21,649 for Sherlock; N = 18 and Voxels = 24,477 for Summer). The higher the U value is, the more the variance is uniquely explained by a variable of interest. Like voxel-wise encoding, we applied the sign permutation test (5000 iterations) for the statistical inference. Group averaged maps, resulting from U values, were thresholded at $P_{FDR} < 0.05$ and plotted on the cortical surface. Most brain areas were labeled with automated anatomical labeling atlas (Tzourio-Mazoyer et al., 2002). The location of STS and temporoparietal junction (TPJ) were compared with the templates created from the previous studies (Deen et al., 2015; Mars et al., 2012).

3. Results

3.1. Perceptual and social-affective features accurately predict voxel-wise activation throughout the brain

Two sets of subjects (N = 17 and N = 18) viewed two different movies (the first episode of the Sherlock BBC TV series and 500 Days of Summer) while their BOLD responses were recorded in fMRI. We extracted a range of perceptual and social features from each movie using a combination of automatic methods and human labeling. The perceptual features consisted of low-level sensory features including HSV, motion energy, audio amplitude and pitch, as well as higher-level perceptual features including the presence of faces, indoor vs. outdoor scenes, written words, the presence of music and features from the fifth (final) convolutional layer of DNN (see 2.3 Movie analysis and annotations). Social-affective features consisted of social interactions, an agent speaking, theory of mind, perceived valence, and arousal for both studies, and additionally the touch feature for the Summer data. We performed voxel-wise encoding analyses to learn the relationship between these features and fMRI movie data, and then predicted held out BOLD responses using the features and beta values learned in training (Fig 1A). We focused our analyses on voxels with shared stimuli-evoked responses across participants as measured by ISC (see 2.5 Inter-subject brain correlation).

In both fMRI datasets, the performance of the full model, consisting of all perceptual and social-affective features, was significantly better than chance at predicting voxel-wise responses throughout the brain. The full model significantly predicted BOLD responses in 100% and 99.99% of voxels inside of the ISC mask for the Sherlock (range of prediction performance (r) = 0.08 ~ 0.51, mean performance = 0.25, standard deviation (std) = 0.07) and Summer fMRI data (range = 0 ~ 0.43, mean = 0.17, std = 0.06), respectively. Note that prediction performances differ across these voxels as reflected in Fig. 3. The highest model performance was observed in the left STS in both studies (Peak MNI coordinates X, Y, Z = −63, −24, −3 for Sherlock; X, Y, Z = −66, −18, 1 for Summer). High prediction performance was also found in the right STS (highest performance in right STS = 0.48 for Sherlock (X, Y, Z = 51, −33, 0); 0.43 for Summer (X, Y, Z = 67, −21, −3)).

The perceptual model, consisting of the visual and auditory features listed above, significantly explained BOLD responses in 100% and 99.99% of voxels inside of the ISC mask in the Sherlock (range of prediction performance = 0.05–0.38, mean = 0.21, std = 0.06) and Summer data (range = 0–0.35, mean = 0.14, std = 0.05), respectively. The highest model performance was observed in the visual cortex in both experiments – the right fusiform gyrus (X, Y, Z = 27, −69, −12) in Sherlock and the early visual cortex (X, Y, Z = 13, −87, 7) in Summer (Fig. 4).

A social-affective model, consisting of the social-affective features listed above (an agent speaking, social interactions, theory of mind, per-

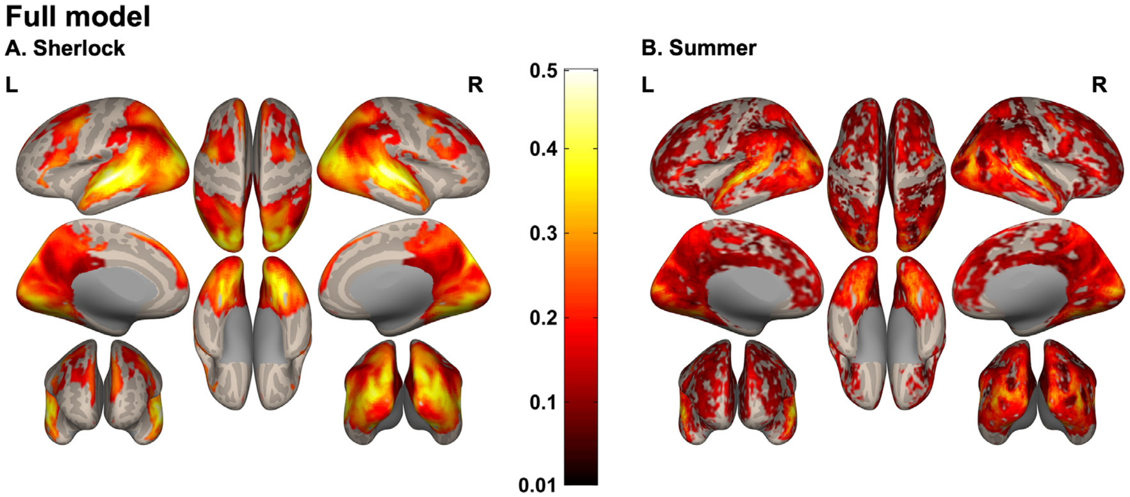


Fig. 3. The full model prediction performance observed in Sherlock (A) and Summer (B) data. Group averaged performance scores are mapped on inflated cortices using the CONN software (Whitfield-Gabrieli and Nieto-Castanon, 2012) ($P_{FDR} < 0.05$, minimum cluster size > 10 voxels). The color bar indicates the prediction performance score (0 = chance, 1 = perfect prediction). The full model predicts neural responses in the bilateral STS particularly well in both studies. L = left hemisphere, R = right hemisphere, FDR = the false discovery rate.

