

# 1           **Functional selectivity for naturalistic social interaction perception in the** 2                                   **human superior temporal sulcus**

3  
4                                   Haemy Lee Masson<sup>1</sup> and Leyla Isik<sup>1</sup>

5                                   Department of Cognitive Science, Johns Hopkins University, United States

6  
7           Correspondence should be addressed to Haemy Lee Masson, Department of Cognitive Science,  
8           Johns Hopkins University, Maryland, 21218, United States, E-mail: [haemy.leemasson@jhu.edu](mailto:haemy.leemasson@jhu.edu)

## 9 10   **Abstract**

11   Recognizing others' social interactions is a crucial human ability. Using simple stimuli, previous studies have shown  
12   that social interactions are processed in the superior temporal sulcus (STS) and the medial prefrontal cortex (mPFC).  
13   However, it remains unknown to what extent social interaction selectivity is observed for real world stimuli where  
14   social interactions covary with other perceptual and social information, such as faces, voices, and theory of mind. By  
15   using a naturalistic fMRI movie paradigm and advanced machine learning methods, we found that broad social-  
16   affective features predict neural responses in social brain regions, including the STS and mPFC. However, only the  
17   STS showed robust and unique selectivity specifically to social interactions, independent from other covarying  
18   features. This selectivity generalized across two separate fMRI datasets. These findings suggest that naturalistic social  
19   interaction perception recruits dedicated neural circuitry in the STS and is a critical dimension of human social  
20   understanding.

## 21   **Introduction**

22   Humans learn a great deal about the social world by observing others' social interactions, defined here as action or  
23   communication between two or more people that are directed at and contingent upon each other. This social  
24   observation helps us form impressions of interacting people (e.g., whether they are prosocial or antisocial), their social  
25   relationships, and relative social status (Quadflieg and Koldewyn, 2017). Our ability to extract social information  
26   about interacting people appears to emerge in infancy. For example, infants show preference for a person or character  
27   who helps others achieve their goals over someone who hinders (Hamlin, 2015; Hamlin et al., 2007; Hamlin and  
28   Wynn, 2011). These findings underscore the importance of observing others' social interactions for making social  
29   judgements.

30           Previous studies have reported that the posterior superior temporal sulcus (pSTS) (Isik et al., 2017; Walbrin  
31   et al., 2018; Walbrin and Koldewyn, 2019) and the cortical midline structure (the medial prefrontal cortex (mPFC)  
32   and the precuneus) are involved in recognizing others' social interactions (Iacoboni et al., 2004; Petrini et al., 2014).  
33   Importantly, social interactions appear to be processed selectively, in a manner that is functionally distinct from other  
34   related social functions, including detecting faces, animacy, other agents' goals and theory of mind. Most prior studies,

35 however, have used simple, artificial stimuli, such as point-light figures or staged images that are largely devoid of  
36 other social information. It thus remains unknown how the humans brain processes social interactions in real world  
37 scenes, where the interactions are highly confounded with other perceptual and social properties, including faces and  
38 theory of mind. The goal of the current study is to identify the brain mechanisms underlying real-world social  
39 interaction perception by adopting a naturalistic movie viewing paradigm.

40 A growing body of evidence in visual neuroscience suggests that naturalistic neuroimaging paradigms better  
41 uncover neural representations of objects, faces, scenes, and actions (Haxby et al., 2020; Nishimoto et al., 2011; Wen  
42 et al., 2018). Natural movies elicit stronger brain responses and higher inter-subject reliability than traditional  
43 experiments (Hasson et al., 2010; Sonkusare et al., 2019). In contrast, social neuroscience has not yet fully exploited  
44 these methods, although improving ecological generalizability in (social) cognitive neuroscience is considered an  
45 urgent challenge (Nastase et al., 2020; Redcay and Moraczewski, 2020). While several recent social neuroscience  
46 studies have adopted this paradigm to investigate theory of mind (Jacoby et al., 2016; Richardson, 2019; Richardson  
47 et al., 2018), only two have investigated how the human brain processes naturalistic social interactions during movie  
48 viewing. One study identified the STS in general social perception, including social interactions (Lahnakoski et al.,  
49 2012), while a second study revealed the involvement of the mPFC in social interaction perception (Wagner et al.,  
50 2016), conjecturing that others' social interactions invite a viewer to infer others' personality and intention. These  
51 studies provided inconsistent evidence, with each study finding evidence for one brain region and not the other, and  
52 neither study spoke to the selectivity for social interactions in the brain. Critically, no prior studies of naturalistic  
53 social perception have identified or controlled for covarying perceptual features.

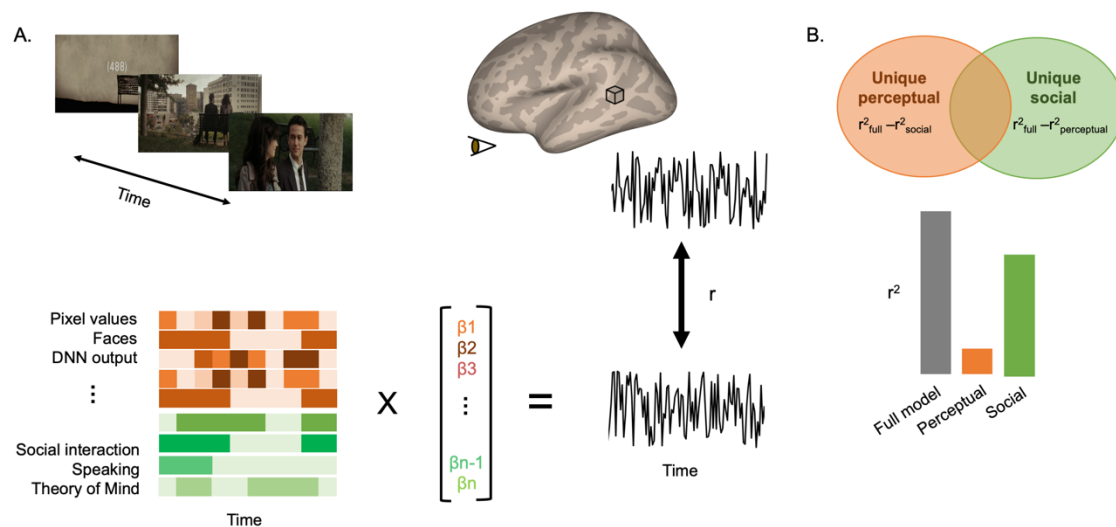
54 Here, we implemented computer vision techniques, machine-learning-based encoding model analyses, and  
55 variance partitioning to identify the unique contribution of social interactions to responses in the human brain. To  
56 improve ecological generalizability, we performed the same analyses on two different fMRI datasets (Aliquo et al.,  
57 2020; Chen et al., 2017) collected by different labs showing different movies from different genres (crime vs. romance)  
58 to different groups of participants living in different countries (US vs. UK) on different MRI scanners (3T vs. 1.5T).  
59 Our findings reveal that that social-affective features, consisting of an agent speaking, social interactions, theory of  
60 mind, perceived valence, and arousal, independently contribute to predicting brain responses in the STS and the mPFC.  
61 However, we find that the STS, but not mPFC, shows unique selectivity for others' social interactions in particular,  
62 independent of other co-varying perceptual and social features.

## 63 Results

64 *Perceptual and social-affective features accurately predict voxel-wise activation throughout the*  
65 *brain.*

66 Two sets of subjects (N = 17 and N = 18) viewed two different movies (the first episode of the Sherlock BBC TV  
67 series and 500 Days of Summer) while their blood-oxygen-level-dependent (BOLD) responses were recorded in fMRI.  
68 We extracted a range of perceptual and social features from each movie using a combination of automatic methods  
69 and human labeling. The perceptual features consisted of low-level sensory features including hue, saturation, value  
70 for each pixel (HSV), motion energy, audio amplitude and pitch, as well as higher-level perceptual features including  
71 the presence of faces, indoor vs. outdoor scenes, written words, the presence of music and features from the fifth (final)

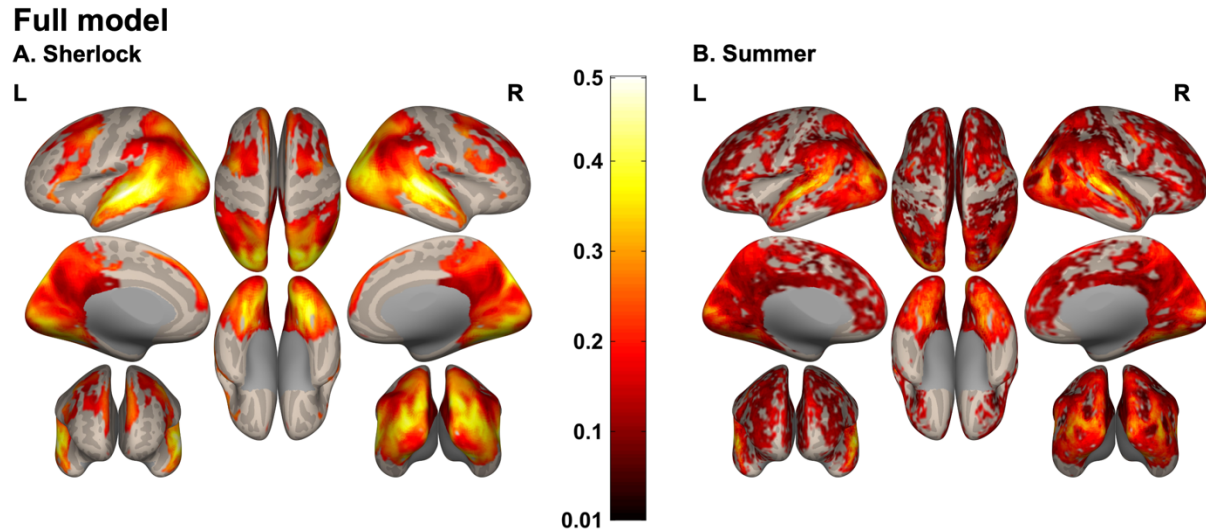
72 convolutional layer of a deep neural network (DNN) (see Movie analysis and annotations in Materials and Methods).  
 73 We included DNN features as they have recently been shown to explain a great deal of variance throughout both early  
 74 and later stages of visual cortex (Wen et al., 2018). We selected the final convolutaional layer in particular because its  
 75 activations were highly correlated to those of early convolutional layers and the final fully connected layer (see  
 76 Convolutional neural network in Materials and Methods). Social-affective features consisted of social interactions, an  
 77 agent speaking, theory of mind, perceived valence, and arousal for both studies, and additionally the touch feature for  
 78 the Summer data. We performed voxel-wise encoding analyses to learn the relationship between these features and  
 79 fMRI movie data, and then predicted held out BOLD responses using the features and beta values learned in training  
 80 (Fig 1A). We focused our analyses on voxels with shared stimuli-evoked responses across participants as measured  
 81 by inter-subject correlation (ISC, see Inter-subject brain correlation in Materials and Methods).



