COMMON AND DISTINCT NEURAL CORRELATES OF SOCIAL INTERACTION PERCEPTION AND THEORY OF MIND

Zizhuang Miao[1]
Heejung Jung[1]
Philip A. Kragel[2]
Patrick Sadil[3]
Martin A. Lindquist[3]
Tor D. Wager[1*]

[1] Dartmouth College
[2] Emory University
[3] Johns Hopkins University

Running Head: SOCIAL INTERACTION PERCEPTION AND THEORY OF MIND

* Please address correspondence to:

Tor D. Wager
Diana L. Taylor Distinguished Professor
Presidential Cluster in Neuroscience and
Department of Psychological and Brain Sciences, HB 6207
Dartmouth College
Hanover, NH 03755
603-646-2196

SUMMARY FOR THIS DRAFT
6 figures
2 Tables
5 supp. Figures (online)
5 supp. Tables (online)

Character count:
59,595 characters with spaces (9,037 words)
Introduction + Results + Discussion + Conclusion: 36,152 characters (5,361 words)
Figures: 7,532 characters.

Abstract

Social cognition spans from perceiving agents and their interactions to making inferences based on theory of mind (ToM). Despite their frequent co-occurrence in real life, the commonality and distinction between social interaction perception and ToM at behavioral and neural levels remain unclear. Here, participants (N = 231) provided moment-by-moment ratings of four text and four audio narratives on social interactions and ToM engagement. Social interaction and ToM ratings were reliable (split-half r = .98 and .92, respectively) but only modestly correlated across time (r = .32). In a second sample (N = 90), we analyzed co-variation between normative social interaction and ToM ratings and functional magnetic resonance (fMRI) activity during narrative reading (text) and listening (audio). Social interaction perception and ToM activity maps generalized across text and audio presentation (r = .83 and .57 between unthresholded t maps, respectively). When ToM was held constant, merely perceiving social interactions activated all regions canonically associated with ToM under both modalities (FDR q < .01), including temporoparietal junction, superior temporal sulcus, medial prefrontal cortex, and precuneus. ToM activated these regions as well, indicating a shared, modality-general system for social interaction perception and ToM. Furthermore, ToM uniquely engaged lateral occipitotemporal cortex, left anterior intraparietal sulcus, and right premotor cortex. These results imply that perceiving social interactions automatically engages regions implicated in mental state inferences. In addition, ToM is distinct from social interaction perception in its recruitment of regions associated with higher-level cognitive processes, including action understanding and executive functions.

Author Note

Thanks to Bogdan Petre, Yaroslav O. Halchenko, David M. Gantz, Sydney L. Shohan, Xinming Xu, Maryam Amini, Bethany J. Hunt, and Eilis I. Murphy for data collection and management. This project was supported by grants NIBIB R01EB026549. Matlab code for analyses is available at: https://github.com/canlab, with documentation and examples at https://canlab.github.io.

Common and distinct neural correlates of social interaction perception and theory of mind

As social animals, we are surrounded by others whose appearances, expressions, and actions we can perceive, and whose thoughts, beliefs, and emotions we can infer. Based on a processing-stage theoretical framework (or dual-process model; see Lieberman, 2007; Bohl & van den Bos, 2012; Lieberman, 2022), there are two types of social cognitive processes: early perceptual processes that are automatic, inflexible, and specific to sensory modality, and late inferential processes that are controlled, flexible, and operate on multi-modal information. Although qualitatively different, the two processes often happen simultaneously and transition seamlessly. For example, people can readily infer animacy and mental states while perceiving the motions and interactions of completely inanimate geometric shapes (Heider & Simmel, 1944). However, social cognition research has been undergoing a separation, with some studies focusing on automatic social perception (e.g., Pitcher & Ungerleider, 2021) and others on controlled social inferences (e.g., Schurz et al., 2021). One prominent example is the distinction between social interaction perception, which involves perceiving the interactions of other agents from a third-person perspective (e.g., Isik et al., 2017), and theory of mind (ToM), which involves deliberating over other agents' mental states (e.g., Frith & Frith, 2005). The separation in research topics has promoted the idea that social interaction perception and ToM are distinct mental processes implemented in separate brain regions (McMahon & Isik, 2023). Given their frequent co-occurrence (Heider & Simmel, 1944) and potential confounding (e.g., Varrier & Finn (2022) and Moessnang et al. (2016) studied social interaction perception and ToM, respectively, using similar stimuli), results from different studies may not reflect their actual differences. It thus remains unclear how common or unique social interaction perception and ToM are at both neural and behavioral levels.