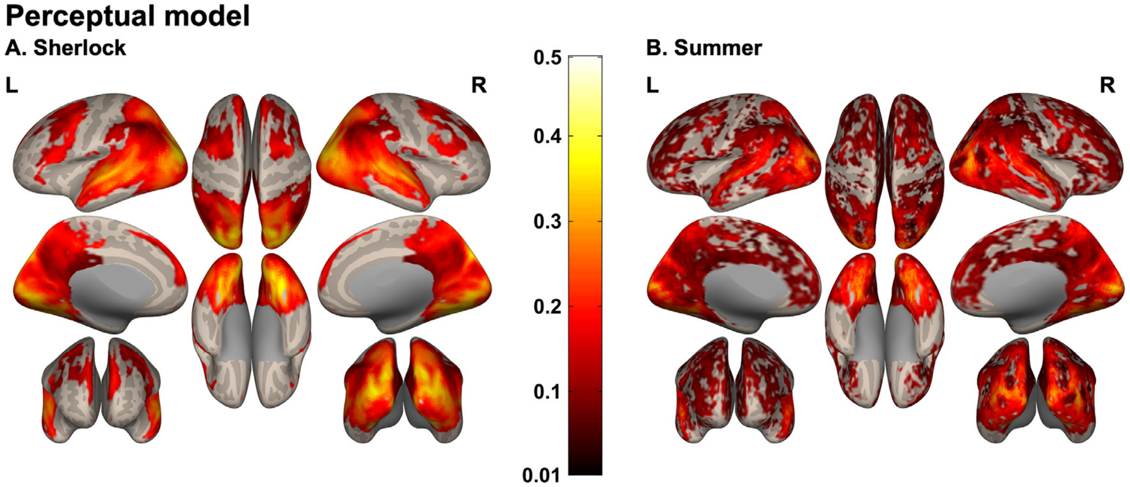


Fig 4. The perceptual model prediction performance observed in Sherlock (A) and Summer (B) data. The perceptual model significantly predicted neural responses in the visual cortex in both studies. Group averaged performance maps are thresholded at $P_{FDR} < 0.05$ and the minimum cluster size > 10 voxels.

ceived valence, and arousal) also produced significantly above chance performance throughout the whole brain in both studies. The social-affective model explained significant variance in 100 and 99.92% of voxels inside of the ISC mask in the Sherlock (range of prediction performance = 0.04–0.55, mean = 0.20, std = 0.08) and Summer data (range = 0–0.39, mean = 0.12, std = 0.05), respectively. The highest model performance was observed in the left STS in both experiments (X, Y, Z = −60, −24, −3 for Sherlock; X, Y, Z = −63, −18, 1 for Summer). Again, findings are bilateral (highest performance = 0.49 in the right STS (X, Y, Z = 51, −33, 0) for Sherlock; 0.38 for Summer (X, Y, Z = 64, −24, 1)) (Fig. 5).

3.2. *Social interaction features are the strongest predictor of STS activity, while deep neural network features are the strongest nearly everywhere else*

We next performed preference mapping analyses to measure the relative contribution of each stimulus feature in predicting held out BOLD responses for every voxel. Fig. 6 illustrates winning features that best (or second best) capture voxel-wise responses in each movie. The DNN fifth layer was the most predictive feature for most voxels throughout the brain (pink in Fig 6A,B). Notably, however, this was not true in the bilateral STS where two social interaction-related features, an agent speaking (green in Fig 6) and presence of social interactions (green-yellow in Fig 6), are the most or second most preferred stimulus features in both studies. As expected, the auditory cortex was best explained by the audio amplitude feature (yellow-orange in Fig 6) in both studies. Visual features, i.e., motion energy, faces, and written words, are the second-best features explaining neural responses in ventral and dorsal visual pathways (red colors in Fig 6C,D). TPJ and mPFC were second-best explained by social-affective features (theory of mind, speaking, social interaction, and arousal) (green and purple colors in Fig 6C,D). Frequency of feature occurrence does not seem to predict voxel preference. Although faces, indoor scenes, and background music are frequently present in both movies (Table 1), these features are not the ones preferred by most voxels.

3.3. *The perceptual model uniquely predicts responses in the visual and auditory cortex, while the social-affective model does so in the temporal and frontal regions*

While above results indicate that STS and nearby regions in the temporal lobe are best explained by social features, they do not reveal to

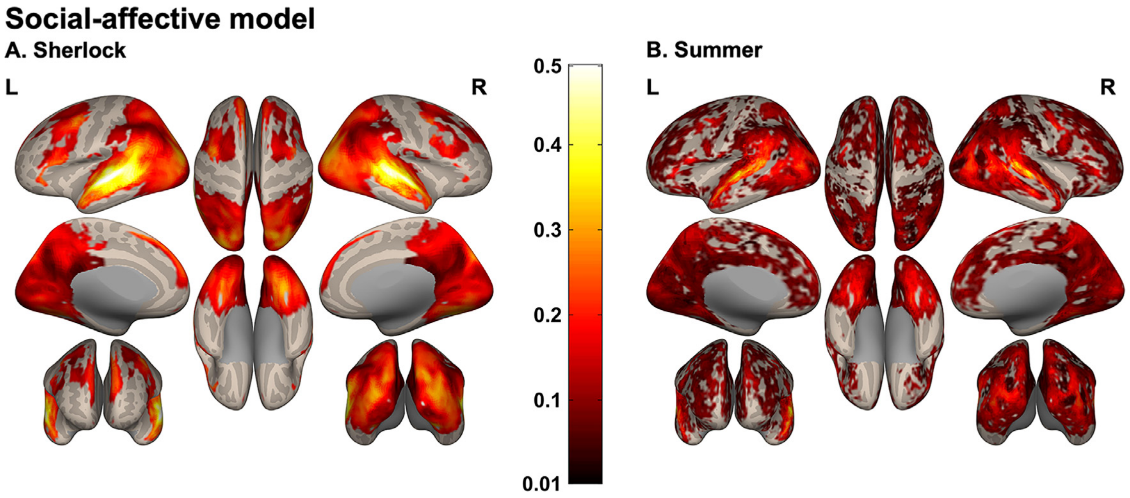


Fig. 5. The social-affective model prediction performance observed in Sherlock (A) and Summer data (B). Group averaged performance scores are plotted on inflated cortices ($P_{FDR} < 0.05$, minimum cluster size > 10 voxels). Like the full model, the social-affective model significantly predicted neural responses in the bilateral STS in both experiments.

