Functional selectivity for naturalistic social interaction perception in the human superior temporal sulcus

Haemy Lee Masson[1] and Leyla Isik[1]

Department of Cognitive Science, Johns Hopkins University, United States

Correspondence should be addressed to Haemy Lee Masson, Department of Cognitive Science, Johns Hopkins University, Maryland, 21218, United States, E-mail: haemy.leemasson@jhu.edu

Abstract

Recognizing others' social interactions is a crucial human ability. Using simple stimuli, previous studies have shown that social interactions are processed in the superior temporal sulcus (STS) and the medial prefrontal cortex (mPFC). However, it remains unknown to what extent social interaction selectivity is observed for real world stimuli where social interactions covary with other perceptual and social information, such as faces, voices, and theory of mind. By using a naturalistic fMRI movie paradigm and advanced machine learning methods, 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 generalized across two separate fMRI datasets. These findings suggest that naturalistic social interaction perception recruits dedicated neural circuity in the STS and is a critical dimension of human social understanding.

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. For example, infants show preference for a person or character who helps others achieve their goals over someone who hinders (Hamlin, 2015; Hamlin et al., 2007; Hamlin and Wynn, 2011). These findings underscore the importance of observing others' social interactions for making social judgements.

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

however, have used simple, artificial stimuli, such as point-light figures or staged images that are largely devoid of other social information. It thus remains unknown how the humans brain processes social interactions in real world scenes, where the interactions are highly confounded with other perceptual and social properties, including 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., 2020; Nishimoto et al., 2011; Wen et al., 2018). Natural movies elicit stronger brain responses and higher inter-subject reliability than traditional experiments (Hasson et al., 2010; Sonkusare et al., 2019). In contrast, social neuroscience has not yet fully exploited these methods, although improving ecological generalizability 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), only two have investigated how the human brain processes naturalistic social interactions during movie viewing. One study identified the STS in general social perception, including social interactions (Lahnakoski et al., 2012), while a second study revealed the involvement of the mPFC in social interaction perception (Wagner et al., 2016), conjecturing that others' social interactions invite a viewer to infer others' personality and intention. These studies provided inconsistent evidence, with each study finding evidence for one brain region and not the other, and neither study spoke to the selectivity for social interactions in the brain. Critically, no prior studies of naturalistic social perception have identified or controlled for covarying perceptual features.

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 generalizability, we performed the same analyses on two different fMRI datasets (Aliko et al., 2020; Chen et al., 2017) collected by different labs showing different movies from different genres (crime vs. romance) to different groups of participants living in different countries (US vs. UK) on different MRI scanners (3T vs. 1.5T). Our findings reveal that 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 perceptual and social features.

Results

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 blood-oxygen-level-dependent (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 hue, saturation, value for each pixel (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 a deep neural network (DNN) (see Movie analysis and annotations in Materials and Methods). 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). We selected the final convolutaional layer in particular because its activations were highly correlated to those of early convolutional layers and the final fully connected layer (see Convolutional neural network in Materials and Methods). 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 inter-subject correlation (ISC, see Inter-subject brain correlation in Materials and Methods).

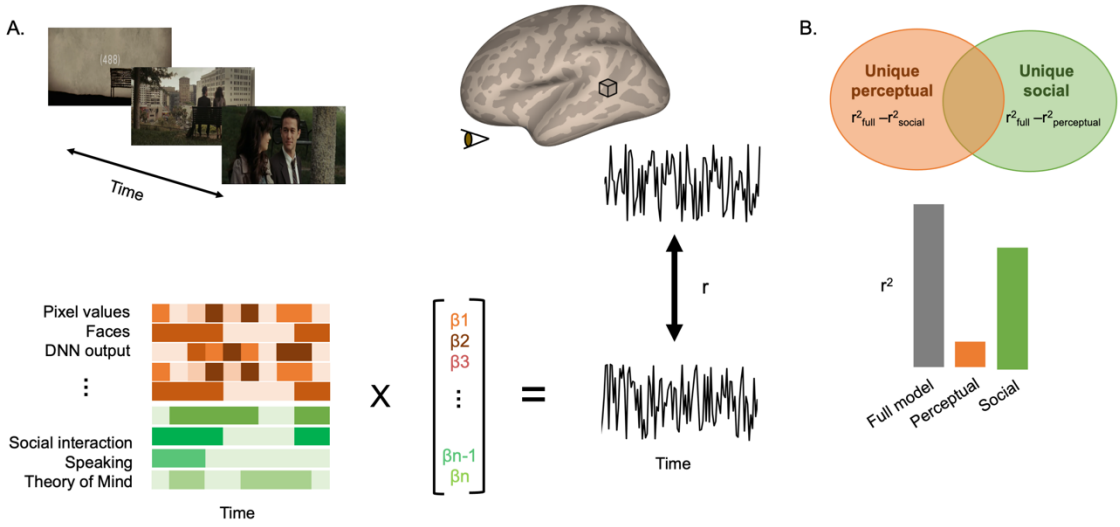


Fig 1. A. Encoding model overview. We labeled perceptual and social-affective features from a movie that participants viewed in the MRI scanner. The features include hue, saturation, value for each pixel, motion energy, the presence/absence of written words, indoor vs. outdoor scenes, the presence/absence 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 accuracy score (r) of the model is assigned to each voxel. B. Variance partitioning analysis overview. Prediction accuracy 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.