82  
 83 Fig 1. A. Encoding model overview. We labeled perceptual and social-affective features from a movie that participants viewed in  
 84 the MRI scanner. The features include hue, saturation, value for each pixel, motion energy, the presence/absence of written words,  
 85 indoor vs. outdoor scenes, the presence/absence of faces, audio amplitude, pitch, the presence/absence of music, the  
 86 presence/absence of social interactions, the presence/absence of an agent speaking, the presence/absence of agent talking about  
 87 others' mental states (theory of mind), perceived valence, and arousal of the scene for both movies, and additionally the social  
 88 touch feature for the Summer movie. During model training, we learned a set of beta weights linking the features to fMRI BOLD  
 89 responses over time. We then predicted the response to held-out movie data by multiplying the movie feature vectors by their  
 90 corresponding beta weights, and correlated these predictions with the actual responses extracted while participants viewed the  
 91 movie. As a result, a prediction accuracy score ( $r$ ) of the model is assigned to each voxel. B. Variance partitioning analysis overview.  
 92 Prediction accuracy  $r$  values from the full, perceptual, and social-affective models are used to calculate unique variance explained  
 93 by the perceptual or social-affective model for each voxel.

94 In both fMRI datasets, the performance of the full model, consisting of all perceptual and social-affective  
 95 features, was significantly better than chance at predicting voxel-wise responses throughout the brain. The full model  
 96 significantly predicted BOLD responses in 100% and 99.99% of voxels inside of the ISC mask for the Sherlock (range  
 97 of prediction accuracy ( $r$ ) = 0.08 ~ 0.51, mean accuracy = 0.25, standard deviation (std) = 0.07) and Summer fMRI  
 98 data (range = 0 ~ 0.43, mean = 0.17, std = 0.06), respectively. Note that prediction accuracies differ across these voxels

99 as reflected in Figure 2. The highest model performance was observed in the left STS in both studies (Peak Montreal  
100 Neurological Institute (MNI) coordinates X, Y, Z = -63, -24, -3 for Sherlock; X, Y, Z = -66, -18, 1 for Summer). High  
101 prediction accuracy was also found in the right STS (highest accuracy in right STS = 0.48 for Sherlock (X, Y, Z = 51,  
102 -33, 0); 0.43 for Summer (X, Y, Z = 67, -21, -3)).



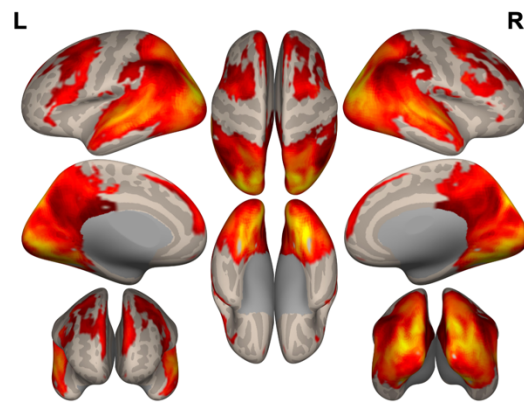
103

104 Fig 2. The full model prediction accuracy observed in Sherlock (A) and Summer (B) data. Group averaged accuracy scores are  
105 mapped on inflated cortices using the CONN software (Whitfield-Gabrieli and Nieto-Castanon, 2012) ( $P_{FDR} < 0.05$ , minimum  
106 cluster size > 10 voxels). The color bar indicates the prediction accuracy score (0 = chance, 1 = perfect prediction). The full model  
107 predicts neural responses in the bilateral STS particularly well in both studies. L = left hemisphere, R = right hemisphere, FDR =  
108 the false discovery rate.

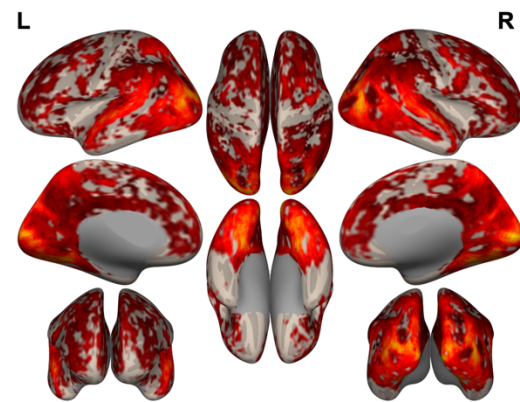
109 The perceptual model, consisting of the visual and auditory features listed above, significantly explained  
110 BOLD responses in 100% and 99.99% of voxels inside of the ISC mask in the Sherlock (range of prediction accuracy  
111 = 0.05 ~ 0.38, mean = 0.21, std = 0.06) and Summer data (range = 0 ~ 0.35, mean = 0.14, std = 0.05), respectively.  
112 The highest model performance was observed in the visual cortex in both experiments – the right fusiform gyrus (X,  
113 Y, Z = 27, -69, -12) in Sherlock and the early visual cortex (X, Y, Z = 13, -87, 7) in Summer (Fig 3).

## Perceptual model

### A. Sherlock



### B. Summer



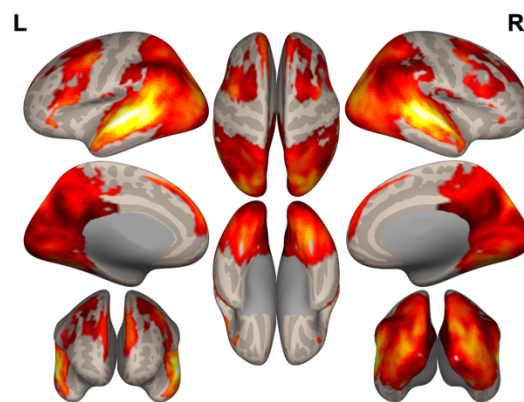
114

115 Fig 3. The perceptual model prediction accuracy observed in Sherlock (A) and Summer (B) data. The perceptual model significantly  
116 predicted neural responses in the visual cortex in both studies. Group averaged accuracy maps are thresholded at  $P_{FDR} < 0.05$  and  
117 the minimum cluster size  $> 10$  voxels.

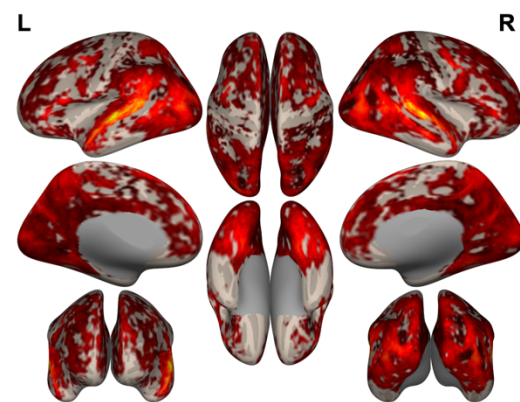
118 A social-affective model, consisting of the social-affective features listed above (an agent speaking, social  
119 interactions, theory of mind, perceived valence, and arousal) also produced significantly above chance performance  
120 throughout the whole brain in both studies. The social-affective model explained significant variance in 100% and  
121 99.92% of voxels inside of the ISC mask in the Sherlock (range of prediction accuracy = 0.04 ~ 0.55, mean = 0.20,  
122 std = 0.08) and Summer data (range = 0 ~ 0.39, mean = 0.12, std = 0.05), respectively. The highest model performance  
123 was observed in the left STS in both experiments (X, Y, Z = -60, -24, -3 for Sherlock; X, Y, Z = -63, -18, 1 for  
124 Summer). Again, findings are bilateral (highest accuracy = 0.49 in the right STS (X, Y, Z = 51, -33, 0) for Sherlock;  
125 0.38 for Summer (X, Y, Z = 64, -24, 1)) (Fig 4).

## Social-affective model

### A. Sherlock



### B. Summer

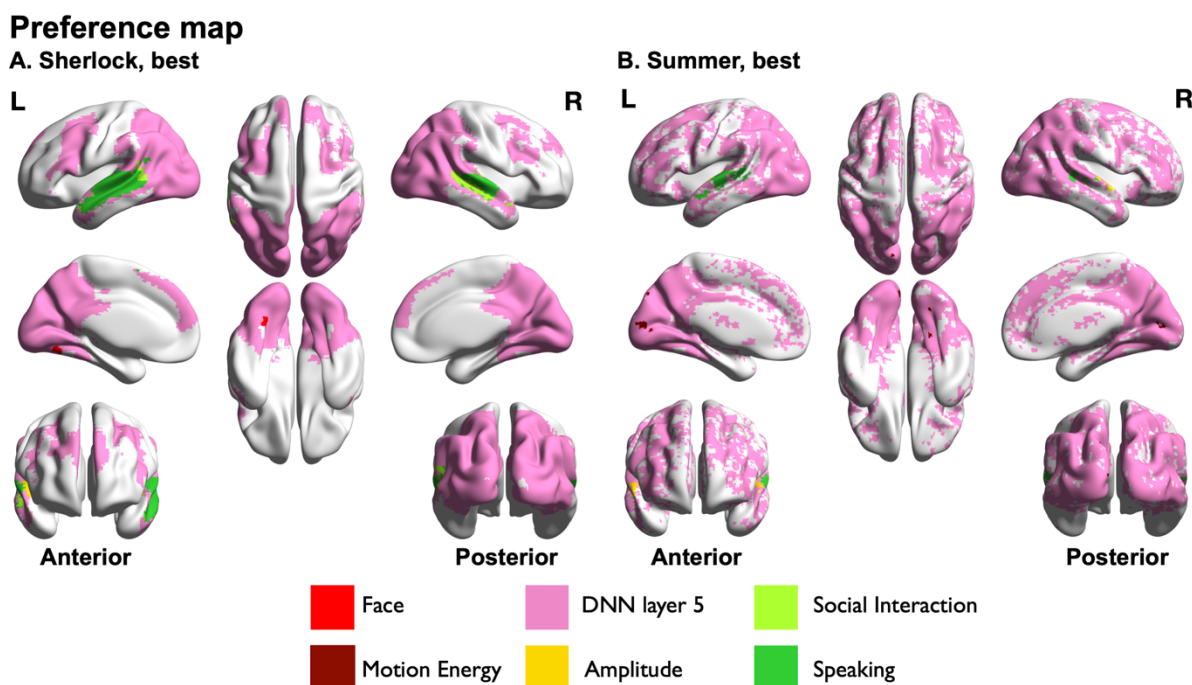


126

127 Fig 4. The social-affective model prediction accuracy observed in Sherlock (A) and Summer data (B). Group averaged accuracy  
128 scores are plotted on inflated cortices ( $P_{FDR} < 0.05$ , minimum cluster size  $> 10$  voxels). Like the full model, the social-affective  
129 model significantly predicted neural responses in the bilateral STS in both experiments.