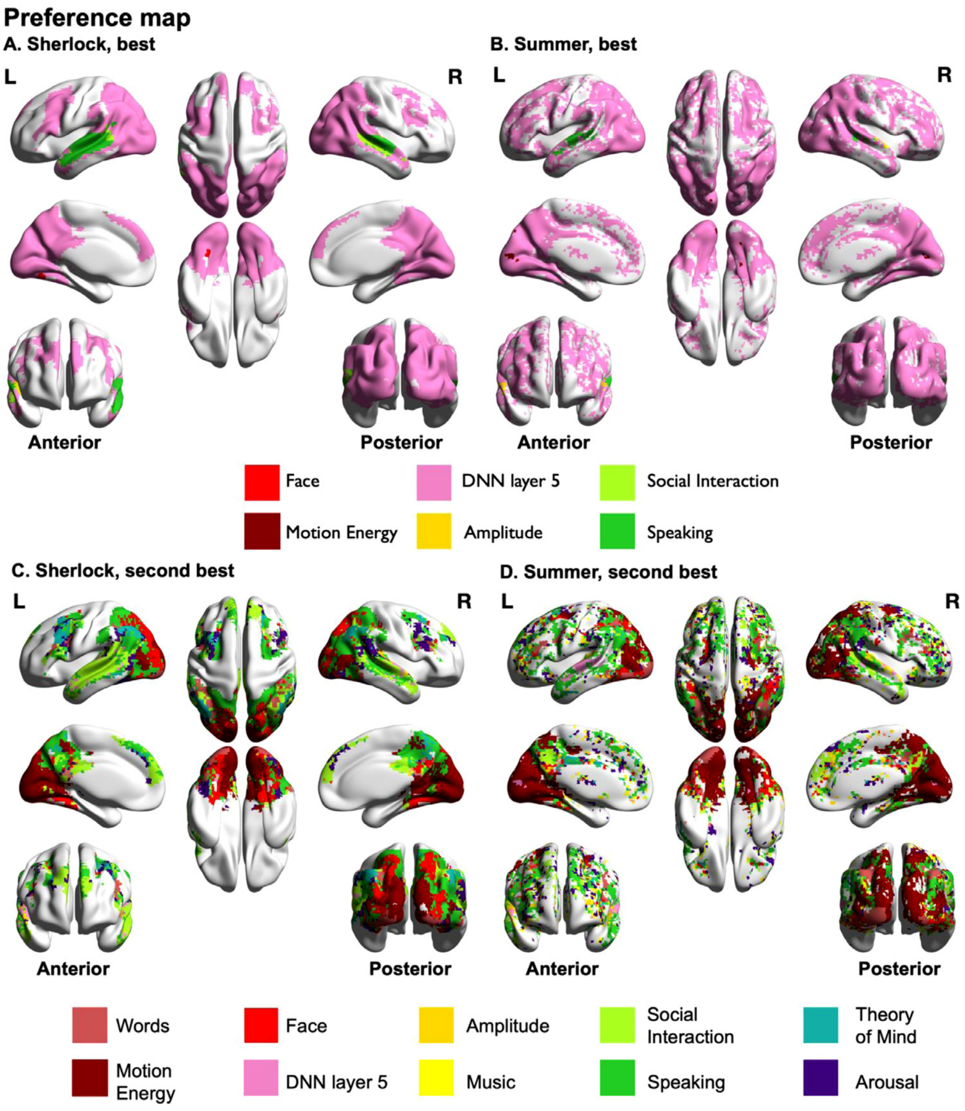


Fig. 6. Preference maps for the best (panel A and B) and second-best features (panel C and D) in each voxel for Sherlock (A and C) and Summer (B and D) data. Feature color codes indicated at bottom. DNN features best explain the neural responses throughout most of the brain in both studies, except the STS, whose neural responses are best explained by social features. Overall, social-affective features (green and purple colors) are the most or at least the second most preferred in the temporal and frontal lobe in both studies, whereas visual features (red colors), in addition to DNN features, are preferred in the visual cortex (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

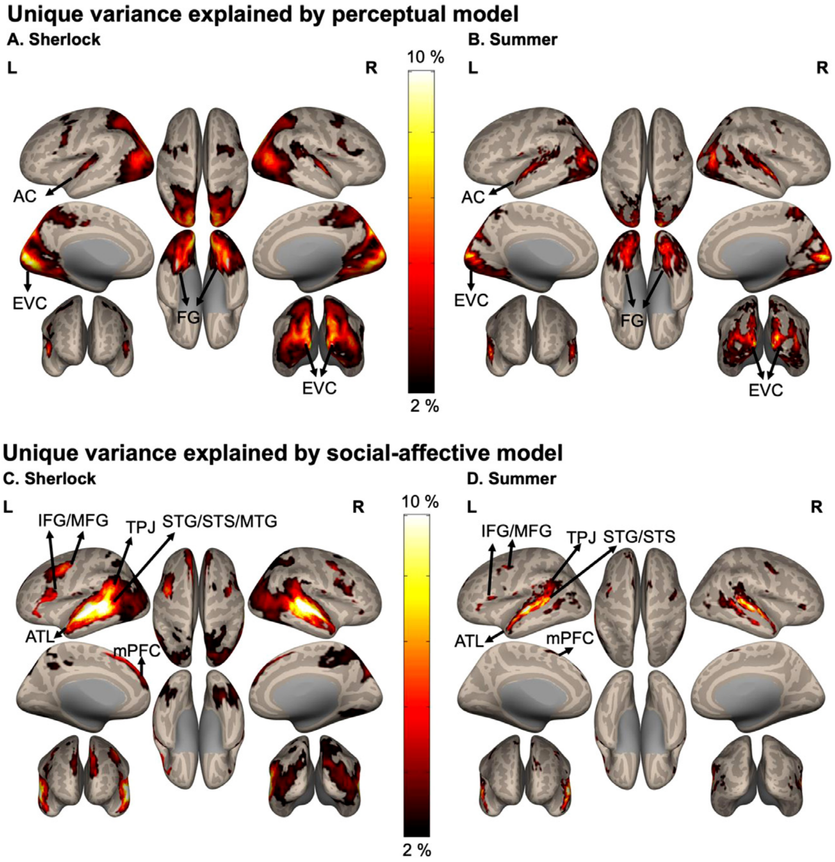

Unique variance explained by perceptual model
A. Sherlock
B. Summer
L
R
10 %
2 %
AC
EVC
FG
Unique variance explained by social-affective model
C. Sherlock
D. Summer
IFG/MFG
TPJ
STG/STS/MTG
STG/STS
ATL
mPFC