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 accuracy (r) = 0.08 ~ 0.51, mean accuracy = 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 accuracies differ across these voxels

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

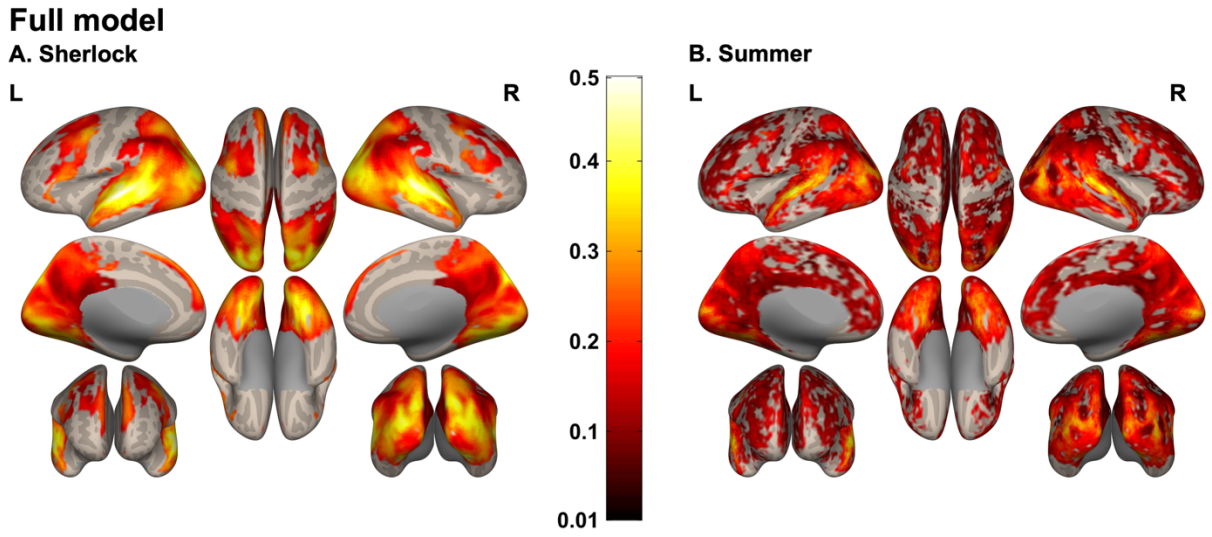


Fig 2. The full model prediction accuracy observed in Sherlock (A) and Summer (B) data. Group averaged accuracy scores are mapped on inflated cortices using the CONN software (Whitfield-Gabrieli and Nieto-Castanon, 2012) ($P_{FDR} < 0.05$, minimum cluster size > 10 voxels). The color bar indicates the prediction accuracy 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.

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 accuracy = 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 3).

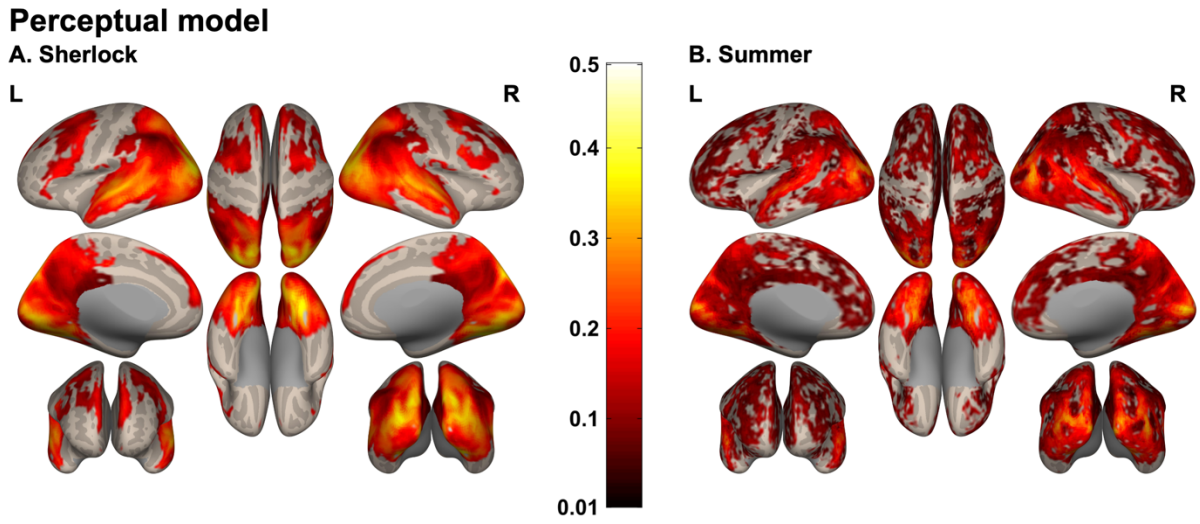


Fig 3. The perceptual model prediction accuracy observed in Sherlock (A) and Summer (B) data. The perceptual model significantly predicted neural responses in the visual cortex in both studies. Group averaged accuracy maps are thresholded at $P_{FDR} < 0.05$ and the minimum cluster size > 10 voxels.

A social-affective model, consisting of the social-affective features listed above (an agent speaking, social interactions, theory of mind, perceived 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 accuracy = 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 accuracy = 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 4).