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

132 We next performed preference mapping analyses to measure the relative contribution of each stimulus feature in  
133 predicting held out BOLD responses for every voxel. Fig 5 illustrates the winning features that best capture voxel-  
134 wise responses in each movie. The DNN fifth layer was the most predictive feature for most voxels throughout the  
135 brain (pink in Fig 5). Notably, however, this was not true in the bilateral STS where two social interaction-related  
136 features, an agent speaking (green in Fig 5) and presence of social interactions (green-yellow in Fig 5), are the most  
137 or second most preferred stimulus features in both studies (see Fig S2 for the second winning features). As expected,  
138 the auditory cortex was best explained by the audio amplitude feature (yellow-orange in Fig 5) in both studies. Visual  
139 features, i.e., motion energy, faces, and written words, are the second-best features explaining neural responses in  
140 ventral and dorsal visual pathways (red colors in Fig S2). TPJ and mPFC were second-best explained by social-  
141 affective features (theory of mind, speaking, social interaction, and arousal) (green and purple colors in Fig S2).



142  
143 Fig 5. Preference maps for the best features in each voxel for Sherlock (A) and Summer (B) data. Feature color codes indicated at  
144 bottom. DNN features best explain the neural responses throughout most of the brain in both studies, except the STS, whose neural  
145 responses are best explained by social features. Overall, social-affective features (green and purple colors) are the most or at least  
146 the second most preferred (Fig S2) in the temporal and frontal lobe in both studies, whereas visual features (red colors, Fig S2), in  
147 addition to DNN features, are preferred in the visual cortex.

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

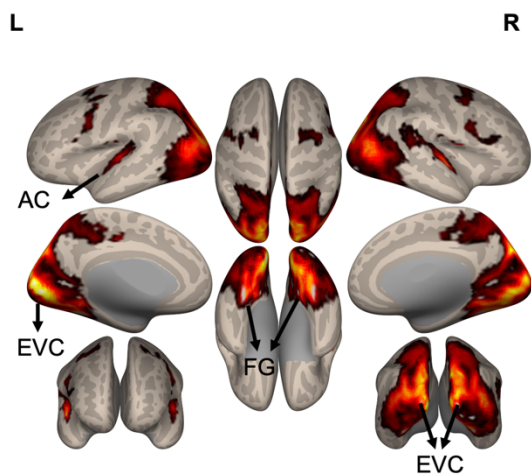
150 While above results indicate that STS and nearby regions in the temporal lobe are best explained by social features,  
151 they do not reveal to what extent perceptual and social-affective features contribute to the neural response,  
152 independently of their co-varying perceptual features. By conducting variance partitioning analyses (Fig 1B), we  
153 examined the unique contribution of the perceptual and social-affective feature models to the prediction of BOLD  
154 responses throughout the brain. This analysis is particularly important as many features are at least somewhat  
155 correlated with each other (e.g., rank correlation between social interactions and the presence of faces  $r = 0.48$  in  
156 Sherlock and  $r = 0.35$  in Summer, see Fig S1 in Supplementary Material).

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

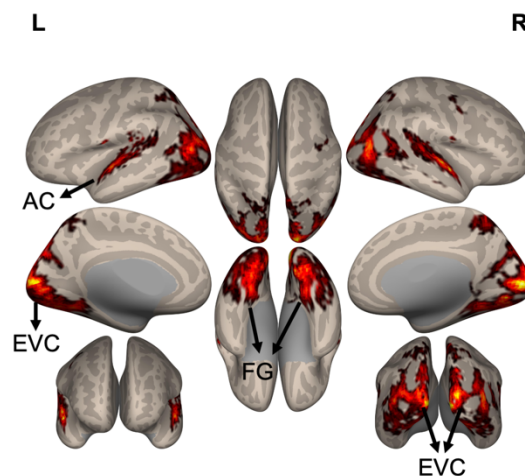
160 On the other hand, the social-affective model uniquely predicted the voxel-wise responses in high-level social  
161 cognitive regions, including the STS, temporoparietal junction (TPJ), anterior temporal lobe (ATL), and mPFC in  
162 both studies (Fig 6C-D). The largest portion of unique variance explained was in the left STS at the same or  
163 immediately adjacent MNI coordinates where the social-affective model shows the highest performance. Note that a  
164 substantial portion of the shared variance explained by both models was in the temporal lobe and occipital lobe (Fig  
165 S3). The fact that perceptual and social-affective features covary may explain this finding.

### Unique variance explained by perceptual model

#### A. Sherlock

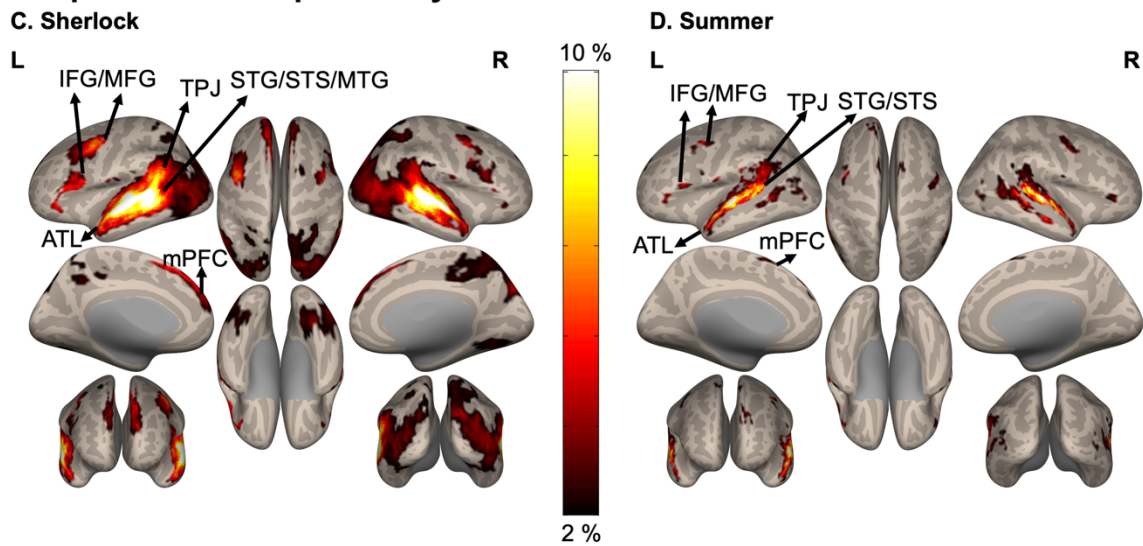


#### B. Summer



166

## Unique variance explained by social-affective model



167  
168 Fig 6. The portion of unique variance, expressed as percentages, explained by the perceptual (A, B) and social-affective model (C,  
169 D) in the Sherlock and Summer data. Color bar indicates percentage of unique variance explained. The perceptual model explained  
170 unique variance of neural responses in the visual and auditory cortices. In contrast, the social-affective model explained the unique  
171 variance of social brain responses, including the STS, TPJ, ATL, and mPFC activations. Group averaged variance partitioning  
172 maps are thresholded at  $P_{FDR} < 0.05$ , the minimum cluster size  $> 10$  voxels, and the portion of explained unique variance  $> 2\%$ .  
173 AC = auditory cortex, EVC = early visual cortex, FG = fusiform gyrus, IFG = inferior frontal gyrus, MFG = middle frontal gyrus,  
174 STG = superior temporal gyrus, STS = superior temporal sulcus, MTG = middle temporal gyrus, TPJ = temporoparietal junction,  
175 ATL = anterior temporal lobe, mPFC = medial prefrontal cortex. For labeling methods, see Variance Partitioning in Materials and  
176 Methods.

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

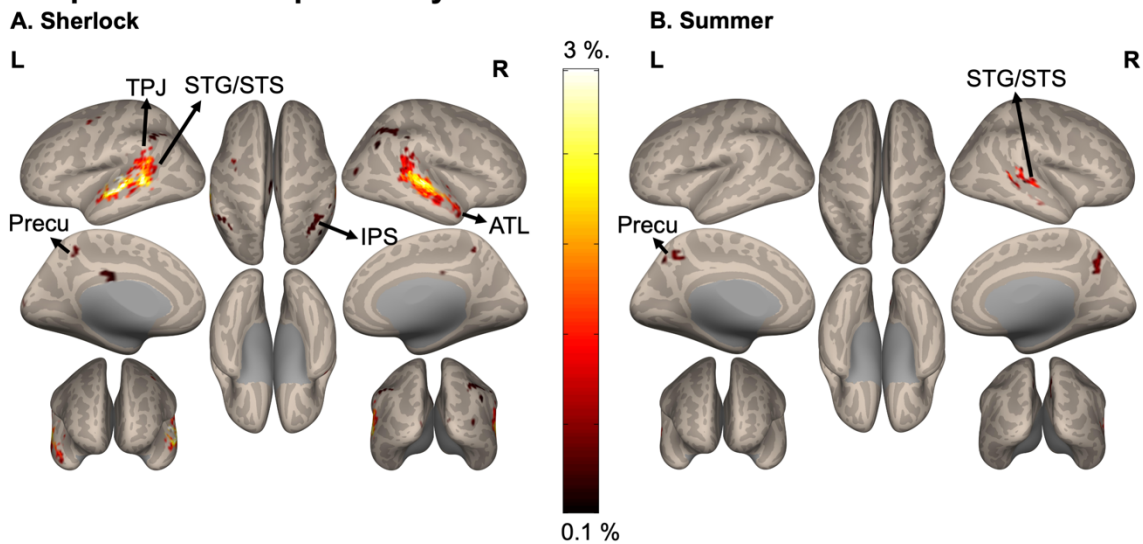
178 Finally, we sought to identify whether two major, independent (though often correlated) social features, social  
179 interactions and theory of mind, independently predict brain activity during natural movie viewing. For the Sherlock  
180 data, the presence versus absence of social interactions uniquely predicted brain responses throughout the temporal  
181 lobe and to some extent in precuneus and inferior parietal sulcus (IPS), with the bilateral STS showing the strongest  
182 selectivity (Fig 7A). The largest portion of unique contribution was observed in the left STS at the same MNI  
183 coordinates, where the social-affective model showed the highest performance. The Summer fMRI data showed  
184 similar results, although the predicted brain areas were confined to right STS and bilateral precuneus, with the right  
185 STS ( $X, Y, Z = 52, -42, 7$ ) showing the strongest selectivity (Fig 7B). This supports and extends the results of our  
186 feature preference mapping (Fig 5), highlighting the role of social interaction processing in the STS, independent of  
187 all other features, including spoken language.