We will divide this question into two parts and seek answers by modeling social interaction perception and ToM in the same neuroimaging study. First, we will investigate to what extent social interaction perception is a purely perceptual process or also involves conceptual processes. Social interaction perception has been predominantly studied using visual or audiovisual stimuli, such as animations of geometric shapes (e.g., Varrier & Finn, 2022), animations of virtual characters (e.g., Schilbach et al., 2006), videotapes (e.g., Ainsworth et al., 2021), and movie clips (e.g., Lahnakoski et al., 2012). It activates many heavily visual regions, such as middle temporal cortex or lateral occipitotemporal cortex (e.g., Landsiedel et al., 2022; Wurm & Caramazza, 2019) and cuneus (e.g., Schilbach et al., 2006; Wolf et al., 2018). Accumulating behavioral and neural

evidence suggests that social interaction perception is a visually grounded perceptual process (McMahon & Isik, 2023). However, other neuroimaging studies have opposite findings: brain regions that respond to social interactions overlap with those responding to ToM (Arioli & Canessa, 2019), include multimodal regions in anterior temporal cortex (McMahon et al., 2023), and show generalizability across sensory modality (Landsiedel & Koldewyn, 2023) and task modality (Lee Masson et al., 2024b). These findings implicate an alternative view that social interaction perception has cross-modality components and could involve conceptual processing. To further compare these two views, we investigate whether any brain regions that respond to social interactions in visual stimuli in previous studies also respond to visually and auditorily presented verbal stimuli in the current study. We also compare the neural correlates of social interaction perception with those of ToM which, by definition, involves conceptual inferences.

Second, we will ask to what extent ToM is a controlled or automatic process. Much developmental and evolutionary evidence (e.g., Frith & Frith, 2005; Richardson et al., 2018; Deen et al., 2023) has shown that ToM is a more resource-demanding mental function compared to more primitive functions such as face perception. Neuroimaging evidence supports this by showing that ToM is primarily implemented in the multimodal neocortex, such as temporoparietal junction, medial prefrontal cortex, and precuneus (e.g., Schurz et al., 2021; Dufour et al., 2013). Based on this view, existing studies have operationalized ToM as a controlled mental process, such as deliberately reasoning about the beliefs of characters in stories (Saxe & Kanwisher, 2003) and the intentions of people in pictures (Spunt & Adolphs, 2014). However, when researchers do not ask participants to deliberately use ToM while watching a movie, the ToM annotations correlate with neural activity in largely the same brain regions found in controlled experiments (Jacoby et al., 2016). These regions are also engaged by many other pre-reflective subjective construal processes (Lieberman, 2022). The findings suggest that ToM may be used in an effortless, automatic way during naturalistic experience. To better understand how controlled or automatic ToM is, we study it in situations akin to real life using narratives and compare it with another more automatic process, social interaction perception.

Addressing similar questions, several recent studies have made breakthroughs by directly comparing the neural correlates of social interaction perception and ToM and found that they are separable (Isik et al., 2017; Lee Masson et al., 2021; Lee Masson et al., 2024a). This suggests that social interaction perception and ToM are at different processing stages, as separation in neural implementations appears to be a sufficient condition for separation at the mental level. However, a methodological limitation is worth noting: previous studies have operationalized ToM in controlled experiments as one condition versus another (Isik et al., 2017), and in movie stimuli