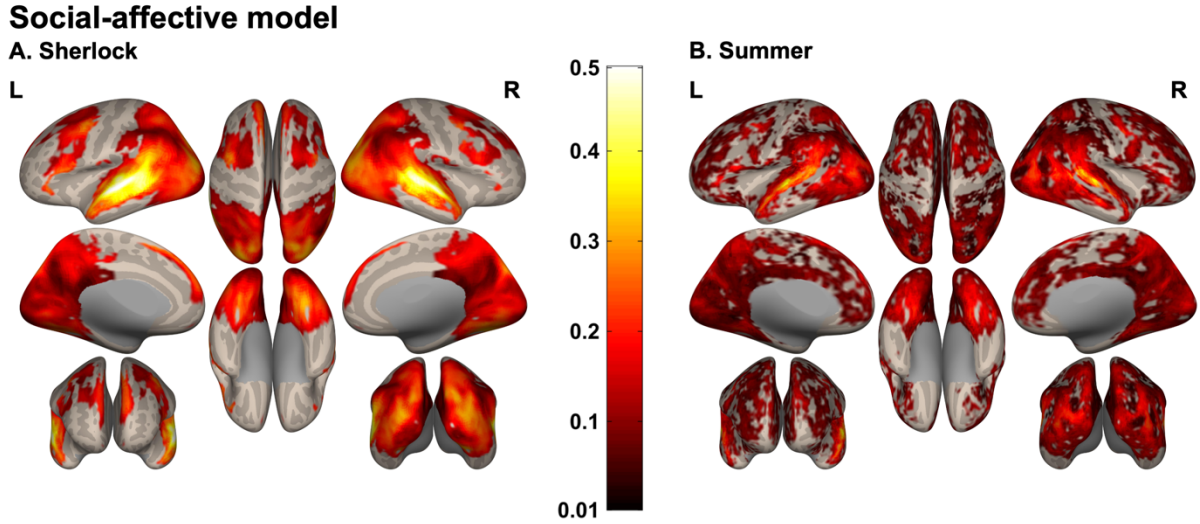


Fig 4. The social-affective model prediction accuracy observed in Sherlock (A) and Summer data (B). Group averaged accuracy scores are plotted on inflated cortices ($P_{FDR} < 0.05$, minimum cluster size > 10 voxels). Like the full model, the social-affective model significantly predicted neural responses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

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 5 illustrates the winning features that 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 5). Notably, however, this was not true in the bilateral STS where two social interaction-related features, an agent speaking (green in Fig 5) and presence of social interactions (green-yellow in Fig 5), are the most or second most preferred stimulus features in both studies (see Fig S2 for the second winning features). As expected, the auditory cortex was best explained by the audio amplitude feature (yellow-orange in Fig 5) 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 S2). 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 S2).