188 In the Sherlock data the theory of mind feature uniquely predicted neural responses of other social brain  
189 regions, mainly located in the theory of mind network (Dufour et al., 2013) – the precuneus, mPFC, and TPJ (Fig 7C),  
190 with the precuneus showing the strongest selectivity ( $X, Y, Z = 6, -60, 51$ ). However, this distinct contribution of the  
191 theory of mind feature, observed in the Sherlock data, did not generalize to the Summer fMRI data. Only a few voxels,  
192 fewer than 10 voxels in each cluster, in STS, TPJ, and mPFC showed selectivity to the theory of mind feature (Fig



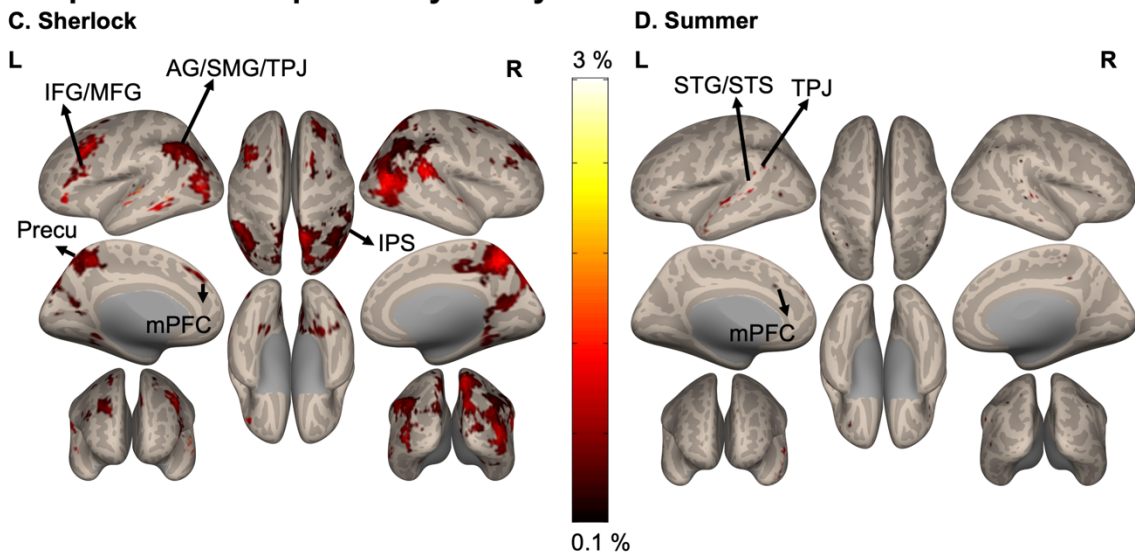
193 7D). Nonetheless, uncorrected variance partitioning results (i.e.,  $P$  uncorrected  $< 0.05$ ) included the same brain regions  
194 reported in Sherlock, such as TPJ, precuneus, and mPFC, and were largely non-overlapping with the unique social  
195 interaction voxels. In addition, no voxel showed the shared variance explained by both social interaction and theory  
196 of mind in Sherlock data. We did not run this analysis in Summer data considering the weak results found in the theory  
197 of mind feature. Overall, the results show distinct functional and anatomical divisions between social interaction and  
198 theory of mind processing in natural movies. In particular, the STS shows strong, unique selectivity to scenes with  
199 social interactions, and the precuneus, mPFC, and TPJ show selectivity for scenes where characters infer others'  
200 thoughts and emotions.

### Unique variance explained by social interaction



201

### Unique variance explained by theory of mind



202

203 Fig 7. The amount of the unique variance explained by the social interaction (A, B) and theory of mind feature (C, D) in the  
204 Sherlock and Summer studies. Color bar indicates percentage of unique variance explained. The social interaction feature explained  
205 unique variance in neural responses in the STS and precuneus. The theory of mind feature explained the unique variance of neural

206 responses in the theory of mind network, such as TPJ, precuneus, and mPFC. Group averaged variance partitioning maps are  
207 thresholded at  $P_{FDR} < 0.05$  and the minimum cluster size  $> 10$  voxels (except for the panel D, where the minimum cluster size  $> 1$   
208 voxel). STG = superior temporal gyrus, STS = superior temporal sulcus, MFG = middle frontal gyrus, ATL = anterior temporal  
209 lobe, mPFC = medial prefrontal cortex, TPJ = temporoparietal junction, AG = angular gyrus, SMG = supramarginal gyrus, IPS =  
210 intraparietal sulcus, Precu = precuneus

## 211 Discussion

212 Here we uncovered the brain mechanisms underlying naturalistic social interaction perception in an ecologically  
213 generalizable context. Combining voxel-wise encoding and variance partitioning analyses (Fig 1), we identified the  
214 brain regions showing unique selectivity for general social-affective information (Fig 6C-D) and those particularly  
215 selective to social interactions (Fig 7A-B). We also demonstrated how auditory and visual features, including audio  
216 amplitude, faces, written words, and motion energy are encoded during natural movie viewing (Fig 6A-B), replicating  
217 prior findings (Brugge et al., 2009; Cohen et al., 2002; Hart et al., 2003; Kanwisher et al., 1997; Sunaert et al., 1999)  
218 in an ecologically valid context. Importantly, our findings generalized across both sets of movie data that came from  
219 different genres, subjects, MRI scanners, and labs.

### 220 *Social-affective information processing during naturalistic movie viewing*

221 Movie viewing paradigms have recently been highlighted as essential tools in cognitive neuroscience due to their  
222 richness, dynamics, and comparable complexity to the real world (Redcay and Moraczewski, 2020; Sonkusare et al.,  
223 2019). Most prior social neuroscience studies with movies have used reverse correlation analyses to investigate  
224 phenomena such as theory of mind (Richardson et al., 2018) and social interaction perception (Wagner et al., 2016).  
225 These methods present a major advantage in that they do not require extensive movie labeling, but are prone to reverse  
226 inference errors as the event labeling happens post-hoc. This is particularly challenging in movies where the  
227 distribution of social information is often imbalanced (e.g., the majority of movie scenes contain a social interaction  
228 or a face on the screen). In the current study we densely labeled movies, using a combination of automatic and human  
229 annotations, and used the extracted movie features to perform cross-validated voxel-wise encoding analysis. This  
230 method is less prone to reverse inference error as the event coding happens prior to fMRI analysis (Redcay and  
231 Moraczewski, 2020).

232 Using these methods, we found several brain regions that coded for social-affective information – social  
233 interaction, an agent speaking, mentalizing, perceived valence, and arousal – independent of co-varying perceptual  
234 information. In particular, we identified unique variance explained in brain regions previously attributed to social  
235 perception (STS), theory of mind (TPJ, mPFC, ATL), and action observation (inferior frontal gyrus (IFG)) (Fig 6C-  
236 D). Previous studies using controlled experiments have shown an increased response in STS to social stimuli, such as  
237 point light figures (Vangeneugden et al., 2014), social touch interaction videos (Lee Masson et al., 2018), eye gaze  
238 cues (Hooker et al., 2003), affective voice/speech (Pegado et al., 2018; Wildgruber et al., 2006), and affective touch  
239 (Lee Masson et al., 2020a; Voos et al., 2013), all of which are ubiquitous in movies. The role of the theory of mind  
240 network in social cognition has also been extensively studied in social neuroscience (see the review in (Schurz et al.,  
241 2020)). This network is consistently recruited during a variety of social tasks that involve thinking about others’  
242 thoughts, emotions, intentions, and beliefs (Saxe and Kanwisher, 2003; Wolf et al., 2010). Lastly, we identified the

243 unique contribution of general social-affective features in predicting IFG activations. In the context of social  
244 processing, IFG is implicated in understanding others' actions (Carr et al., 2003). Furthermore, a stimulus showing a  
245 joint action increases its activation more than an independent action, implying its essential role in processing socially  
246 relevant actions (Centelles et al., 2011). However, compared to the above regions, the contribution of IFG in social  
247 cognition is less investigated; results are often mixed (Yang et al., 2015), and its distinctive role compared to STS  
248 remains unknown (Quadflieg and Koldewyn, 2017).

249 Surprisingly, the preferred feature across most of the brain, including many of the above social-affective  
250 regions, was the fifth layer of a DNN pre-trained on an object recognition task (Fig 5). This result is hard to interpret  
251 as most prior neuroimaging studies using DNNs have focused on their match to voxel responses in the ventral visual  
252 stream (Bonner and Epstein, 2018; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Zeman et  
253 al., 2020). Although one prior study found that object-trained DNNs well predict voxel-wise brain activity in the  
254 temporal, parietal, and frontal lobes, including the TPJ (Wen et al., 2018). The authors conjectured that later DNN  
255 layers, including the last convolutional layer 5 (the one tested here), might capture high-level semantic information in  
256 naturalistic visual scenes. Further investigation is needed to better understand the substantial contribution of late-stage  
257 DNN layers in explaining brain activations beyond the visual cortex. Excluding the DNN contribution, the results  
258 were largely confirmatory of prior studies, including faces in the fusiform gyrus, motion energy in early visual cortex  
259 and MT, and theory of mind in the TPJ (Fig S2). Intriguingly, the only region where non-DNN features were  
260 consistently preferred was the STS. The presence of a social interaction and a character speaking were the top two  
261 features in both studies. While these results were strongly suggestive of social interaction processing in the STS, it is  
262 difficult to rule out the effect of social interactions above the highly correlated effects of spoken language processing  
263 (Deen et al., 2015; Wilson et al., 2018) in the STS with this analysis alone. Thus, we turned to variance partitioning,  
264 to identify the unique contribution of social interactions in the brain.

### 265 *Social interaction perception during movie viewing*

266 In both studies, right STS and, to a lesser extent, the precuneus showed unique selectivity for naturalistic social  
267 interactions (Fig 7A-B). With the variance partitioning analysis, we were able to rule out the effect of other features,  
268 including strongly covarying faces, speaking, and theory of mind features (Fig S1). As described above, the STS  
269 processes a wide range of social features in a modality-independent manner (Deen et al., 2015), and the posterior  
270 portion of the STS has been implicated in social interaction perception in simplified contexts (Isik et al., 2017; Lee  
271 Masson et al., 2018; Walbrin et al., 2018). The current study provides the first evidence that unique variance in STS  
272 is explained by the presence/absence of social interactions in naturalistic viewing conditions. Although we replicated  
273 our findings across two movie datasets, we observed a slight discrepancy with respect to the extent of STS activity  
274 and the lateralization of this functional selectivity. Bilateral STS showed unique selectivity for social interaction in  
275 Sherlock data, whereas only right hemispheric lateralization was observed in Summer data. This discrepancy has also  
276 been observed in other studies. Some studies found the right hemispheric lateralization of STS during the observation  
277 of non-verbal social interaction, such as dyadic point-light displays and bodily movements (Isik et al., 2017; Lee  
278 Masson et al., 2018; Walbrin et al., 2018), while some did not (Lahnakoski et al., 2012; Walbrin et al., 2020; Walbrin

279 and Koldewyn, 2019). Our findings highlight that although STS is known as a hub for general social processing, social  
280 interaction is a critical feature that uniquely contributes to STS responses.