as the annotations of where the viewer is led to think about the characters' thoughts (Jacoby et al., 2016) or where the characters are inferring other characters' mental states (Lee Masson & Isik, 2021; Lee Masson et al., 2024a). These operational definitions are more like a stimulus property, or "ToM demands," rather than the underlying cognitive process, or ToM *per se*. This discrepancy between measurements and the mental process could impact the validity and sensitivity of linking ToM to neural activity (Zaki & Ochsner, 2009). Indeed, in the same dataset, voxel-wise modeling reveals separable neural correlates between ToM and social interaction perception (Lee Masson & Isik, 2021), but regions-of-interest-based analysis finds some overlaps (Lee Masson et al., 2024a), partly because the latter method has larger power. To overcome this limitation, we use self-report measurement for ToM during narratives and compare it with the annotation measurement used in previous studies.

Taken together, we aim to investigate to what extent social interaction perception is a modality-specific perceptual process and ToM is a controlled process. For this, we took a "cross-modal replication strategy" by presenting written narratives to participants as both texts and audios, and compared neural correlates across the two modalities. We also sought to identify specific regions for social interaction perception and ToM regardless of sensory inputs. We correlated social interactions and ToM with neural activity in a single model and compared their neural correlates when controlling for each other. Spatially separable neural correlates of social interaction perception and ToM can support the hypothesis that social interaction perception and ToM are two distinct mental processes, whereas shared neural correlates between the two will challenge it.

Results

We collected and analyzed two datasets in the current study. The first dataset came from an online behavioral experiment where participants (N = 231) provided ratings about how much they thought there were social interactions (N = 114) or how much they were using ToM (N = 117) in real time as they read four written narratives and listened to another four. We extracted the medians of those ratings to serve as continuous estimates of the subjective perception of social interactions and engagement of ToM for general populations. The second dataset came from a neuroimaging experiment where another group of participants (N = 90) passively experienced the same narratives under functional magnetic resonance imaging (fMRI) scanning without providing any ratings. We then correlated the medians of online behavioral ratings with fMRI data using voxel-wise random-effect general linear models (GLM; Figure 1a). All inferential statistical tests were two-tailed. Significance in multiple comparisons in voxel-wise modeling of neural data was

corrected by false discovery rate (FDR; Benjamini & Hochberg, 1995) $q < .01$. Significance in other multiple comparisons was corrected using Bonferroni correction.

In the following sections, we first validate the methods by checking whether written narratives and self-report measurement could reliably reveal neural correlates of social interaction perception and ToM. Then we directly compare the neural correlates of social interaction perception and ToM and use Bayes Factors to classify each voxel as responding to both social interactions and ToM or only to one of them. We found that the neural correlates of social interaction perception were highly similar across modalities. Furthermore, the voxels that responded to social interactions across modalities were also associated with ToM, while ToM was uniquely associated with other brain regions.