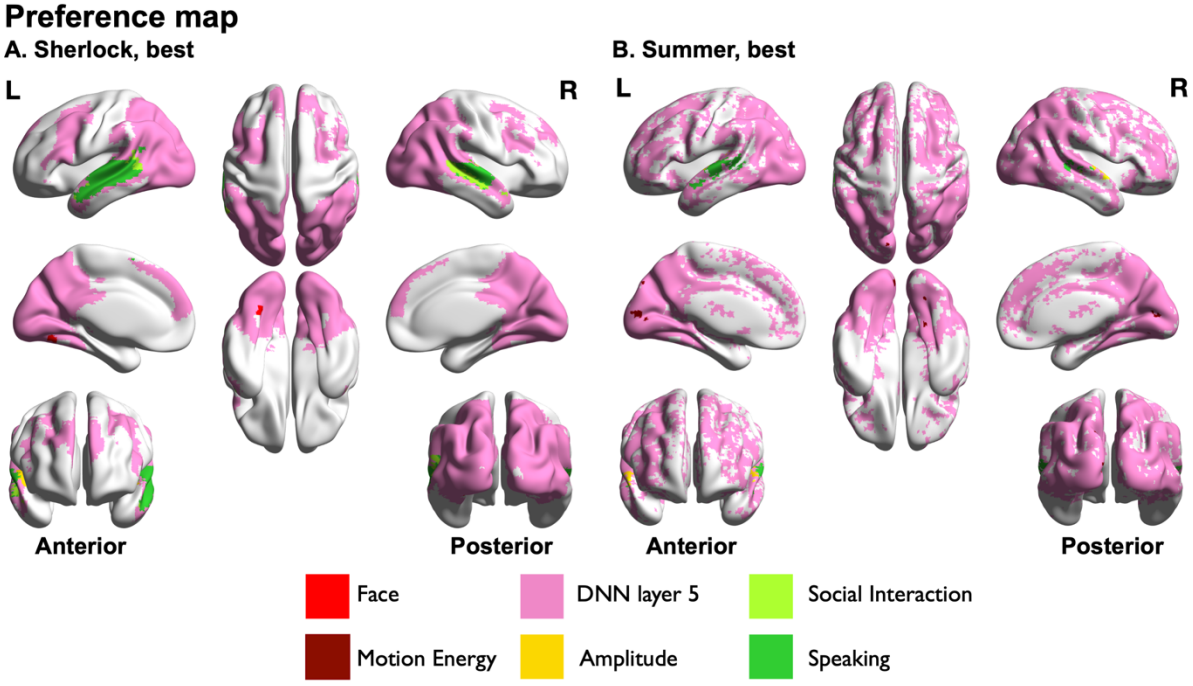


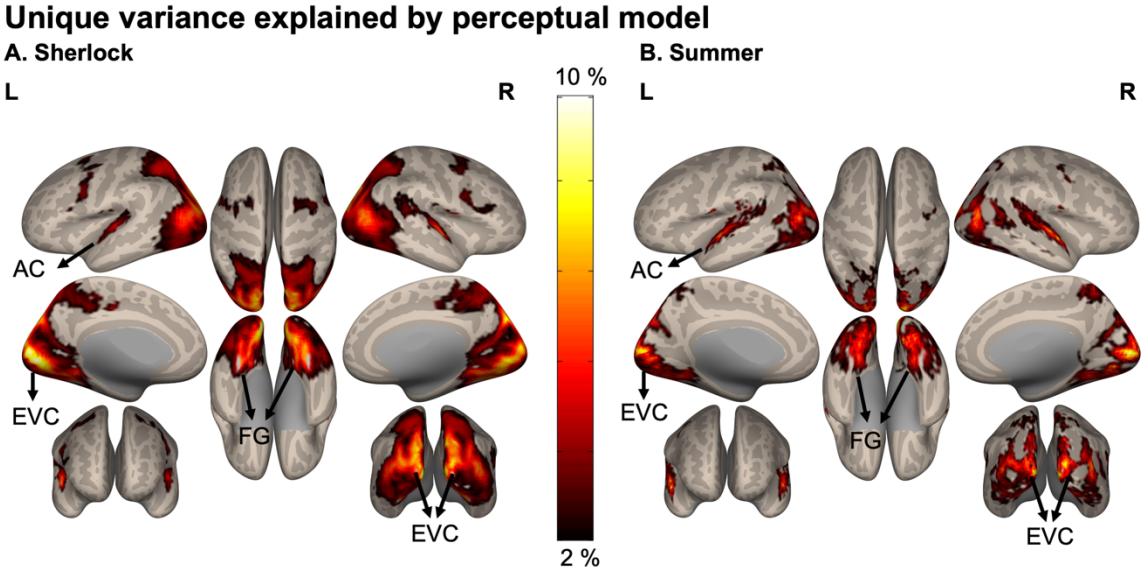
Fig 5. Preference maps for the best features in each voxel for Sherlock (A) and Summer (B) data. Feature color codes indicated at bottom. DNN features best explain the neural responses throughout most of the brain in both studies, except the STS, whose neural responses are best explained by social features.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

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 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, see Fig S1 in Supplementary Material).

The results indicated that the perceptual model significantly explained unique variance in brain regions implicated in visual or auditory processing (Fig 6A-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, temporoparietal junction (TPJ), anterior temporal lobe (ATL), and mPFC in both studies (Fig 6C-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 in the temporal lobe and occipital lobe (Fig S3). The fact that perceptual and social-affective features covary may explain this finding.



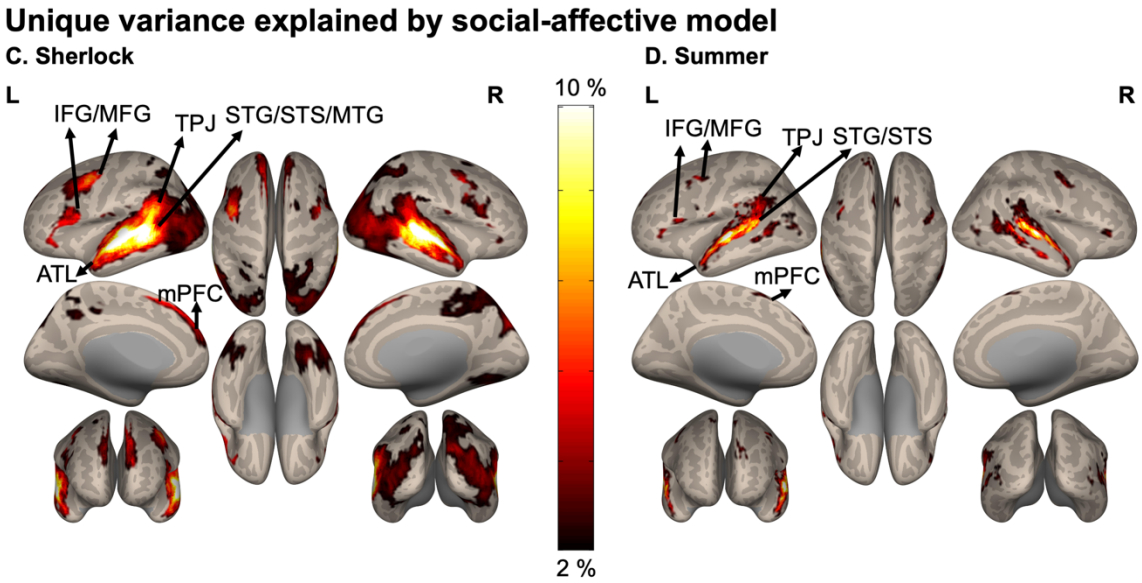


Fig 6. The portion of unique variance, expressed as percentages, explained by the perceptual (A, B) and social-affective model (C, D) in the Sherlock and Summer data. Color bar indicates percentage of unique variance explained. The perceptual model explained unique variance of neural responses in the 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.