Fig. 7. The portion of unique variance, expressed as percentages, explained by the perceptual (A, B) and social-affective model (C, D) in the Sherlock and Summer data. Color bar indicates percentage of unique variance explained. The perceptual model explained unique variance of neural responses in the visual and auditory cortices. In contrast, the social-affective model explained the unique variance of social brain responses, including the STS, TPJ, ATL, and mPFC activations. Group averaged variance partitioning maps are thresholded at $P_{FDR} < 0.05$, the minimum cluster size > 10 voxels, and the portion of explained unique variance > 2%. AC = auditory cortex, EVC = early visual cortex, FG = fusiform gyrus, IFG = inferior frontal gyrus, MFG = middle frontal gyrus, STG = superior temporal gyrus, STS = superior temporal sulcus, MTG = middle temporal gyrus, TPJ = temporoparietal junction, ATL = anterior temporal lobe, mPFC = medial prefrontal cortex. For labeling methods, see Variance Partitioning in Materials and Methods.

what extent perceptual and social-affective features contribute to the neural response, independently of their co-varying perceptual features. By conducting variance partitioning analyses (Fig. 1B), we examined the unique contribution of the perceptual and social-affective feature models to the prediction of BOLD responses throughout the brain. This analysis is particularly important as many features are at least somewhat correlated with each other (e.g., rank correlation between social interactions and the presence of faces $r = 0.48$ in Sherlock and $r = 0.35$ in Summer, Fig 2).

The results indicated that the perceptual model significantly explained unique variance in brain regions implicated in visual or auditory processing (Fig 7A,B). The largest portion of unique variance explained was found in the early visual cortex in both studies (X, Y, Z = −6, −90, 3 for Sherlock; X, Y, Z = 13, −87, 7 for Summer).

On the other hand, the social-affective model uniquely predicted the voxel-wise responses in high-level social cognitive regions, including the STS, TPJ, anterior temporal lobe (ATL), and mPFC in both studies (Fig 7C,D). The largest portion of unique variance explained was in the left STS at the same or immediately adjacent MNI coordinates where the social-affective model shows the highest performance. Note that a substantial portion of the shared variance explained by both models was observed in the temporal lobe and occipital lobe (data not shown). The fact that perceptual and social-affective features covary may explain this finding.

3.4. *The presence of a social interaction uniquely predicts brain response in the STS*

Finally, we sought to identify whether two major, independent (though often correlated) social features, social interactions and theory of mind, independently predict brain activity during natural movie viewing. For the Sherlock data, the presence versus absence of social interactions uniquely predicted brain responses throughout the temporal lobe and to some extent in precuneus and inferior parietal sulcus (IPS), with the bilateral STS showing the strongest selectivity (Fig. 8A). The largest portion of unique contribution was observed in the left STS at the same MNI coordinates, where the social-affective model showed the highest performance. The Summer fMRI data showed similar results, although the predicted brain areas were confined to right STS and bilateral precuneus, with the right STS (X, Y, Z = 52, −42, 7) showing the strongest selectivity (Fig. 8B). This supports and extends the results of our feature preference mapping (Fig 6), highlighting the role of social interaction processing in the STS, independent of all other features, including spoken language. Supplementary statistical analysis quantifying the difference between the two movies yielded similar results to those seen qualitatively in Fig. 8A,B. Social interactions explained significantly more unique variance in the bilateral STS in Sherlock compared to Summer (Fig S1).

In the Sherlock data the theory of mind feature uniquely predicted neural responses of other social brain regions, mainly located in the theory of mind network (Dufour et al., 2013) – the precuneus, mPFC, and TPJ (Fig. 8C), with the precuneus showing the strongest selectivity (X, Y, Z = 6, −60, 51). However, this distinct contribution of the theory of mind feature, observed in the Sherlock data, was not replicated in the Summer fMRI data. Only a few voxels, fewer than 10 voxels in each cluster, in STS, TPJ, and mPFC showed selectivity to the theory of mind feature (Fig. 8D). We confirmed this finding by quantifying the statistical difference in unique variance explained by ToM across the two movies (Fig S1). Nonetheless, uncorrected variance partitioning results (i.e., $P_{uncorrected} < 0.05$) included the same brain regions reported in Sherlock, such as TPJ, precuneus, and mPFC, and were largely non-overlapping with the unique social interaction voxels.

In addition, no voxel showed the shared variance explained by both social interaction and theory of mind in Sherlock data. We did not run this analysis in Summer data considering the weak results found in the theory of mind feature. Overall, the results show distinct func-