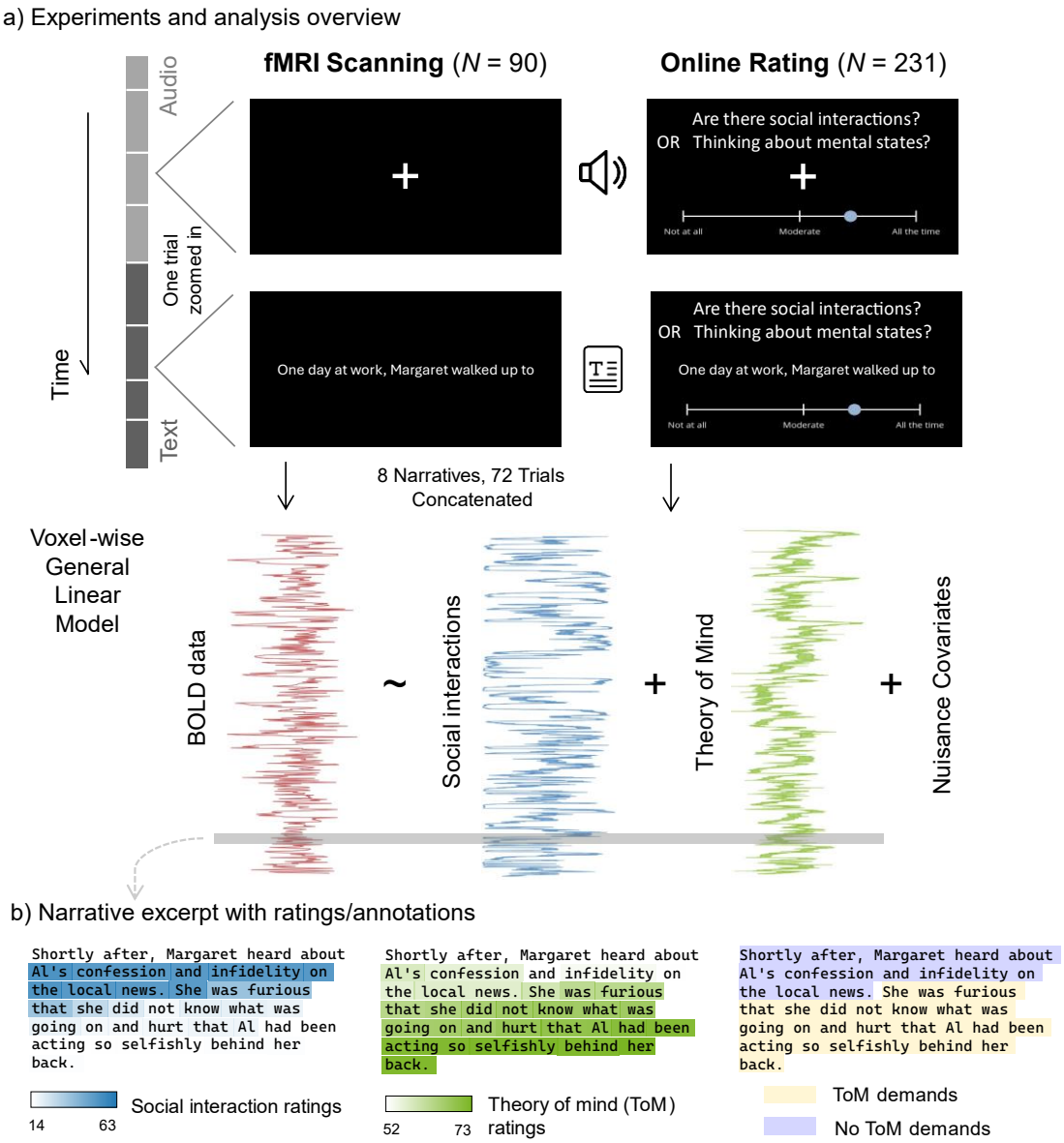


Figure 1. Study overview. a) The stimuli used in this study were eight narratives, with the first four presented as audios and the last four presented as texts. One group of participants (*N* = 90) experienced the narratives during functional magnetic resonance imaging (fMRI) scan. Another group of participants (*N* = 231) experienced the narratives online on their personal computers and simultaneously provided moment-by-moment ratings about social interactions in the narratives (*N* = 114) or their use of theory of mind (ToM) (*N* = 117). The black rectangles illustrate the screen that participants saw when narratives were presented. One online participant only saw one of the two questions shown in this figure throughout the study. We concatenated blood-oxygen-level-dependent (BOLD) signals in fMRI scanning (one fake BOLD time series was plotted for illustration only) and the medians of online ratings across all trials (actual data were plotted). Then we located the neural correlates of social interaction perception and ToM by two-stage voxel-wise general linear models. Shaded areas around the lines of online ratings indicate confidence intervals calculated by median absolute difference. b) An example narrative excerpt extracted from the eighth trial of Narrative #1 (see Supplementary Materials for all narrative texts). The medians of social interaction and ToM ratings during this excerpt, as well as researchers' annotations of ToM demands, were displayed as the highlight color of each word. The approximate time window of this excerpt was shown by the horizontal gray bar in Panel a.