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 7A). 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 7B). This supports and extends the results of our feature preference mapping (Fig 5), highlighting the role of social interaction processing in the STS, independent of all other features, including spoken language.

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 7C), 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, did not generalize to 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

7D). 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 functional 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.

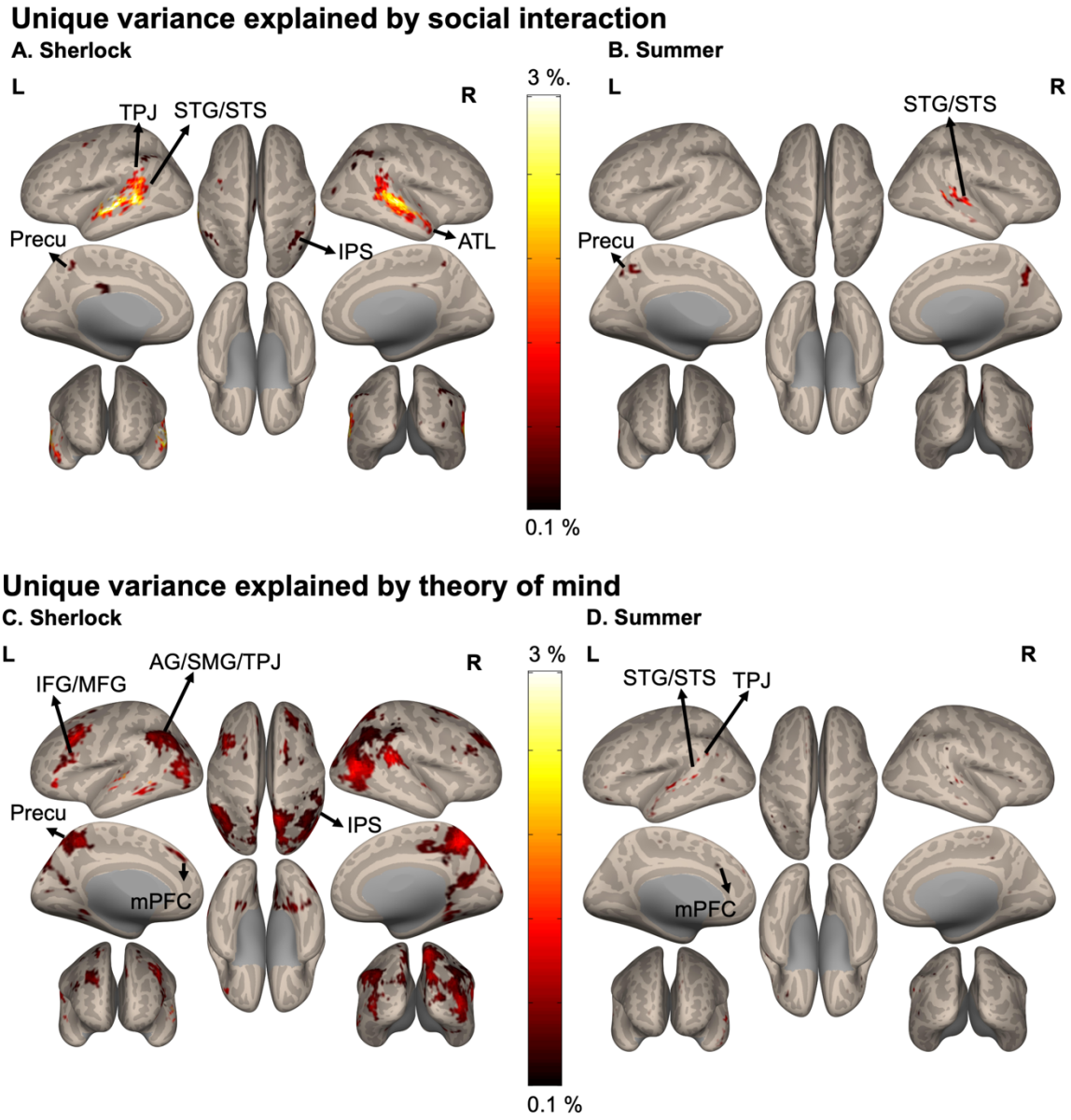


Fig 7. The amount of the unique variance explained by the social interaction (A, B) and theory of mind feature (C, D) in the Sherlock and Summer studies. Color bar indicates percentage of unique variance explained. The social interaction feature explained unique variance in neural responses in the STS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

Discussion

Here we uncovered the brain mechanisms underlying naturalistic social interaction perception in an ecologically generalizable 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 6C-D) and those particularly selective to social interactions (Fig 7A-B). We also demonstrated how auditory and visual features, including audio amplitude, faces, written words, and motion energy are encoded during natural movie viewing (Fig 6A-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 an ecologically valid context. Importantly, our findings generalized across both sets of movie data that came from different genres, subjects, MRI scanners, and labs.