281 Unique variance in the precuneus was also explained by the presence of a social interaction, but to a lesser  
282 degree than STS. The precuneus has been implicated in a wide range of cognitive processes, such as visuo-spatial  
283 imagery, episodic memory retrieval, and social cognition (see review in (Cavanna and Trimble, 2006)). In the context  
284 of social cognition, the precuneus belongs to the theory of mind network and is recruited during false belief tasks  
285 (Jacoby et al., 2016; Saxe and Kanwisher, 2003), social trait judgment tasks (Farrer and Frith, 2002; Iacoboni et al.,  
286 2004), and while viewing social interactions (Iacoboni et al., 2004; Lahnakoski et al., 2012). A possible interpretation  
287 of our findings on the selective involvement of STS and precuneus is that these regions work in concert to  
288 spontaneously extract socially relevant information about others, such as traits of the interacting people. This notion  
289 is supported by previous studies showing increased precuneus's connectivity with other social brain areas, including  
290 STG and MTG, during the observation of dyadic interaction compared to human-object manipulation (Lee Masson et  
291 al., 2020b) and social evaluation task on others' faces (McCormick et al., 2018).

292 A prior study found a peak in the dorsal part of mPFC (dmPFC) activity during social interaction scenes of  
293 a natural movie, concluding that a viewer may spontaneously infer movie characters' thoughts and intentions (Wagner  
294 et al., 2016). Contrary to that finding, we did not observe dmPFC involvement specific to social interaction perception  
295 after controlling for the effects of perceptual and social features, including speaking and theory of mind. Instead,  
296 dmPFC was selectively tuned to more general social-affective features (Fig 6C-D) and theory of mind (Fig 7C-D).  
297 Notably, in Wagner and colleagues' study, movie scenes that evoked strong dmPFC responses were speaking scenes,  
298 and they did not consider theory of mind features in their analysis. Given that social interaction scenes genuinely  
299 invite theory of mind (Dziobek et al., 2006; Grainger et al., 2019; Roeyers et al., 2001), it may make more sense to  
300 compare their findings to our findings on general social-affective features that include speaking, social interaction,  
301 and theory of mind.

302 These results raise further questions regarding the shared and distinct role of precuneus and dmPFC in social  
303 cognition. Like the precuneus, dmPFC plays a central role in social attribution processes, including social trait  
304 judgments (Hensel et al., 2015). Given their similar functional implications in social cognition, it is unclear why we  
305 observed selectivity for social interactions in precuneus, but not dmPFC. We propose two possible interpretations.  
306 Social interaction perception may alone invite spontaneous theory of mind (precuneus activations) rather than effortful  
307 and explicit theory of mind (dmPFC activations). Other studies investigating different types of theory of mind support  
308 this interpretation (Boccardo et al., 2019; Cheong et al., 2017). Another interpretation concerns the broader role of  
309 precuneus in cognition. Given the involvement of precuneus in episodic memory retrieval (Cavanna and Trimble,  
310 2006), we cautiously conjecture that the specific recruitment of precuneus during social interaction perception may  
311 help form memory-informed social judgment during a naturalistic movie viewing. However, these two interpretations  
312 are based on reverse inference, and relatively weak social interaction response in the precuneus compared to STS. We  
313 believe interaction between social interaction perception and theory of mind processing in real-world settings is a rich  
314 area for future research.

315 *Theory of Mind during movie viewing*

316 Recent results have shown increased responses in the theory of mind network, including TPJ, precuneus, and mPFC,  
317 during movie viewing (Jacoby et al., 2016; Richardson et al., 2018). While we largely replicated these results in the  
318 Sherlock data, they did not generalize to the Summer data. This discrepancy may be related to the substantially lower  
319 signal-to-noise ratio in Summer movie data (maximum ISC value 0.73 vs. 0.56, see inter-subject brain correlation in  
320 Methods) perhaps due to the fact that the data were collected on a 1.5 T MRI scanner (Aliko et al., 2020). Another  
321 important distinction is how we annotated the theory of mind feature. Unlike previous studies (Jacoby et al., 2016;  
322 Richardson et al., 2018) which relied on reverse correlation analysis, our theory of mind annotation was labeled in  
323 advance based on whether the scene contained a person speaking about others' mental states. We believe this criterion  
324 to be more objective than trying to guess whether the subjects were engaged in mentalization. However, it may pose  
325 issues for certain movie scenes. For example, 500 Days of Summer (unlike Sherlock) contains many scenes with a  
326 narrator describing characters' mental states in voice-over monologues. We did not distinguish between scenes with  
327 the narrator and examples of mentalization that happened more naturally in the course of the movie.

## 328 Future directions

329 The methods and findings presented in this study open the door for many avenues of naturalistic social neuroscience  
330 research. Our results suggest that, even in a real-world setting, social interactions are processed in a manner that is  
331 distinct from other perceptual and social features. In both simple visual displays (Su et al., 2016) and natural images  
332 (Skripkauskaitė et al., 2021), there is an attentional bias for social interactions. Why do social interactions capture our  
333 attention and how do we use them to reason about interacting individuals? Is there a computational advantage to  
334 selectively processing social interactions? Social information has strong implications for visual neuroscience (Papeo  
335 et al., 2019; Pitcher and Ungerleider, 2021; Vestner et al., 2019). Efficient processing of social interactions seems to  
336 emerge in the visual system: two facing bodies, which occur often during and are proposed as a precursor to social  
337 interactions, are processed faster than two bodies facing away (Papeo et al., 2019; Vestner et al., 2019), evoke stronger  
338 neural responses in the body-selective visual cortex (Abassi and Papeo, 2020), and induce reciprocal brain  
339 communication between body-selective visual cortex and STS (Bellot et al., 2021). Follow-up investigations are  
340 needed to understand how these visual precursors are used to recognize real-world social interactions. Our results  
341 provide a framework to move these questions beyond controlled settings to ecologically valid experiments.

342 Finally, our method of isolating brain areas selective to naturalistic social interaction in MRI may be  
343 particularly advantageous for studies of typical and atypical development, including autism. Movies are entertaining  
344 and engaging making them an attractive stimulus for groups beyond neurotypical adults. Further, many high  
345 functioning individuals with autism look identical to neurotypical adults in simple social psychology and neuroscience  
346 tasks despite differences in their real-world social abilities (Hendriks et al., 2021; Moessnang et al., 2020; Scheeren  
347 et al., 2013). Movies may help us close the gap between simple lab-based tasks and the real world. Indeed one prior  
348 study using a naturalistic movie fMRI paradigm found lower inter-subject neural synchrony in autism (Byrge et al.,  
349 2015). The tools provided here allow us to link specific social features to brain activity in a real-world setting, opening  
350 the door to a range of new studies in autism research and social neuroscience.

## 351 Materials and Methods

### 352 fMRI data sources

353 We re-analyzed two publicly available fMRI datasets from two different studies. In the first study (Chen et al., 2017),  
354 17 participants watched the first episode of the Sherlock BBC TV series (duration ~ 45 mins) in the scanner. In the  
355 second study (Aliko et al., 2020), 86 participants watched 10 different movies from 10 genres. We selected 20  
356 participants who watched a commercial movie, 500 Days of Summer (duration ~ 90 mins). We excluded two  
357 participants from the second study as one participant (ID 14 in the original study) was scanned with a different head  
358 coil, and another participant (ID 16 in the original study) was offered glasses only after the first run.

### 359 *fMRI data acquisition and preprocessing*

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

367 We obtained preprocessed fMRI data from the authors of each original study. In the original Sherlock study,  
368 preprocessing steps included slice-timing correction, motion correction, linear detrending, temporal high-pass filtering  
369 (140s cut off), spatial normalization to a MNI space with a re-sampling size of  $3 \times 3 \times 3$  mm<sup>3</sup>, and spatial smoothing  
370 with a 6-mm full width at half maximum (FWHM) Gaussian kernel. fMRI data were also shifted by 4.5 s (3 TRs)  
371 from the stimulus onset to account for the hemodynamic delay. Lastly, timeseries BOLD signals were z-score  
372 standardized.

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

### 381 *Movie analysis and annotations*

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

390 Using MATLAB (R2020a, The Mathworks, Natick, MA) built-in functions, we extracted low-level visual  
391 features, namely hue, saturation, pixel values (HSV) and motion energy, and auditory features, namely amplitude and  
392 pitch. Specifically, we computed HSV ('rgb2hsv') and motion energy ('opticalFlowFarneback') in each pixel for all  
393 frames (640 × 360 pixels and 38 frames for Sherlock, 720 × 576 pixels and 75 frames for Summer) and averaged these  
394 values over pixels and frames. For amplitude ('audioread') and pitch ('pitch') of the audio, we averaged values over  
395 the audio samples and channels (66150 samples and two channels for Sherlock and 132300 samples and two channels  
396 for Summer). This computation allowed us to obtain one value per feature for each video segment (the duration of 1  
397 TR).

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

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

409 Social features were labeled by two human raters – whether or not a video segment contains social  
410 interactions (presence of social interactions = 1, absence = 0), whether or not a person speaks in the scene (yes = 1,  
411 no = 0), and whether or not a person infers mental states of others (yes = 1, no = 0). The annotation of the theory of  
412 mind feature is based on whether a movie character inferring other characters' thoughts and emotions in each scene  
413 (an example in Sherlock – Ms. Patterson is seated at a table at a press conference, reading her statement saying: "He  
414 loved his family and his work"; an example in Summer – The main character seen sitting next to other colleagues at  
415 a meeting and a narrator says: "The boy, Tom Hansen of Margate of New Jersey, grew up believing that he'd never  
416 truly be happy until the day he met the one."). We selected this criterion to be as objective as possible, and to avoid  
417 raters guessing about whether the subjects were engaged in mentalization. This type of second-order theory of mind  
418 task activates theory of mind network (Tholen et al., 2020). As in prior studies (Saxe and Powell, 2006), the description  
419 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  
420 not considered theory of mind. For Summer, we additionally included the touch feature – whether or not a person  
421 makes physical contact with another person (social touch = 1), him(her)self or an object (nonsocial touch = -1), or not  
422 (absence of touch = 0). We were unable to include the touch feature in the Sherlock data as there were less than 10  
423 video segments displaying social touch.

424 Considering that the time-scale of high-level cognitive events is longer than a low-level perceptual event  
425 (Baldassano et al., 2017), we merged each pair of consecutive Sherlock 1.5s length segments into one for annotations  
426 of social features, resulting in 3s length segments for both studies.

427 Lastly, we added valence and arousal ratings, offered by the authors of another Sherlock fMRI study (Kim  
428 et al., 2020). In their study, 4.5s video segments, parsed from the full-length Sherlock episode, were rated by 113  
429 participants with a 1 – 9 Likert scale. 35 participants rated all video segments, and the remaining participants rated  
430 only a quarter, yielding about 55 ratings per video segment. Group mean valence and arousal ratings were used in the  
431 current study. For Summer segments, four human raters judged how pleasant and arousing the scene is using a 1 – 5  
432 Likert scale. Despite the small number of raters, inter-rater consistencies are relatively high (valence Spearman  $r$  ( $r_s$ )  
433 = 0.62, arousal  $r_s$  = 0.35), compared to that (valence  $r$  = 0.30) reported in the Sherlock study (Kim et al., 2020). The  
434 figure illustrating correlations between features can be found in supplementary material (Fig S1).