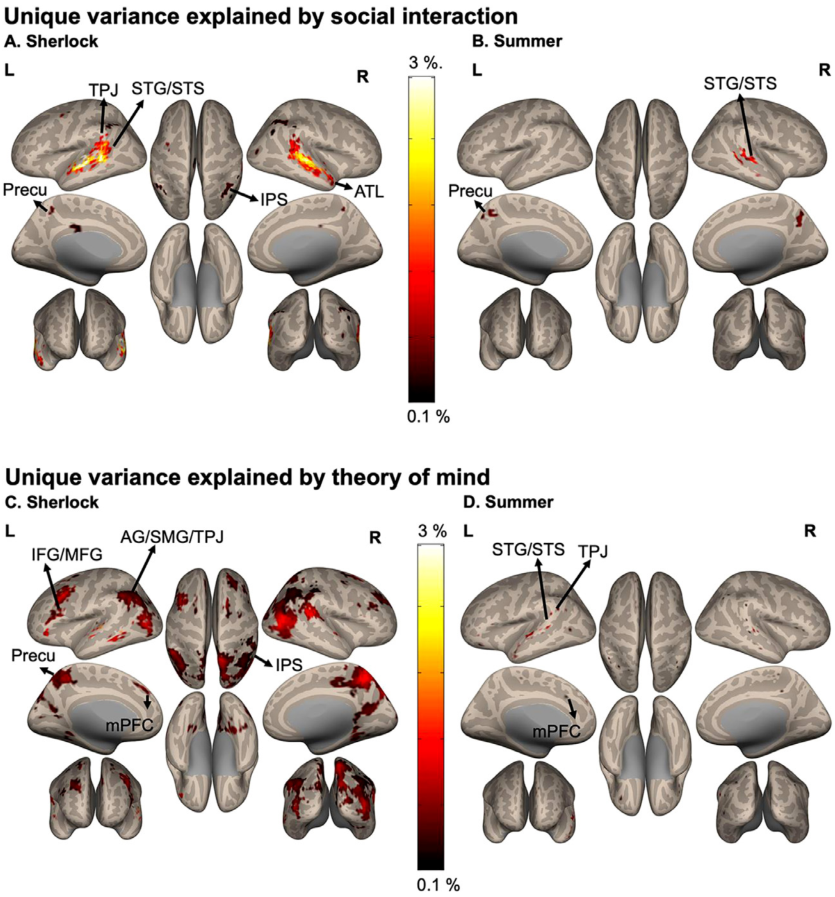


Fig 8. The amount of the unique variance explained by the social interaction (A, B) and theory of mind feature (C, D) in the Sherlock and Summer studies. Color bar indicates percentage of unique variance explained. The social interaction feature explained unique variance in neural responses in the STS and precuneus. The theory of mind feature explained the unique variance of neural responses in the theory of mind network, such as TPJ, precuneus, and mPFC. Group averaged variance partitioning maps are thresholded at $P_{FDR} < 0.05$ and the minimum cluster size > 10 voxels (except for the panel D, where the minimum cluster size > 1 voxel). STG = superior temporal gyrus, STS = superior temporal sulcus, MFG = middle frontal gyrus, ATL = anterior temporal lobe, mPFC = medial prefrontal cortex, TPJ = temporoparietal junction, AG = angular gyrus, SMG = supramarginal gyrus, IPS = intraparietal sulcus, Precu = precuneus.

tional and anatomical divisions between social interaction and theory of mind processing in natural movies. In particular, the STS shows strong, unique selectivity to scenes with social interactions, and the precuneus, mPFC, and TPJ show selectivity for scenes where characters infer others' thoughts and emotions.

4. Discussion

Here we investigated the brain mechanisms underlying naturalistic social interaction perception in a more ecologically valid and replicable context. Combining voxel-wise encoding and variance partitioning analyses (Fig. 1), we identified the brain regions showing unique selectivity for general social-affective information (Fig 7C,D) and those particularly selective to social interactions (Fig. 8A,B). We also demonstrated how auditory and visual features are encoded (Fig 7A,B), replicating prior findings (Brugge et al., 2009; Cohen et al., 2002; Hart et al., 2003; Kanwisher et al., 1997; Sunaert et al., 1999) in a natural setting. Importantly, our findings were replicated across both sets of movie data that came from different genres, subjects, and labs.

4.1. *Social-affective information processing during natural viewing*

Movie viewing paradigms have recently been highlighted as essential tools in cognitive neuroscience due to their richness and comparable complexity to the real world (Redcay and Moraczewski, 2020; Sonkusare et al., 2019). Most prior social neuroscience studies with movies have used reverse correlation analyses to investigate phenomena such as theory of mind (Richardson et al., 2018) and social interaction perception (Wagner et al., 2016). These methods present a major advantage in that they do not require extensive movie labeling, but are prone to reverse inference errors as the event labeling happens post-hoc. This is particularly challenging in movies where the distribution of social information is often imbalanced (see Table 1). In the current study we densely labeled movies, using a combination of automatic and human annotations, and used the extracted movie features to perform cross-validated voxel-wise encoding analysis, which is less prone to reverse inference errors (Redcay and Moraczewski, 2020).

Using these methods, we found several brain regions that coded for social-affective information-social interaction, an agent speaking, mentalizing, perceived valence, and arousal-independent of co-varying perceptual information (Fig 7C,D). Specifically, we identified unique variance explained in brain regions previously attributed to social perception (STS) (Hooker et al., 2003; Lee Masson et al., 2020a, 2018; Pegado et al., 2018; Vangeneugden et al., 2014), theory of mind (TPJ, mPFC, ATL) (see the review in (Schurz et al., 2020), and action observation (inferior frontal gyrus (IFG)) (Carr et al., 2003; Centelles et al., 2011).

Surprisingly, the preferred feature across most of the brain, including many of the above social-affective regions, was the fifth layer of a DNN pre-trained on an object recognition task (Fig 6). This result is hard to interpret as most prior neuroimaging studies using DNNs have focused on their match to voxel responses in the ventral visual stream (Bonner and Epstein, 2018; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Zeman et al., 2020). One prior study found that object-trained DNNs predict voxel-wise responses in the TPJ (Wen et al., 2018), concluding that later DNN layers might capture high-level semantic information in naturalistic visual scenes. Further investigation is needed to better understand the substantial contribution of late-stage DNN layers in explaining social brain activity. Intriguingly, the most notable exception to DNN preference was preference for social interactions in the STS. It is difficult however, to rule out the effect of spoken language processing (Deen et al., 2015; Wilson et al., 2018) on STS activity with this analysis alone. Although we tried to capture this information with our auditory and spoken language features, it is likely these do not capture all of the rich semantic content present in the movie. As the semantic network largely overlaps with brain areas observed in the current study (de Heer et al., 2017; Huth et al., 2016), future investigations should use language models to understand how semantic content overlaps with different types of social perception. With this limitation in mind, we turned to variance partitioning, to identify the unique contribution of social interactions to brain responses.

4.2. Social interaction perception during natural viewing

In both studies, the right STS and, to a lesser extent, the precuneus showed unique selectivity for naturalistic social interactions (Fig 8A,B), independent of all other covarying features. These results provide the first evidence that unique variance in STS is explained by the presence/absence of social interactions during natural viewing. Although we replicated our findings across two movie datasets, we observed a slight discrepancy with respect to the extent of STS activity and the lateralization (see Fig S1 for more information). This discrepancy has also been observed in other studies (e.g., Isik et al. 2017, Lee Masson et al., 2018, Walbrin et al., 2018) found right lateralization, while others did not (Lahnakoski et al., 2012; Walbrin et al., 2020; Walbrin and Koldewyn, 2019). Our findings highlight that although STS is known as a hub for general social processing, social interaction is a critical feature that uniquely contributes to STS responses.

Unique variance in the precuneus was also explained by social interaction, but to a lesser degree than STS. The precuneus has been implicated in a wide range of cognitive processes, including false belief tasks (Jacoby et al., 2016; Saxe and Kanwisher, 2003), social trait judgment tasks (Farrer and Frith, 2002; Iacoboni et al., 2004), and while viewing social interactions (Iacoboni et al., 2004; Lahnakoski et al., 2012). A possible interpretation of our findings is that the STS and precuneus work in concert to spontaneously extract socially relevant information about others. This notion is supported by previous studies showing increased precuneus's connectivity with other social brain areas during the observation of dyadic interaction compared to human-object manipulation (Lee Masson et al., 2020b) and social evaluation task on others' faces (McCormick et al., 2018).