Cross-modal generalizability of social interaction perception

Because few existing studies have presented social interactions in written narratives, we tested whether the narratives were valid and sensitive to probe the neural correlates of perceiving social interactions.

Subheader 1. Consistent and reliable ratings of social interactions

In the online study, participants' ratings were similar to each other (Figure 2a left). This was corroborated by the moderate to high pairwise correlations between participants' ratings (Median = .47). Participants most frequently used the two ends of the scale to answer the social interaction question (Figure 2a right), suggesting that they tended to have high confidence about the presence or absence of social interactions. More importantly, participants provided highly reliable ratings, as suggested by large split-half correlations ([.96, .99], *M* = .98; estimated from 2000 permutation samples and corrected by the Spearman-Brown formula (Brown, 1910; Spearman, 1910)). Overall, these results suggest that participants perceive social interactions in the written narratives in a consistent and reliable way.

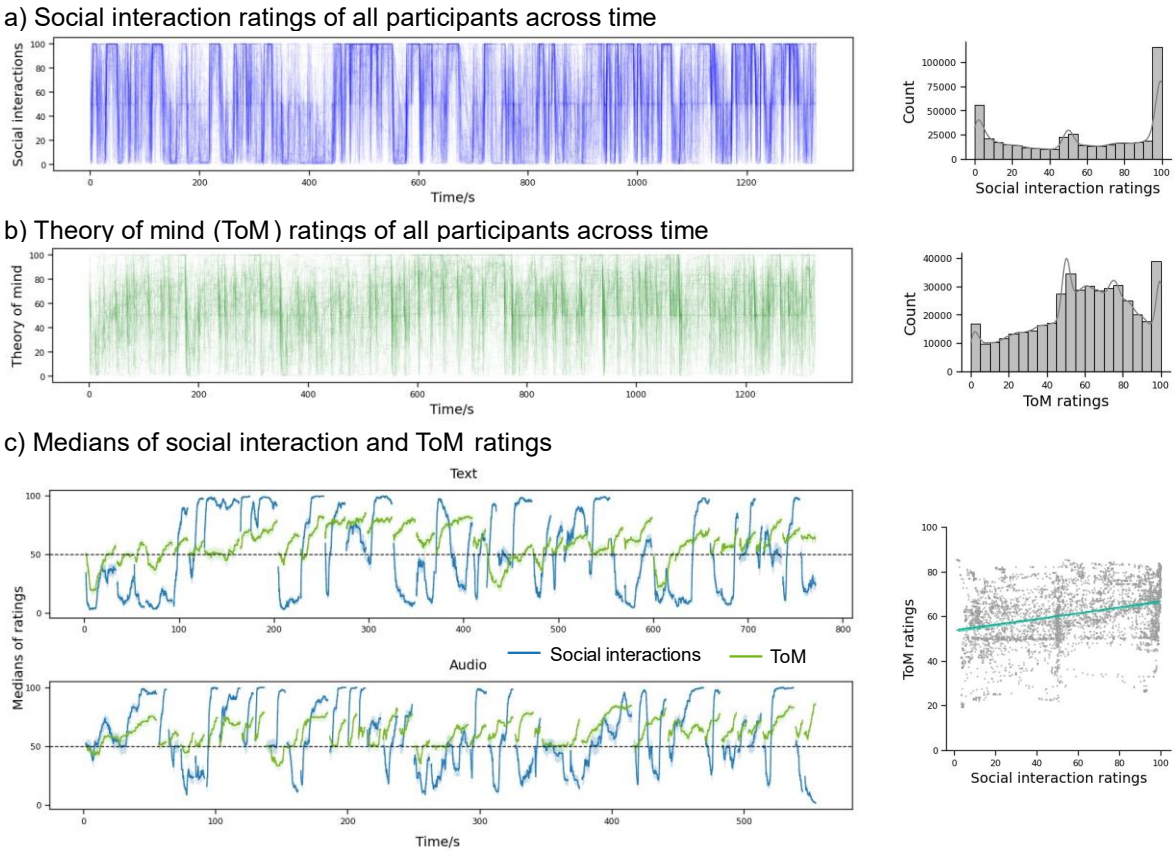


Figure 2. Ratings of online participants about social interactions and theory of mind (ToM). a) and b) In the line plots on the left, the moment-by-moment ratings of one participant were plotted as a thin quasi-transparent line. Thicker/darker parts indicate that more participants shared ratings at those time points. In the histograms on the right, all ratings were binned and plotted. c) In the line plots on the left, medians of ratings were shown by solid lines, and confidence intervals of the medians were shown by shaded areas around the line. We divided all ratings into two modalities and showed them in two plots. Note: the lines have many discontinuity points because a few data points at the start of each trial were removed from further analysis due to inadequate valid ratings (see Methods for details) and we show the version after removal in this figure. In the scatter plot on the right, each dot corresponds to a time point in the line plots. The cyan line is the best fit line estimated by ordinary least squares regression (r = .323).

Subheader 2. Similar neural correlates of social interaction perception across modalities and with previous findings in narratives

We used medians of online participants' ratings in the Text and Audio modality separately to regress each fMRI participant's neural activity. Then we performed group-level t-tests (N = 90)