### 435 *Convolutional neural network*

436 We used a deep convolutional neural network (CNN) to extract visual features from both stimuli. Specifically,  
437 PyTorch (version 1.4.0) implementation of AlexNet (Krizhevsky et al., 2012), pretrained on the ImageNet dataset  
438 (Russakovsky et al., 2015), was used. AlexNet consists of eight layers with five convolutional layers and three fully  
439 connected layers. AlexNet in PyTorch adopted 64, 192, 384, 256, and 256 kernels from the first through fifth layers  
440 instead of 96, 256, 384, 384, and 256 reported in the original study (Krizhevsky et al., 2012). The size of the kernels  
441 is the same as original AlexNet –  $11 \times 11$ ,  $5 \times 5$ ,  $3 \times 3$ ,  $3 \times 3$ ,  $3 \times 3$  from the 1<sup>st</sup> to 5<sup>th</sup> layers, respectively.

442 The first frame of each video segment was, cropped, normalized, and fed into the first layer of AlexNet, with  
443 the image input size  $3 \times 224 \times 224$ . The output of the previous layer served as the input of the following layer. We  
444 captured visual features of each video segment by taking the activations of all units from the fifth layer right before  
445 the rectified linear activation function and the max-pooling layer during the forward pass. We selected the fifth layer  
446 since activations were highly correlated to those of early ( $r > 0.75$  with second layer) and late layers ( $r > 0.5$  seventh  
447 layer). The output size of the fifth layer is  $256 \times 13 \times 13$ , which was flattened to an array with 43264 elements. Unlike  
448 other features ( $1 \times$  the number of video segments), the features from the DNN layer are high dimensional (i.e.,  $43264$   
449  $\times$  the number of video segments). We reduced the dimensionality without losing crucial information by applying  
450 principal component analysis (PCA) to the DNN activations. Components first through  $N^{\text{th}}$  were selected until the  
451 amount of explained variance reached 70%. 147 and 150 components were added to the feature matrix of the fifth  
452 layer for Sherlock and Summer segments, respectively.

453 In summary, perceptual features consist of components from the DNN fifth layer, HSV, motion energy, faces,  
454 indoor/outdoor, written words, amplitude, pitch, and music. Social-affective features consist of social interactions,  
455 speaking, theory of mind, valence, and arousal for both experiments, and additionally the touch feature for the Summer  
456 data.

### 457 *Inter-subject brain correlation*

458 ISC analysis was implemented to create a brain mask containing voxels showing shared stimuli-evoked responses  
459 across participants. ISC is a well-validated fMRI method that identifies voxels with reliable neural responses across  
460 time while rejecting idiosyncratic and noisy voxels (Nastase et al., 2019). Using the brain imaging analysis kit's  
461 (BrainIAK, version 0.1.0) built-in functions ('isc', 'permutation\_isc'), we measured ISC with a leave-one-subject-out  
462 approach. In other words, for every voxel, time-series BOLD responses for all but one subject were averaged and  
463 correlated with that of the remaining subject. We repeated this process for every participant. Fisher-Z transformed



464 correlation values were averaged across folds. Encoding model analysis was masked to include only voxels with ISC  
465 value  $> 0.25$  in the case of the Sherlock dataset, based on the previous permutation results ( $P_{FDR} < 0.001$ ) (Baldassano  
466 et al., 2017). For the Summer dataset, we ran a sign permutation test (5,000 iterations) across participants, as  
467 implemented in BrainIAK, and created a mask consisting of voxels passing the statistical threshold,  $P_{FDR} < 0.005$   
468 after the multiple comparison correction with the FDR. In other words, subjects' ISC values were randomly multiplied  
469 with +1 or -1 for 5,000 times, which resulted in empirical null distribution of ISC values for each voxel. From this  
470 distribution, two-tailed P values were calculated and adjusted with FDR correction. We used a less stringent threshold  
471 ( $P_{FDR} < 0.005$  as opposed to  $P_{FDR} < 0.001$ ) on Summer ISC to have a mask with the similar number of voxels with  
472 that of Sherlock (25468 and 22044 voxels, respectively). Voxels outside of the mask or whose BOLD responses did  
473 not change over time were excluded from the further encoding analyses.

#### 474 *Voxel-wise encoding modeling*

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

487 Data were first split into 10 folds. Among them, nine folds were used to estimate beta weights and optimize  
488 the regularization parameter (range  $0.1 \sim 10,000$ )  $\lambda_1$  and  $\lambda_2$ . Optimal  $\lambda$  values were selected per voxel via inner loop  
489 5-fold cross-validation. In other words, training data from nine folds were again split into five folds. Four folds were  
490 used for the model estimation with various  $\lambda$  values, and unseen data from the remaining fold in the inner loop were  
491 used for selecting the  $\lambda$  values that, on average, yielded the smallest mismatch between predicted and actual BOLD  
492 responses. After the model estimation and regularization parameter optimization, unseen data from the remaining  
493 tenth fold in the outer loop were used for evaluating the performance of the model. We measured model performance  
494 by computing the correlation between predicted and actual BOLD responses. For each voxel separately, this process  
495 was repeated 10 times, and the performance of the model was averaged over the repetitions.

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

501  $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$  (1)

502

503 , where  $\beta_1$  and  $\beta_2$  are a  $F_1$  or  $F_2$  sized array containing the beta weights for each feature.

504

505 
$$\hat{\beta}_{\text{banded ridge}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left( \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} + \begin{bmatrix} \lambda_1^2 I_{F_1} & 0 \\ 0 & \lambda_2^2 I_{F_2} \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} Y$$
 (2)

506

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

513

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

515

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

517

518 where  $Y_{\text{predicted}}$  is an array of size  $T_e$  (number of total fMRI volumes from the testing set) consisting of estimated  
519 BOLD signals. Let  $X = [X_1 \ X_2]$ , 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  
520 calculate correlation coefficient value between the predicted and actual BOLD signal to evaluate the model  
521 performance. A high  $r$  value can be interpreted as high prediction accuracy.

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

524 1) A full model includes all features listed in the sub-section (movie analysis and annotations). In this model,  
525  $X_1$  is composed of high-dimensional DNN components and  $X_2$  is composed of the rest. Beta weights  
526 estimated from this model were used in preference mapping analysis described below.

527 2) A perceptual model includes all visual and auditory features.  $X_1$  is composed of DNN components and  
528  $X_2$  is composed of the rest of visual and auditory features.

529 3) A social-affective model includes all social-affective features. In the social-affective model, the feature  
530 space is low-dimensional, so classic ridge regression (where  $\lambda_1 = \lambda_2$ ) was used. Thus,  $X$  is composed  
531 of all social-affective features.

532 We calculated the prediction accuracy map across participants for every model. Random-effect group-level analyses  
533 were conducted with a nonparametric permutation test to identify voxels showing significantly above chance  
534 prediction accuracy. Before the permutation test, we removed voxels that were not shared across participants, which  
535 resulted in the final number of voxels being slightly less than the total number of voxels contained in the mask (21649  
536 voxels for Sherlock; 24477 voxels for Summer). Similar to previous studies (Cichy et al., 2017; Hebart et al., 2018),

537 and the above-mentioned ISC analysis, we conducted a sign permutation test (5,000 iterations). From the empirical  
538 null distribution of a prediction accuracy, one-tailed P values were calculated and adjusted with FDR correction. Group  
539 averaged prediction accuracy maps of each model were thresholded at  $P_{\text{FDR}} < 0.05$  and plotted on the cortical surface.

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

542 4) One model includes all features except the presence of a social interaction

543 5) The final model includes all features except the mentalization feature.

544 Prediction accuracies obtained from these five models were used in variance partitioning analysis. All code for voxel-  
545 wise encoding and variance partitioning analysis associated with the current study is available at  
546 [https://github.com/haemyleemasson/voxelwise\\_encoding](https://github.com/haemyleemasson/voxelwise_encoding).

### 547 *Preference Mapping*

548 We performed preference mapping analysis to visualize the stimulus feature that were best at explaining each voxel's  
549 activation. Beta weights generated from a full model were used to predict the withheld BOLD responses for each  
550 voxel via a cross-validation as described above. During the testing session, we used the beta weight(s) of a feature of  
551 interest (e.g., the amplitude of audio) to estimate unseen BOLD responses while assigning 0 values to beta weights of  
552 other features. This method is superior to having a separate model for each feature when evaluating the relative  
553 contribution of each individual feature in predicting voxel activations (Nunez-Elizalde et al., 2019). Moreover, this  
554 method is suitable for examining the prediction accuracy for high-dimensional feature spaces, such as DNN units,  
555 which produces hundreds of beta weights. We repeated this procedure for every feature and participant. In the end,  
556 the winning feature that yielded the highest group averaged prediction accuracy were selected for each voxel. We  
557 colored every voxel to reflect which feature predicts BOLD responses of that voxel the best. The first (Fig 5) and  
558 second preference maps (Fig S2) were plotted on the cortical surface.

### 559 *Variance Partitioning*

560 For every voxel, we conducted a variance partitioning analysis to determine the unique contribution of each encoding  
561 model in predicting the BOLD responses in withheld fMRI data. To this end, we compared the amount of variance  
562 explained by a perceptual or social-affective model with that of a full model consisting of all features. We also  
563 measured the unique contribution of two important social features, the presence of a social interaction and  
564 mentalization, in predicting the BOLD responses (for shared variance, see Supplementary Material). The amount of  
565 unique variance explained by a model/feature of interest was calculated as:

$$566 \quad U_{\text{voi}} = r_X^2 - r_{X-\text{voi}}^2$$

567 X reflects all features listed in the sub-section (movie analysis and annotations). VOI reflects a variable of interest  
568 (e.g., all social-affective features or a mentalization feature).  $r^2$  is the squared prediction accuracy value obtained from  
569 the encoding model, which can be interpreted as the variance explained by a model consisting of either all features X  
570 or all features except the variable of interest X – VOI.  $U_{\text{voi}}$  is the amount of unique variance explained by VOI. We  
571 computed  $U_{\text{voi}}$  for every participant and voxel. Thus,  $U_{\text{voi}}$  is an  $N \times V$  sized matrix where N reflects the total number  
572  
573

574 of participants and V reflects the total number of voxels included in the encoding model analysis for each dataset (i.e.,  
575  $N=17$  and Voxels = 21649 for Sherlock;  $N=18$  and Voxels = 24477 for Summer). The higher the U value is, the more  
576 the variance is uniquely explained by a variable of interest. Like voxel-wise encoding, we applied the sign permutation  
577 test (5,000 iterations) for the statistical inference. Group averaged maps, resulting from U values, were thresholded at  
578  $P_{FDR} < 0.05$  and plotted on the cortical surface. Most brain areas were labeled with automated anatomical labeling  
579 atlas (Tzourio-Mazoyer et al., 2002). The location of STS and TPJ were compared with the templates created from  
580 the previous studies (Deen et al., 2015; Mars et al., 2012).