A prior study found increased response to social interaction scenes of a movie in mPFC, concluding that a viewer may spontaneously infer movie characters' thoughts and intentions (Wagner et al., 2016). Contrary to that finding, we did not observe mPFC involvement specific to social interaction perception after controlling for the effects of perceptual and social features, including speaking and theory of mind. Instead, mPFC was selectively tuned to more general social-affective features (Fig 7C-D) and theory of mind (Fig 8C,D). Notably, in Wagner and colleagues' study, movie scenes that evoked strong mPFC responses were speaking scenes identified with reverse correlation analysis, and they did not consider theory of mind features in their analysis. Given that social interaction scenes genuinely invite theory of mind (Dziobek et al., 2006; Grainger et al., 2019; Roeyers et al., 2001), it may make more sense to compare their findings to our findings on general social-affective features that include speaking, social interaction, and theory of mind. Our variance partitioning results show for the first time that social interaction perception is separable from theory of mind processing, even in natural conditions.

4.3. Theory of mind during natural viewing

Recent results have shown increased responses in the theory of mind network, including TPJ, precuneus, and mPFC, during movie viewing (Jacoby et al., 2016; Richardson et al., 2018). While we largely replicated these results in the Sherlock data, we did not observe the same in the Summer data. This discrepancy may be related to the substantially lower signal-to-noise ratio, which may be driven from the usage of 1.5T MRI scanner and larger inter-subject variability, in the Summer movie data (maximum ISC value 0.73 vs. 0.56, see inter-subject brain correlation in Methods). To examine this discrepancy further, we quantified a difference between the results of the two movies (see Fig S1 for the Sherlock > Summer contrast). It is unlikely that this discrepancy is due to large inter-subject variability in Summer data as we took subject-level noise into account by using the ISC for noise-ceiling normalization. However, this method does not correct for global noise, such as noise from the MRI scanner (1.5T instead of 3T) shared across participants. Thus, it is difficult to determine whether this discrepancy is caused by a difference in the movies versus global noise levels shared by participants. Another important distinction is how we annotated the theory of mind feature. Unlike previous studies (Jacoby et al., 2016; Richardson et al., 2018), our theory of mind annotation was labeled in advance based on whether the scene contained a person speaking about others' mental states. We believe this criterion to be more objective than trying to guess whether the subjects were engaged in mentalization. However, it may pose issues for certain movie scenes. For example, 500 Days of Summer (unlike Sherlock) contains many scenes with a narrator describing characters' mental states in voice-over monologues. We did not distinguish between scenes with the narrator and examples of mentalization that happened more naturally in the course of the movie. In future studies, our operationalized definition of 'second order' theory of mind could be compared to the subjects' first order mentalization as they view the movie by intermittently asking questions about characters' mental states. Further, the unique contribution of different types of mentalization (from different sources (Koster-Hale et al., 2014) or characters with different motives (Young et al., 2007)) could be compared. Finally, scenes identified via reverse correlation analysis as evoking strong activity in the ToM network (Richardson et al., 2018; Wagner et al., 2016) could be operationalized and labeled in subsequent analyses and compared to the activity evoked by second-order ToM identified here.

5. Conclusions

Using voxel-wise encoding and variance partitioning methods we found that the human brain encodes social interaction of others, independent of other crucial social-affective features such as theory of mind. The methods and findings presented in this study open the door for many avenues of naturalistic social neuroscience research. Our results suggest that, even in a real-world setting, social interactions are processed in a manner that is distinct from other perceptual and social features. In both simple visual displays (Su et al., 2016) and natural images (Skripkauskaite et al., 2021), there is an attentional bias for social interactions. Why do social interactions capture our attention and how do we use them to reason about interacting individuals? Is there a computational advantage to selectively processing social interactions?

Finally, our method of isolating brain areas selective to naturalistic social interaction in MRI may be particularly advantageous for studies of typical and atypical development, including autism. Many individuals with autism look identical to neurotypical adults in simple social psychology and neuroscience tasks despite differences in their real-world social abilities (Moessnang et al., 2020; Scheeren et al., 2013). Using our method with the addition of hyperalignment (Haxby et al., 2020b) would benefit cross participants and group membership prediction in future studies. The methods outlined here may help us close the gap between simple lab-based tasks and the real world.

Data and code availability statement

fMRI data and original movie annotations are available from the original authors for Sherlock (https://dataspace.princeton.edu/handle/88435/dsp01nz8062179) and Summer (https://openneuro.org/datasets/ds002837/, https://www.naturalistic-neuroimaging-database.org/index.html). All code for voxel-wise encoding and variance partitioning analysis along with new movie annotations are available at https://github.com/haemyleemasson/voxelwise_encoding.

Declaration of Competing Interest

The authors have no conflicts of interest.

Credit authorship contribution statement

Haemy Lee Masson: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Leyla Isik:** Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing.

Acknowledgments

We thank Janice Chen, Sarah Aliko, Jongwan Kim and their colleagues for data and annotations used in this study. We would like to thank J. Brendan Ritchie and Alon Hafri for valuable comments. This work was supported with funds from The Clare Boothe Luce Program for Women in STEM.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118741.

References

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., Skipper, J.I., 2020. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. Sci. Data 7, 1–21. doi:10.1038/s41597-020-00680-2.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721. doi:10.1016/j.neuron.2017.06.041, e5.

Bonner, M.F., Epstein, R.A., 2018. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLOS Comput. Biol. 14, e1006111. doi:10.1371/journal.pcbi.1006111.

Brugge, J.F., Nourski, K.V., Oya, H., Reale, R.A., Kawasaki, H., Steinschneider, M., Howard, M.A., 2009. Coding of repetitive transients by auditory cortex on Heschl's gyrus. J. Neurophysiol. 102, 2358–2374. doi:10.1152/jn.91346.2008.

Carr, L., Iacoboni, M., Dubeaut, M.C., Mazziotta, J.C., Lenzi, G.L., 2003. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. Proc. Natl. Acad. Sci. U. S. A. 100, 5497–5502. doi:10.1073/pnas.0935845100.

Centelles, L., Assaiante, C., Nazarian, B., Anton, J.-L., Schmitz, C., 2011. Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: a neuroimaging study. PLoS ONE 6, e15749. doi:10.1371/journal.pone.0015749.

Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., Hasson, U., 2017. Shared memories reveal shared structure in neural activity across individuals. Nat. Neurosci. 20, 115–125. doi:10.1038/nn.4450.

Chen, P.H.A., Jolly, E., Cheong, J.H., Chang, L.J., 2020. Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. Neuroimage 216, 116851. doi:10.1016/j.neuroimage.2020.116851.