over the individual regression coefficients for the two modalities separately. We found that the unthresholded whole-brain effect maps were highly similar across modalities (r = .83; for thresholded maps see Figure 3a left). In a side analysis, we also found that the whole-brain predictive models for social interactions were generalizable across modalities (see Supplementary Results). This suggests that participants use the same neural circuitries to process social interaction information both when reading and when listening to the narratives.

We then located the cross-modal neural correlates of social interaction perception by taking the conjunction of the two effect maps (Figure 3a left). In the conjunction map, a voxel was significant if and only if it was significant in both original maps with the same sign. Many regions implicated in social cognition studies were significant in the conjunction map, including bilateral temporoparietal junction (TPJ), superior temporal sulcus (STS), precuneus (PC), dorsomedial prefrontal cortex (dmPFC), and inferior frontal gyrus (IFG). Moreover, early visual cortex including V3 and V4, lateral occipital cortex (LOC), and parts of ventral visual stream or inferior temporal cortex (ITC) (Glasser et al., 2016) were also significant (Figure 3a right, also see Table S1). This suggests that social interactions in narratives are correlated with neural activity in regions typically seen as responsible for visual processing, even when there is no visual information about social interactions.

The cross-modal regions responding to social interactions in the narratives were similar to those reported in previous studies (e.g., Georgescu et al., 2014; Walbrin et al., 2018; Wagner et al., 2016). To further compare the results with previous findings in a quantitative way, we used the Neurosynth database (Yarkoni et al., 2011) to identify brain masks where all voxels were significantly related to “social interaction”, “social cognition”, or “psts” (Isik et al., 2017; Pitcher & Ungerleider, 2021). Then we extracted the means of regression coefficients (betas) of social interactions within the masks, a metric reflecting the association between within-mask neural activity and behavioral ratings of social interactions. The brain-behavior associations under both modalities within all three masks were significantly above 0 (all $t(89)$s > 6.5, all ps < .001; Figure 3b). This indicates that brain areas associated with social interactions in previous studies are also activated by the social interaction ratings in the current study. We further tested the robustness of the results by annotating social interactions in the narratives following the methods of previous studies (e.g., Nijhof & Willems, 2015; Lee Masson & Isik, 2021) and found similar result patterns (see Supplementary Results). In summary, the results at both behavioral and neural levels support that narratives are valid and sensitive stimuli to probe the neural correlates of social interaction perception.