## 581 Acknowledgements

582 We thank Janice Chen, Sarah Aliko, Jongwan Kim and their colleagues for data and annotations used in this study.  
583 We would like to thank J. Brendan Ritchie and Alon Hafri for valuable comments.

## 584 CRediT authorship contribution statement

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

## 588 References

- 589 Abassi E, Papeo L. 2020. The representation of two-body shapes in the human visual cortex. *J Neurosci* **40**:852–  
590 863. doi:10.1523/JNEUROSCI.1378-19.2019
- 591 Aliko S, Huang J, Gheorghiu F, Meliss S, Skipper JI. 2020. A naturalistic neuroimaging database for understanding  
592 the brain using ecological stimuli. *Sci Data* **7**:1–21. doi:10.1038/s41597-020-00680-2
- 593 Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. 2017. Discovering Event Structure in  
594 Continuous Narrative Perception and Memory. *Neuron* **95**:709-721.e5. doi:10.1016/j.neuron.2017.06.041
- 595 Bellot E, Abassi E, Papeo L. 2021. Moving Toward versus Away from Another: How Body Motion Direction  
596 Changes the Representation of Bodies and Actions in the Visual Cortex. *Cereb Cortex*.  
597 doi:10.1093/cercor/bhaa382
- 598 Boccadoro S, Cracco E, Hudson AR, Bardi L, Nijhof AD, Wiersema JR, Brass M, Mueller SC. 2019. Defining the  
599 neural correlates of spontaneous theory of mind (ToM): An fMRI multi-study investigation. *Neuroimage*  
600 **203**:116193. doi:10.1016/j.neuroimage.2019.116193
- 601 Bonner MF, Epstein RA. 2018. Computational mechanisms underlying cortical responses to the affordance  
602 properties of visual scenes. *PLOS Comput Biol* **14**:e1006111. doi:10.1371/journal.pcbi.1006111
- 603 Brugge JF, Nourski K V., Oya H, Reale RA, Kawasaki H, Steinschneider M, Howard MA. 2009. Coding of  
604 repetitive transients by auditory cortex on Heschl's gyrus. *J Neurophysiol* **102**:2358–2374.  
605 doi:10.1152/jn.91346.2008
- 606 Byrge L, Dubois J, Tyszka JM, Adolphs R, Kennedy DP. 2015. Idiosyncratic brain activation patterns are associated  
607 with poor social comprehension in autism. *J Neurosci* **35**:5837–5850. doi:10.1523/JNEUROSCI.5182-14.2015
- 608 Carr L, Iacoboni M, Dubeaut MC, Mazziotta JC, Lenzi GL. 2003. Neural mechanisms of empathy in humans: A  
609 relay from neural systems for imitation to limbic areas. *Proc Natl Acad Sci U S A* **100**:5497–5502.

- 610 doi:10.1073/pnas.0935845100
- 611 Cavanna AE, Trimble MR. 2006. The precuneus: a review of its functional anatomy and behavioural correlates.
- 612 *Brain* **129**:564–583. doi:10.1093/brain/awl004
- 613 Centelles L, Assaiante C, Nazarian B, Anton J-L, Schmitz C. 2011. Recruitment of Both the Mirror and the
- 614 Mentalizing Networks When Observing Social Interactions Depicted by Point-Lights: A Neuroimaging Study.
- 615 *PLoS One* **6**:e15749. doi:10.1371/journal.pone.0015749
- 616 Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, Hasson U. 2017. Shared memories reveal shared structure in
- 617 neural activity across individuals. *Nat Neurosci* **20**:115–125. doi:10.1038/nn.4450
- 618 Cheong JH, Jolly E, Sul S, Chang LJ. 2017. Computational models in social neuroscience. *Comput Model brain*
- 619 *Behav* 229–244.
- 620 Cichy RM, Khosla A, Pantazis D, Oliva A. 2017. Dynamics of scene representations in the human brain revealed by
- 621 magnetoencephalography and deep neural networks. *Neuroimage* **153**:346–358.
- 622 doi:10.1016/j.neuroimage.2016.03.063
- 623 Cohen L, Lehericy S, Chochon F, Lemer C, Rivaud S, Dehaene S. 2002. Language-specific tuning of visual cortex?
- 624 Functional properties of the Visual Word Form Area. *Brain* **125**:1054–1069. doi:10.1093/brain/awf094
- 625 Deen B, Koldewyn K, Kanwisher N, Saxe R. 2015. Functional organization of social perception and cognition in the
- 626 superior temporal sulcus. *Cereb Cortex* **25**:4596–4609. doi:10.1093/cercor/bhv111
- 627 Dufour N, Redcay E, Young L, Mavros PL, Moran JM, Triantafyllou C, Gabrieli JDE, Saxe R. 2013. Similar Brain
- 628 Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS One*
- 629 **8**:e75468. doi:10.1371/journal.pone.0075468
- 630 Dziobek I, Fleck S, Kalbe E, Rogers K, Hassenstab J, Brand M, Kessler J, Woike JK, Wolf OT, Convit A. 2006.
- 631 Introducing MASC: A movie for the assessment of social cognition. *J Autism Dev Disord* **36**:623–636.
- 632 doi:10.1007/s10803-006-0107-0
- 633 Farrer C, Frith CD. 2002. Experiencing oneself vs another person as being the cause of an action: The neural
- 634 correlates of the experience of agency. *Neuroimage* **15**:596–603. doi:10.1006/nimg.2001.1009
- 635 Grainger SA, Steinvik HR, Henry JD, Phillips LH. 2019. The role of social attention in older adults' ability to
- 636 interpret naturalistic social scenes. *Q J Exp Psychol* **72**:1328–1343. doi:10.1177/1747021818791774
- 637 Güçlü U, van Gerven MAJ. 2015. Deep neural networks reveal a gradient in the complexity of neural
- 638 representations across the ventral stream. *J Neurosci* **35**:10005–10014. doi:10.1523/JNEUROSCI.5023-
- 639 14.2015
- 640 Hamlin JK. 2015. The case for social evaluation in preverbal infants: gazing toward one's goal drives
- 641 infants' preferences for Helpers over Hinderers in the hill paradigm. *Front Psychol* **5**:1563.
- 642 doi:10.3389/fpsyg.2014.01563
- 643 Hamlin JK, Wynn K. 2011. Young infants prefer prosocial to antisocial others. *Cogn Dev* **26**:30–39.
- 644 doi:10.1016/j.cogdev.2010.09.001
- 645 Hamlin JK, Wynn K, Bloom P. 2007. Social evaluation by preverbal infants. *Nature* **450**:557–559.
- 646 doi:10.1038/nature06288

- 647 Hart H, Palmer A, Hall D. 2003. Amplitude and Frequency-modulated Stimuli Activate Common Regions of  
648 Human Auditory Cortex. *Cereb Cortex* **13**:773–781. doi:10.1093/cercor/13.7.773
- 649 Hasson U, Malach R, Heeger DJ. 2010. Reliability of cortical activity during natural stimulation. *Trends Cogn Sci*.  
650 doi:10.1016/j.tics.2009.10.011
- 651 Haxby J V., Gobbini MI, Nastase SA. 2020. Naturalistic stimuli reveal a dominant role for agentic action in visual  
652 representation. *Neuroimage* **216**:116561. doi:10.1016/j.neuroimage.2020.116561
- 653 Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM. 2018. The representational dynamics of task and object  
654 processing in humans. *Elife* **7**. doi:10.7554/eLife.32816
- 655 Hendriks MHA, Dillen C, Vettori S, Vercammen L, Daniels N, Steyaert J, Op de Beeck H, Boets B. 2021. Neural  
656 processing of facial identity and expression in adults with and without autism: A multi-method approach.  
657 *NeuroImage Clin* **29**:102520. doi:10.1016/j.nicl.2020.102520
- 658 Hensel L, Bzdok D, Müller VI, Zilles K, Eickhoff SB. 2015. Neural correlates of explicit social judgments on vocal  
659 stimuli. *Cereb Cortex* **25**:1152–1162. doi:10.1093/cercor/bht307
- 660 Hooker CI, Paller KA, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ. 2003. Brain networks for analyzing eye  
661 gaze. *Cogn Brain Res* **17**:406–418. doi:10.1016/S0926-6410(03)00143-5
- 662 Iacoboni M, Lieberman MD, Knowlton BJ, Molnar-Szakacs I, Moritz M, Throop CJ, Fiske AP. 2004. Watching  
663 social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases  
664 compared to a resting baseline. *Neuroimage* **21**:1167–1173. doi:10.1016/j.neuroimage.2003.11.013
- 665 Isik L, Koldewyn K, Beeler D, Kanwisher N. 2017. Perceiving social interactions in the posterior superior temporal  
666 sulcus. *Proc Natl Acad Sci* 201714471. doi:10.1073/pnas.1714471114
- 667 Jacoby N, Bruneau E, Koster-Hale J, Saxe R. 2016. Localizing Pain Matrix and Theory of Mind networks with both  
668 verbal and non-verbal stimuli. *Neuroimage* **126**:39–48. doi:10.1016/j.neuroimage.2015.11.025
- 669 Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: A module in human extrastriate cortex  
670 specialized for face perception. *J Neurosci* **17**:4302–4311. doi:10.1523/jneurosci.17-11-04302.1997
- 671 Khaligh-Razavi S-M, Kriegeskorte N. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT  
672 Cortical Representation. *PLoS Comput Biol* **10**:e1003915. doi:10.1371/journal.pcbi.1003915
- 673 Kim J, Weber CE, Gao C, Schulteis S, Wedell DH, Shinkareva S V. 2020. A study in affect: Predicting valence  
674 from fMRI data. *Neuropsychologia* **143**:107473. doi:10.1016/j.neuropsychologia.2020.107473
- 675 Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural  
676 networks *Advances in Neural Information Processing Systems*. pp. 1097–1105.
- 677 Lahnakoski JM, Glerean E, Salmi J, Jääskeläinen IP, Sams M, Hari R, Nummenmaa L. 2012. Naturalistic fMRI  
678 Mapping Reveals Superior Temporal Sulcus as the Hub for the Distributed Brain Network for Social  
679 Perception. *Front Hum Neurosci* **6**:233. doi:10.3389/fnhum.2012.00233
- 680 Lee Masson H, Op de Beeck H, Boets B. 2020a. Reduced task-dependent modulation of functional network  
681 architecture for positive versus negative affective touch processing in autism spectrum disorders. *Neuroimage*.  
682 doi:10.1016/j.neuroimage.2020.117009
- 683 Lee Masson H, Pillet I, Boets B, Op de Beeck H. 2020b. Task-dependent changes in functional connectivity during