Cheney, D.L., Seyfarth, R.M., 1986. The recognition of social alliances by vervet monkeys. Anim. Behav. 34, 1722–1731. doi:10.1016/S0003-3472(86)80259-7.

Cichy, R.M., Khosla, A., Pantazis, D., Oliva, A., 2017. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. Neuroimage 153, 346–358. doi:10.1016/j.neuroimage.2016.03.063.

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., Dehaene, S., 2002. Language-specific tuning of visual cortex? Functional properties of the visual word form area. Brain 125, 1054–1069. doi:10.1093/brain/awf094.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37, 6539–6557. doi:10.1523/JNEUROSCI.3267-16.2017.

Deen, B., Koldewyn, K., Kanwisher, N., Saxe, R., 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cereb. Cortex 25, 4596–4609. doi:10.1093/cercor/bhv111.

Dufour, N., Redcay, E., Young, L., Mavros, P.L., Moran, J.M., Triantafyllou, C., Gabrieli, J.D.E., Saxe, R., 2013. Similar brain activation during false belief tasks in a large sample of adults with and without autism. PLoS ONE 8, e75468. doi:10.1371/journal.pone.0075468.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J.K., Wolf, O.T., Convit, A., 2006. Introducing MASC: a movie for the assessment of social cognition. J. Autism Dev. Disord. 36, 623–636. doi:10.1007/s10803-006-0107-0.

Farrer, C., Frith, C.D., 2002. Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. Neuroimage 15, 596–603. doi:10.1006/nimg.2001.1009.

Grainger, S.A., Steinvik, H.R., Henry, J.D., Phillips, L.H., 2019. The role of social attention in older adults' ability to interpret naturalistic social scenes. Q. J. Exp. Psychol. 72, 1328–1343. doi:10.1177/1747021818791774.

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015.

Hamlin, J.K., Wynn, K., 2011. Young infants prefer prosocial to antisocial others. Cogn. Dev. 26, 30–39. doi:10.1016/j.cogdev.2010.09.001.

Hart, H., Palmer, A., Hall, D., 2003. Amplitude and frequency-modulated stimuli activate common regions of human auditory cortex. Cereb. Cortex 13, 773–781. doi:10.1093/cercor/13.7.773.

Hasson, U., Malach, R., Heeger, D.J., 2010. Reliability of cortical activity during natural stimulation. Trends Cogn. Sci. doi:10.1016/j.tics.2009.10.011.

Haxby, J.V., Gobbini, M.I., Nastase, S.A., 2020a. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. Neuroimage 216, 116561. doi:10.1016/j.neuroimage.2020.116561.

Haxby, J.V., Guntupalli, J.S., Nastase, S.A., Feilong, M., 2020b. Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. Elife 9, 1–26. doi:10.7554/ELIFE.56601.

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., Cichy, R.M., 2018. The representational dynamics of task and object processing in humans. Elife 7. doi:10.7554/eLife.32816.

Hooker, C.I., Paller, K.A., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., Reber, P.J., 2003. Brain networks for analyzing eye gaze. Cogn. Brain Res. 17, 406–418. doi:10.1016/S0926-6410(03)00143-5.

Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. doi:10.1038/nature17637.

Iacoboni, M., Lieberman, M.D., Knowlton, B.J., Molnar-Szakacs, I., Moritz, M., Throop, C.J., Fiske, A.P., 2004. Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. Neuroimage 21, 1167–1173. doi:10.1016/j.neuroimage.2003.11.013.

Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N., 2017. Perceiving social interactions in the posterior superior temporal sulcus. Proc. Natl. Acad. Sci, 201714471 doi:10.1073/pnas.1714471114.

Jacoby, N., Bruneau, E., Koster-Hale, J., Saxe, R., 2016. Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. Neuroimage 126, 39–48. doi:10.1016/j.neuroimage.2015.11.025.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311. doi:10.1523/jneurosci.17-11-04302.1997.

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915. doi:10.1371/journal.pcbi.1003915.

Kim, J., Weber, C.E., Gao, C., Schulteis, S., Wedell, D.H., Shinkareva, S.V., 2020. A study in affect: predicting valence from fMRI data. Neuropsychologia 143, 107473. doi:10.1016/j.neuropsychologia.2020.107473.

Koster-Hale, J., Bedny, M., Saxe, R., 2014. Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. Cognition 133, 65–78. doi:10.1016/J.COGNITION.2014.04.006.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105.

Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012. Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front. Hum. Neurosci. 6, 233. doi:10.3389/fnhum.2012.00233.

Lee Masson, H., Op de Beeck, H., Boets, B., 2020a. Reduced task-dependent modulation of functional network architecture for positive versus negative affective touch processing in autism spectrum disorders. Neuroimage doi:10.1016/j.neuroimage.2020.117009.

Lee Masson, H., Pillet, I., Boets, B., Op de Beeck, H., 2020b. Task-dependent changes in functional connectivity during the observation of social and non-social touch interaction. Cortex doi:10.1016/j.cortex.2019.12.011.

Lee Masson, H., Van De Plas, S., Daniels, N., Op de Beeck, H., 2018. The multidimensional representational space of observed socio-affective touch experiences. Neuroimage 175, 297–314. doi:10.1016/j.neuroimage.2018.04.007.

Mars, R.B., Sallet, J., Schüffelgen, U., Jbabdi, S., Toni, I., Rushworth, M.F.S., 2012. Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. Cereb. Cortex 22, 1894–1903. doi:10.1093/cercor/bhr268.

McCormick, E.M., van Hoorn, J., Cohen, J.R., Telzer, E.H., 2018. Functional connectivity in the social brain across childhood and adolescence. Soc. Cogn. Affect. Neurosci. 13, 819–830. doi:10.1093/scan/nsy064.

va der Meer, J.N., Breakspear, M., Chang, L.J., Sonkusare, S., Cocchi, L., 2020. Movie viewing elicits rich and reliable brain state dynamics. Nat. Commun. 11, 1–14. doi:10.1038/s41467-020-18717-w.