Social-affective information processing during naturalistic movie viewing

Movie viewing paradigms have recently been highlighted as essential tools in cognitive neuroscience due to their richness, dynamics, 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 (e.g., the majority of movie scenes contain a social interaction or a face on the screen). 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. This method is less prone to reverse inference error as the event coding happens prior to fMRI analysis (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. In particular, we identified unique variance explained in brain regions previously attributed to social perception (STS), theory of mind (TPJ, mPFC, ATL), and action observation (inferior frontal gyrus (IFG)) (Fig 6C-D). Previous studies using controlled experiments have shown an increased response in STS to social stimuli, such as point light figures (Vangeneugden et al., 2014), social touch interaction videos (Lee Masson et al., 2018), eye gaze cues (Hooker et al., 2003), affective voice/speech (Pegado et al., 2018; Wildgruber et al., 2006), and affective touch (Lee Masson et al., 2020a; Voos et al., 2013), all of which are ubiquitous in movies. The role of the theory of mind network in social cognition has also been extensively studied in social neuroscience (see the review in (Schurz et al., 2020)). This network is consistently recruited during a variety of social tasks that involve thinking about others' thoughts, emotions, intentions, and beliefs (Saxe and Kanwisher, 2003; Wolf et al., 2010). Lastly, we identified the

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

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 5). 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). Although one prior study found that object-trained DNNs well predict voxel-wise brain activity in the temporal, parietal, and frontal lobes, including the TPJ (Wen et al., 2018). The authors conjectured that later DNN layers, including the last convolutional layer 5 (the one tested here), 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 brain activations beyond the visual cortex. Excluding the DNN contribution, the results were largely confirmatory of prior studies, including faces in the fusiform gyrus, motion energy in early visual cortex and MT, and theory of mind in the TPJ (Fig S2). Intriguingly, the only region where non-DNN features were consistently preferred was the STS. The presence of a social interaction and a character speaking were the top two features in both studies. While these results were strongly suggestive of social interaction processing in the STS, it is difficult to rule out the effect of social interactions above the highly correlated effects of spoken language processing (Deen et al., 2015; Wilson et al., 2018) in the STS with this analysis alone. Thus, we turned to variance partitioning, to identify the unique contribution of social interactions in the brain.

Social interaction perception during movie viewing

In both studies, right STS and, to a lesser extent, the precuneus showed unique selectivity for naturalistic social interactions (Fig 7A-B). With the variance partitioning analysis, we were able to rule out the effect of other features, including strongly covarying faces, speaking, and theory of mind features (Fig S1). As described above, the STS processes a wide range of social features in a modality-independent manner (Deen et al., 2015), and the posterior portion of the STS has been implicated in social interaction perception in simplified contexts (Isik et al., 2017; Lee Masson et al., 2018; Walbrin et al., 2018). The current study provides the first evidence that unique variance in STS is explained by the presence/absence of social interactions in naturalistic viewing conditions. 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 of this functional selectivity. Bilateral STS showed unique selectivity for social interaction in Sherlock data, whereas only right hemispheric lateralization was observed in Summer data. This discrepancy has also been observed in other studies. Some studies found the right hemispheric lateralization of STS during the observation of non-verbal social interaction, such as dyadic point-light displays and bodily movements (Isik et al., 2017; Lee Masson et al., 2018; Walbrin et al., 2018), while some 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 the presence of a social interaction, but to a lesser degree than STS. The precuneus has been implicated in a wide range of cognitive processes, such as visuo-spatial imagery, episodic memory retrieval, and social cognition (see review in (Cavanna and Trimble, 2006)). In the context of social cognition, the precuneus belongs to the theory of mind network and is recruited during 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 on the selective involvement of STS and precuneus is that these regions work in concert to spontaneously extract socially relevant information about others, such as traits of the interacting people. This notion is supported by previous studies showing increased precuneus's connectivity with other social brain areas, including STG and MTG, 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 a peak in the dorsal part of mPFC (dmPFC) activity during social interaction scenes of a natural movie, concluding that a viewer may spontaneously infer movie characters' thoughts and intentions (Wagner et al., 2016). Contrary to that finding, we did not observe dmPFC involvement specific to social interaction perception after controlling for the effects of perceptual and social features, including speaking and theory of mind. Instead, dmPFC was selectively tuned to more general social-affective features (Fig 6C-D) and theory of mind (Fig 7C-D). Notably, in Wagner and colleagues' study, movie scenes that evoked strong dmPFC responses were speaking scenes, 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.

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

Theory of Mind during movie 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, they did not generalize to the Summer data. This discrepancy may be related to the substantially lower signal-to-noise ratio in Summer movie data (maximum ISC value 0.73 vs. 0.56, see inter-subject brain correlation in Methods) perhaps due to the fact that the data were collected on a 1.5 T MRI scanner (Aliko et al., 2020). Another important distinction is how we annotated the theory of mind feature. Unlike previous studies (Jacoby et al., 2016; Richardson et al., 2018) which relied on reverse correlation analysis, 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.

Future directions

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? Social information has strong implications for visual neuroscience (Papeo et al., 2019; Pitcher and Ungerleider, 2021; Vestner et al., 2019). Efficient processing of social interactions seems to emerge in the visual system: two facing bodies, which occur often during and are proposed as a precursor to social interactions, are processed faster than two bodies facing away (Papeo et al., 2019; Vestner et al., 2019), evoke stronger neural responses in the body-selective visual cortex (Abassi and Papeo, 2020), and induce reciprocal brain communication between body-selective visual cortex and STS (Bellot et al., 2021). Follow-up investigations are needed to understand how these visual precursors are used to recognize real-world social interactions. Our results provide a framework to move these questions beyond controlled settings to ecologically valid experiments.