- 684 the observation of social and non-social touch interaction. *Cortex*. doi:10.1016/j.cortex.2019.12.011
- 685 Lee Masson H, Van De Plas S, Daniels N, Op de Beeck H. 2018. The multidimensional representational space of  
686 observed socio-affective touch experiences. *Neuroimage* **175**:297–314. doi:10.1016/j.neuroimage.2018.04.007
- 687 Mars RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS. 2012. Connectivity-based subdivisions of the  
688 human right “temporoparietal junction area”: Evidence for different areas participating in different cortical  
689 networks. *Cereb Cortex* **22**:1894–1903. doi:10.1093/cercor/bhr268
- 690 McCormick EM, van Hoorn J, Cohen JR, Telzer EH. 2018. Functional connectivity in the social brain across  
691 childhood and adolescence. *Soc Cogn Affect Neurosci* **13**:819–830. doi:10.1093/scan/nsy064
- 692 Moessnang C, Baumeister S, Tillmann J, Goyard D, Charman T, Ambrosino S, Baron-Cohen S, Beckmann C, Bölte  
693 S, Bours C, Crawley D, Dell’Acqua F, Durston S, Ecker C, Frouin V, Hayward H, Holt R, Johnson M, Jones  
694 E, Lai MC, Lombardo M V., Mason L, Oldenhinkel M, Persico A, Cáceres ASJ, Spooren W, Loth E, Murphy  
695 DGM, Buitelaar JK, Banaschewski T, Brandeis D, Tost H, Meyer-Lindenberg A. 2020. Social brain activation  
696 during mentalizing in a large autism cohort: The Longitudinal European Autism Project. *Mol Autism* **11**:17.  
697 doi:10.1186/s13229-020-0317-x
- 698 Nastase SA, Gazzola V, Hasson U, Keysers C. 2019. Measuring shared responses across subjects using intersubject  
699 correlation. *Soc Cogn Affect Neurosci* **14**:669–687. doi:10.1093/scan/nsz037
- 700 Nastase SA, Goldstein A, Hasson U. 2020. Keep it real: rethinking the primacy of experimental control in cognitive  
701 neuroscience. *Neuroimage* **222**. doi:10.1016/j.neuroimage.2020.117254
- 702 Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. 2011. Reconstructing visual experiences from  
703 brain activity evoked by natural movies. *Curr Biol* **21**:1641–1646. doi:10.1016/j.cub.2011.08.031
- 704 Nunez-Elizalde AO, Huth AG, Gallant JL. 2019. Voxelwise encoding models with non-spherical multivariate  
705 normal priors. *Neuroimage* **197**:482–492. doi:10.1016/j.neuroimage.2019.04.012
- 706 Papeo L, Goupil N, Soto-Faraco S. 2019. Visual Search for People Among People. *Psychol Sci* **30**:1483–1496.  
707 doi:10.1177/0956797619867295
- 708 Pegado F, Hendriks MHA, Amelynck S, Daniels N, Bulthé J, Lee Masson H, Boets B, Op de Beeck H. 2018. A  
709 Multitude of Neural Representations Behind Multisensory “Social Norm” Processing. *Front Hum Neurosci*  
710 **12**:153. doi:10.3389/fnhum.2018.00153
- 711 Petrini K, Piwek L, Crabbe F, Pollick FE, Garrod S. 2014. Look at those two!: The precuneus role in unattended  
712 third-person perspective of social interactions. *Hum Brain Mapp* **35**:5190–5203. doi:10.1002/hbm.22543
- 713 Pitcher D, Ungerleider LG. 2021. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends*  
714 *Cogn Sci*. doi:10.1016/j.tics.2020.11.006
- 715 Quadflieg S, Koldewyn K. 2017. The neuroscience of people watching: how the human brain makes sense of other  
716 people’s encounters. *Ann N Y Acad Sci* **1396**:166–182. doi:10.1111/nyas.13331
- 717 Redcay E, Moraczewski D. 2020. Social cognition in context: A naturalistic imaging approach. *Neuroimage*  
718 **216**:116392. doi:10.1016/j.neuroimage.2019.116392
- 719 Richardson H. 2019. Development of brain networks for social functions: Confirmatory analyses in a large open  
720 source dataset. *Dev Cogn Neurosci* **37**:100598. doi:10.1016/j.dcn.2018.11.002

- 721 Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. 2018. Development of the social brain from age three to  
722 twelve years. *Nat Commun* **9**:1027. doi:10.1038/s41467-018-03399-2
- 723 Roeyers H, Buysse A, Ponnet K, Pichal B. 2001. Advancing Advanced Mind-reading Tests: Empathic Accuracy in  
724 Adults with a Pervasive Developmental Disorder. *J Child Psychol Psychiatry* **42**:271–278. doi:10.1111/1469-  
725 7610.00718
- 726 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg  
727 AC, Fei-Fei L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* **115**:211–252.  
728 doi:10.1007/s11263-015-0816-y
- 729 Saxe R, Kanwisher N. 2003. People thinking about thinking people: The role of the temporo-parietal junction in  
730 “theory of mind.” *Neuroimage* **19**:1835–1842. doi:10.1016/S1053-8119(03)00230-1
- 731 Saxe R, Powell LJ. 2006. It’s the thought that counts: Specific brain regions for one component of theory of mind.  
732 *Psychol Sci* **17**:692–699. doi:10.1111/j.1467-9280.2006.01768.x
- 733 Scheeren AM, De Rosnay M, Koot HM, Begeer S. 2013. Rethinking theory of mind in high-functioning autism  
734 spectrum disorder. *J Child Psychol Psychiatry Allied Discip* **54**:628–635. doi:10.1111/jcpp.12007
- 735 Schurz M, Radua J, Tholen MG, Maliske L, Margulies DS, Mars RB, Sallet J, Kanske P. 2020. Toward a  
736 hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and  
737 theory of mind. *Psychol Bull* **undefined**:undefined. doi:10.1037/bul0000303
- 738 Skripkauskaitė S, Mihai I, Koldewyn K. 2021. ATTENTION TO SOCIAL INTERACTIONS 1 Brief report:  
739 Attentional Bias towards Social Interactions during Viewing of Naturalistic Scenes. *bioRxiv*  
740 2021.02.26.433078. doi:10.1101/2021.02.26.433078
- 741 Sonkusare S, Breakspear M, Guo C. 2019. Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends Cogn*  
742 *Sci*. doi:10.1016/j.tics.2019.05.004
- 743 Su J, van Boxtel JJA, Lu H. 2016. Social Interactions Receive Priority to Conscious Perception. *PLoS One*  
744 **11**:e0160468. doi:10.1371/journal.pone.0160468
- 745 Sunaert S, Van Hecke P, Marchal G, Orban GA. 1999. Motion-responsive regions of the human brain. *Exp Brain*  
746 *Res* **127**:355–370. doi:10.1007/s002210050804
- 747 Tholen MG, Trautwein FM, Böckler A, Singer T, Kanske P. 2020. Functional magnetic resonance imaging (fMRI)  
748 item analysis of empathy and theory of mind. *Hum Brain Mapp* **41**:2611–2628. doi:10.1002/hbm.24966
- 749 Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. 2002.  
750 Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI  
751 MRI single-subject brain. *Neuroimage* **15**:273–289. doi:10.1006/nimg.2001.0978
- 752 Vangeneugden J, Peelen M V., Tadin D, Battelli L. 2014. Distinct Neural Mechanisms for Body Form and Body  
753 Motion Discriminations. *J Neurosci* **34**:574–585. doi:10.1523/JNEUROSCI.4032-13.2014
- 754 Vestner T, Tipper SP, Hartley T, Over H, Rueschemeyer SA. 2019. Bound together: Social binding leads to faster  
755 processing, spatial distortion, and enhanced memory of interacting partners. *J Exp Psychol Gen* **148**:1251–  
756 1268. doi:10.1037/xge0000545
- 757 Voos AC, Pelphey KA, Kaiser MD. 2013. Autistic traits are associated with diminished neural response to affective



- 758 touch. *Soc Cogn Affect Neurosci* **8**:378–386. doi:10.1093/scan/nss009
- 759 Wagner DD, Kelley WM, Haxby J V., Heatherton TF. 2016. The dorsal medial prefrontal cortex responds  
760 preferentially to social interactions during natural viewing. *J Neurosci* **36**:6917–6925.  
761 doi:10.1523/JNEUROSCI.4220-15.2016
- 762 Walbrin J, Downing P, Koldewyn K. 2018. Neural responses to visually observed social interactions.  
763 *Neuropsychologia* **112**:31–39. doi:10.1016/j.neuropsychologia.2018.02.023
- 764 Walbrin J, Koldewyn K. 2019. Dyadic interaction processing in the posterior temporal cortex. *Neuroimage*  
765 **198**:296–302. doi:10.1016/j.neuroimage.2019.05.027
- 766 Walbrin J, Mihai I, Landsiedel J, Koldewyn K. 2020. Developmental changes in visual responses to social  
767 interactions. *Dev Cogn Neurosci* **42**:100774. doi:10.1016/j.dcn.2020.100774
- 768 Wen H, Shi J, Zhang Y, Lu K-H, Cao J, Liu Z. 2018. Neural Encoding and Decoding with Deep Learning for  
769 Dynamic Natural Vision. *Cereb Cortex* **28**:4136–4160. doi:10.1093/cercor/bhx268
- 770 Whitfield-Gabrieli S, Nieto-Castanon A. 2012. Conn: a functional connectivity toolbox for correlated and  
771 anticorrelated brain networks. *Brain Connect* **2**:125–141.
- 772 Wildgruber D, Ackermann H, Kreifelts B, Ethofer T. 2006. Cerebral processing of linguistic and emotional prosody:  
773 fMRI studies. *Prog Brain Res*. doi:10.1016/S0079-6123(06)56013-3
- 774 Wilson SM, Bautista A, McCarron A. 2018. Convergence of spoken and written language processing in the superior  
775 temporal sulcus. *Neuroimage* **171**:62–74. doi:10.1016/j.neuroimage.2017.12.068
- 776 Wolf I, Dziobek I, Heekeren HR. 2010. Neural correlates of social cognition in naturalistic settings: A model-free  
777 analysis approach. *Neuroimage* **49**:894–904. doi:10.1016/j.neuroimage.2009.08.060
- 778 Yang DY-J, Rosenblau G, Keifer C, Pelphrey KA. 2015. An integrative neural model of social perception, action  
779 observation, and theory of mind. *Neurosci Biobehav Rev*. doi:10.1016/j.neubiorev.2015.01.020
- 780 Zeman AA, Ritchie JB, Bracci S, Op de Beeck H. 2020. Orthogonal Representations of Object Shape and Category  
781 in Deep Convolutional Neural Networks and Human Visual Cortex. *Sci Rep* **10**:1–12. doi:10.1038/s41598-  
782 020-59175-0
- 783