Moessnang, C., Baumeister, S., Tillmann, J., Goyard, D., Charman, T., Ambrosino, S., Baron-Cohen, S., Beckmann, C., Bölte, S., Bours, C., Crawley, D., Dell'Acqua, F., Durston, S., Ecker, C., Frouin, V., Hayward, H., Holt, R., Johnson, M., Jones, E., Lai, M.C., Lombardo, M.V., Mason, L., Oldenhinkel, M., Persico, A., Cáceres, A.S.J., Spooren, W., Loth, E., Murphy, D.G.M., Buitelaar, J.K., Banaschewski, T., Brandeis, D., Tost, H., Meyer-Lindenberg, A., 2020. Social brain activation during mentalizing in a large autism cohort: the longitudinal European autism project. Mol. Autism 11, 17. doi:10.1186/s13229-020-0317-x.

Nastase, S.A., Gazzola, V., Hasson, U., Keysers, C., 2019. Measuring shared responses across subjects using intersubject correlation. Soc. Cogn. Affect. Neurosci. 14, 669–687. doi:10.1093/scan/nsz037.

Nastase, S.A., Goldstein, A., Hasson, U., 2020. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. Neuroimage 222. doi:10.1016/j.neuroimage.2020.117254.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646. doi:10.1016/j.cub.2011.08.031.

Nunez-Elizalde, A.O., Huth, A.G., Gallant, J.L., 2019. Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage 197, 482–492. doi:10.1016/j.neuroimage.2019.04.012.

Pegado, F., Hendriks, M.H.A., Amelynck, S., Daniels, N., Bulthé, J., Lee Masson, H., Boets, B., Op de Beeck, H., 2018. A multitude of neural representations behind multisensory "social norm" processing. Front. Hum. Neurosci. 12, 153. doi:10.3389/fnhum.2018.00153.

Quadflieg, S., Koldewyn, K., 2017. The neuroscience of people watching: how the human brain makes sense of other people's encounters. Ann. N. Y. Acad. Sci. 1396, 166–182. doi:10.1111/nyas.13331.

Redcay, E., Moraczewski, D., 2020. Social cognition in context: a naturalistic imaging approach. Neuroimage 216, 116392. doi:10.1016/j.neuroimage.2019.116392.

Richardson, H., 2019. Development of brain networks for social functions: confirmatory analyses in a large open source dataset. Dev. Cogn. Neurosci. 37, 100598. doi:10.1016/j.dcn.2018.11.002.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., Saxe, R., 2018. Development of the social brain from age three to twelve years. Nat. Commun. 9, 1027. doi:10.1038/s41467-018-03399-2.

Roeyers, H., Buysse, A., Ponnet, K., Pichal, B., 2001. Advancing advanced mind-reading tests: empathic accuracy in adults with a pervasive developmental disorder. J. Child Psychol. Psychiatry 42, 271–278. doi:10.1111/1469-7610.00718.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. doi:10.1007/s11263-015-0816-y.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind. Neuroimage 19, 1835–1842. doi:10.1016/S1053-8119(03)00230-1.

Saxe, R., Powell, L.J., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol. Sci. 17, 692–699. doi:10.1111/j.1467-9280.2006.01768.x.

Scheeren, A.M., De Rosnay, M., Koot, H.M., Begeer, S., 2013. Rethinking theory of mind in high-functioning autism spectrum disorder. J. Child Psychol. Psychiatry Allied Discip. 54, 628–635. doi:10.1111/jcpp.12007.

Schurz, M., Radua, J., Tholen, M.G., Maliske, L., Margulies, D.S., Mars, R.B., Sallet, J., Kanske, P., 2020. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol. Bull. doi:10.1037/bul0000303.

S. Skripkauskaite, I. Mihai, K. Koldewyn, 2021. Brief report: attentional bias towards social interactions during viewing of naturalistic scenes. bioRxiv 2021.02.26.433078. doi:10.1101/2021.02.26.433078.

Sonkusare, S., Breakspear, M., Guo, C., 2019. Naturalistic stimuli in neuroscience: critically acclaimed. Trends Cogn. Sci. doi:10.1016/j.tics.2019.05.004.

Su, J., van Boxtel, J.J.A., Lu, H., 2016. Social interactions receive priority to conscious perception. PLoS ONE 11, e0160468. doi:10.1371/journal.pone.0160468.

Sunaert, S., Van Hecke, P., Marchal, G., Orban, G.A., 1999. Motion-responsive regions of the human brain. Exp. Brain Res. 127, 355–370. doi:10.1007/s002210050804.

Tholen, M.G., Trautwein, F.M., Böckler, A., Singer, T., Kanske, P., 2020. Functional magnetic resonance imaging (fMRI) item analysis of empathy and theory of mind. Hum. Brain Mapp. 41, 2611–2628. doi:10.1002/hbm.24966.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289. doi:10.1006/nimg.2001.0978.

Vangeneugden, J., Peelen, M.V., Tadin, D., Battelli, L., 2014. Distinct neural mechanisms for body form and body motion discriminations. J. Neurosci. 34, 574–585. doi:10.1523/JNEUROSCI.4032-13.2014.

Wagner, D.D., Kelley, W.M., Haxby, J.V., Heatherton, T.F., 2016. The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. J. Neurosci. 36, 6917–6925. doi:10.1523/JNEUROSCI.4220-15.2016.

Walbrin, J., Downing, P., Koldewyn, K., 2018. Neural responses to visually observed social interactions. Neuropsychologia 112, 31–39. doi:10.1016/j.neuropsychologia.2018.02.023.

Walbrin, J., Koldewyn, K., 2019. Dyadic interaction processing in the posterior temporal cortex. Neuroimage 198, 296–302. doi:10.1016/j.neuroimage.2019.05.027.

Walbrin, J., Mihai, I., Landsiedel, J., Koldewyn, K., 2020. Developmental changes in visual responses to social interactions. Dev. Cogn. Neurosci. 42, 100774. doi:10.1016/j.dcn.2020.100774.

Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z., 2018. Neural encoding and decoding with deep learning for dynamic natural vision. Cereb. Cortex 28, 4136–4160. doi:10.1093/cercor/bhx268.

Whitfield-Gabrieli, S., Nieto-Castanon, A., 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect 2, 125–141.

Wilson, S.M., Bautista, A., McCarron, A., 2018. Convergence of spoken and written language processing in the superior temporal sulcus. Neuroimage 171, 62–74. doi:10.1016/j.neuroimage.2017.12.068.

Young, L., Cushman, F., Hauser, M., Saxe, R., 2007. The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. 104, 8235–8240. doi:10.1073/PNAS.0701408104.

Zeman, A.A., Ritchie, J.B., Bracci, S., Op de Beeck, H., 2020. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. Sci. Rep. 10, 1–12. doi:10.1038/s41598-020-59175-0.