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

Materials and Methods

fMRI data sources

We re-analyzed two publicly available fMRI datasets from two different studies. In the first study (Chen et al., 2017), 17 participants watched the first episode of the Sherlock BBC TV series (duration ~ 45 mins) 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 mins). 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.

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 (140s cut off), spatial normalization to a MNI space with a re-sampling size of 3 x 3×3 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 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 x 3×3 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.

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 3sc 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. For amplitude ('audioread') and pitch ('pitch') of the audio, we averaged values over the audio samples and channels (66150 samples and two channels for Sherlock and 132300 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.

Social features were labeled by two human raters – whether or not a video segment contains social interactions (presence of social interactions = 1, absence = 0), 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 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 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 the 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 1.5s length segments into one for annotations of social features, resulting in 3s length segments for both studies.

Lastly, we added valence and arousal ratings, offered by the authors of another Sherlock fMRI study (Kim et al., 2020). In their study, 4.5s 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). The figure illustrating correlations between features can be found in supplementary material (Fig S1).

Convolutional neural network

We used a deep convolutional neural network (CNN) to extract visual features from both stimuli. Specifically, PyTorch (version 1.4.0) implementation of AlexNet (Krizhevsky et al., 2012), pretrained 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 output of the previous layer served as the input of the following layer. 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. We selected the fifth layer since activations were highly correlated to those of early (r > 0.75 with second layer) and late layers (r > 0.5 seventh layer). The output size of the fifth layer is 256 × 13 × 13, which was flattened to an array with 43264 elements. Unlike other features (1 × the number of video segments), the features from the DNN layer are high dimensional (i.e., 43264 × the number of video segments). We reduced the dimensionality without losing crucial information by applying principal component analysis (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.

In summary, perceptual features consist of components from the DNN fifth layer, 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.

Inter-subject brain 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 ($P_{FDR} < 0.001$) (Baldassano et al., 2017). For the Summer dataset, we ran a sign permutation test (5,000 iterations) across participants, as implemented in BrainIAK, and created a mask consisting of voxels passing the statistical threshold, $P_{FDR} < 0.005$ after the multiple comparison correction with the FDR. In other words, subjects' ISC values were randomly multiplied with +1 or −1 for 5,000 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 ($P_{FDR} < 0.005$ as opposed to $P_{FDR} < 0.001$) on Summer ISC to have a mask with the similar number of voxels with that of Sherlock (25468 and 22044 voxels, respectively). Voxels outside of the mask or whose BOLD responses did not change over time were excluded from the further encoding analyses.

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. 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 accuracy (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.

Procedurally the process of encoding model was carried out 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:

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

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

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

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 × 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:

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

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 accuracy.

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 accuracy 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 accuracy. 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 (21649 voxels for Sherlock; 24477 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 (5,000 iterations). From the empirical null distribution of a prediction accuracy, one-tailed P values were calculated and adjusted with FDR correction. Group averaged prediction accuracy maps of each model were thresholded at $P_{FDR} < 0.05$ and plotted on the cortical surface.

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.

4) One model includes all features except the presence of a social interaction
5) The final model includes all features except the mentalization feature.

Prediction accuracies obtained from these five models were used in variance partitioning analysis. All code for voxel-wise encoding and variance partitioning analysis associated with the current study is available at https://github.com/haemyleemasson/voxelwise_encoding.

Preference Mapping

We performed preference mapping analysis to visualize the stimulus feature that were best at explaining each voxel's activation. 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 accuracy 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 accuracy were selected for each voxel. We colored every voxel to reflect which feature predicts BOLD responses of that voxel the best. The first (Fig 5) and second preference maps (Fig S2) were plotted on the cortical surface.

Variance Partitioning

For every voxel, we conducted a variance partitioning analysis to determine the unique contribution of each encoding model in predicting the BOLD responses in withheld fMRI data. 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. We also measured the unique contribution of two important social features, the presence of a social interaction and mentalization, in predicting the BOLD responses (for shared variance, see Supplementary Material). 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 (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 accuracy 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 = 21649 for Sherlock; N=18 and Voxels = 24477 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 (5,000 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 TPJ were compared with the templates created from the previous studies (Deen et al., 2015; Mars et al., 2012).

Acknowledgements

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.

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.

References

Abassi E, Papeo L. 2020. The representation of two-body shapes in the human visual cortex. *J Neurosci* **40**:852–863. doi:10.1523/JNEUROSCI.1378-19.2019

Aliko S, Huang J, Gheorghiu F, Meliss S, Skipper JI. 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 JW, Hasson U, Norman KA. 2017. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95**:709-721.e5. doi:10.1016/j.neuron.2017.06.041

Bellot E, Abassi E, Papeo L. 2021. Moving Toward versus Away from Another: How Body Motion Direction Changes the Representation of Bodies and Actions in the Visual Cortex. *Cereb Cortex*. doi:10.1093/cercor/bhaa382

Boccadoro S, Cracco E, Hudson AR, Bardi L, Nijhof AD, Wiersema JR, Brass M, Mueller SC. 2019. Defining the neural correlates of spontaneous theory of mind (ToM): An fMRI multi-study investigation. *Neuroimage* **203**:116193. doi:10.1016/j.neuroimage.2019.116193

Bonner MF, Epstein RA. 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 JF, Nourski K V., Oya H, Reale RA, Kawasaki H, Steinschneider M, Howard MA. 2009. Coding of repetitive transients by auditory cortex on Heschl's gyrus. *J Neurophysiol* **102**:2358–2374. doi:10.1152/jn.91346.2008

Byrge L, Dubois J, Tyszka JM, Adolphs R, Kennedy DP. 2015. Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *J Neurosci* **35**:5837–5850. doi:10.1523/JNEUROSCI.5182-14.2015

Carr L, Iacoboni M, Dubeaut MC, Mazziotta JC, Lenzi GL. 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.

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, Lehéricy 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 oneâ€™s goal drives
641 infantsâ€™ 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

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 DJ. 2010. Reliability of cortical activity during natural stimulation. *Trends Cogn Sci*. doi:10.1016/j.tics.2009.10.011

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

Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM. 2018. The representational dynamics of task and object processing in humans. *Elife* **7**. doi:10.7554/eLife.32816

Hendriks MHA, Dillen C, Vettori S, Vercammen L, Daniels N, Steyaert J, Op de Beeck H, Boets B. 2021. Neural processing of facial identity and expression in adults with and without autism: A multi-method approach. *NeuroImage Clin* **29**:102520. doi:10.1016/j.nicl.2020.102520

Hensel L, Bzdok D, Müller VI, Zilles K, Eickhoff SB. 2015. Neural correlates of explicit social judgments on vocal stimuli. *Cereb Cortex* **25**:1152–1162. doi:10.1093/cercor/bht307

Hooker CI, Paller KA, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ. 2003. Brain networks for analyzing eye gaze. *Cogn Brain Res* **17**:406–418. doi:10.1016/S0926-6410(03)00143-5

Iacoboni M, Lieberman MD, Knowlton BJ, Molnar-Szakacs I, Moritz M, Throop CJ, Fiske AP. 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 MM. 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 CE, Gao C, Schulteis S, Wedell DH, Shinkareva S V. 2020. A study in affect: Predicting valence from fMRI data. *Neuropsychologia* **143**:107473. doi:10.1016/j.neuropsychologia.2020.107473

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networksAdvances in Neural Information Processing Systems. pp. 1097–1105.

Lahnakoski JM, Glerean E, Salmi J, Jääskeläinen IP, 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 RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS. 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 EM, van Hoorn J, Cohen JR, Telzer EH. 2018. Functional connectivity in the social brain across childhood and adolescence. *Soc Cogn Affect Neurosci* **13**:819–830. doi:10.1093/scan/nsy064

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 MC, Lombardo M V., Mason L, Oldenhinkel M, Persico A, Cáceres ASJ, Spooren W, Loth E, Murphy DGM, Buitelaar JK, 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 SA, 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 SA, 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 AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. 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 AO, Huth AG, Gallant JL. 2019. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage* **197**:482–492. doi:10.1016/j.neuroimage.2019.04.012

Papeo L, Goupil N, Soto-Faraco S. 2019. Visual Search for People Among People. *Psychol Sci* **30**:1483–1496. doi:10.1177/0956797619867295

Pegado F, Hendriks MHA, 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

Petrini K, Piwek L, Crabbe F, Pollick FE, Garrod S. 2014. Look at those two!: The precuneus role in unattended third-person perspective of social interactions. *Hum Brain Mapp* **35**:5190–5203. doi:10.1002/hbm.22543

Pitcher D, Ungerleider LG. 2021. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends Cogn Sci*. doi:10.1016/j.tics.2020.11.006

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 AC, 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 LJ. 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 AM, De Rosnay M, Koot HM, 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 MG, Maliske L, Margulies DS, Mars RB, 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* **undefined**:undefined. doi:10.1037/bul0000303

Skripkauskaite S, Mihai I, Koldewyn K. 2021. ATTENTION TO SOCIAL INTERACTIONS 1 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 JJA, 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 GA. 1999. Motion-responsive regions of the human brain. *Exp Brain Res* **127**:355–370. doi:10.1007/s002210050804

Tholen MG, Trautwein FM, 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

Vestner T, Tipper SP, Hartley T, Over H, Rueschemeyer SA. 2019. Bound together: Social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. *J Exp Psychol Gen* **148**:1251–1268. doi:10.1037/xge0000545

Voos AC, Pelphrey KA, Kaiser MD. 2013. Autistic traits are associated with diminished neural response to affective

touch. *Soc Cogn Affect Neurosci* **8**:378–386. doi:10.1093/scan/nss009

Wagner DD, Kelley WM, Haxby J V., Heatherton TF. 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.

Wildgruber D, Ackermann H, Kreifelts B, Ethofer T. 2006. Cerebral processing of linguistic and emotional prosody: fMRI studies. *Prog Brain Res*. doi:10.1016/S0079-6123(06)56013-3

Wilson SM, 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

Wolf I, Dziobek I, Heekeren HR. 2010. Neural correlates of social cognition in naturalistic settings: A model-free analysis approach. *Neuroimage* **49**:894–904. doi:10.1016/j.neuroimage.2009.08.060

Yang DY-J, Rosenblau G, Keifer C, Pelphrey KA. 2015. An integrative neural model of social perception, action observation, and theory of mind. *Neurosci Biobehav Rev*. doi:10.1016/j.neubiorev.2015.01.020

Zeman AA, Ritchie JB, 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