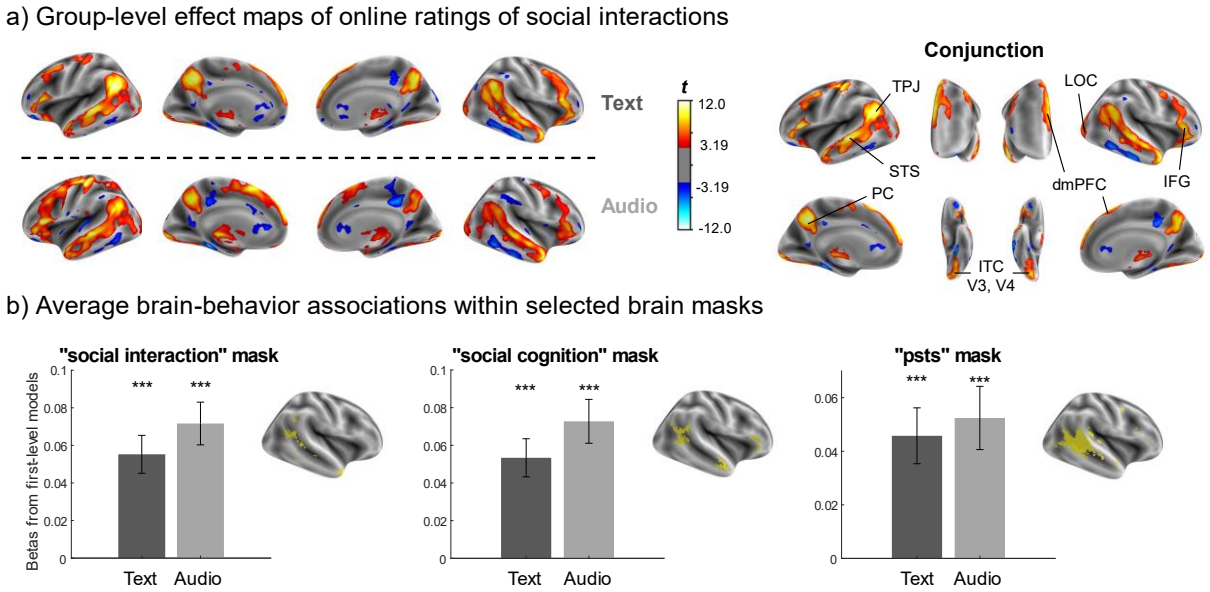


Figure 3. Neural correlates of social interaction perception in narratives. a) Group-level effect maps of social interactions as the medians of online participants' ratings, thresholded at FDR $q < .01$. Maps on the left showed separate patterns for Text and Audio modality, and maps on the right showed the conjunction across modalities. We did analyses in the volumetric space and projected results onto the Freesurfer average surface maps (Fischl, 2012) by linear interpolation for visualization only. The colormap of *t* values applied to all maps. We annotated several key regions that were significant in both hemispheres and both modalities. Abbreviations: TPJ – temporoparietal junction; STS – superior temporal sulcus; PC – precuneus; dmPFC – dorsomedial prefrontal cortex; IFG – inferior frontal gyrus; LOC – lateral occipital cortex; ITC – inferior temporal cortex. b) Group-level effects of social interactions in two modalities within three brain masks. Parts of each mask were illustrated by the yellowish-green fill color in the surface plots of the right hemisphere; the whole masks were used for analysis. Bar heights indicate the means of regression coefficients (betas) from the first-level models within each mask across participants, which reflect the strength of associations between social interaction ratings and neural activity within each mask. Error bars indicate 95% confidence intervals. *** indicates $p < .001$ in one-sample *t*-tests against 0 (after Bonferroni correction).

Validation of self-report measurement of ToM

In the Introduction, we argued that researchers' annotations of ToM events should be thought of as "ToM demands" and may not be sufficiently sensitive to localize ToM-related neural activity. We sought to overcome this limitation by asking participants to report how much they

were thinking about characters' mental states, and refer to this measurement as "ToM engagement". We also annotated the narratives in terms of ToM demands (Figure 1b) which approximated the annotations in published studies (e.g., Lee Masson & Isik, 2021). Below we quantitatively compare the two measurements, ToM engagement and ToM demands, on the behavioral and neural levels.

Subheader 1. Low similarity between ToM engagement and ToM demands

For ToM engagement ratings, participants used the middle to right end portion of the scale more than the social interaction ratings, suggested by denser distributions in the range 50-80 (Figure 2b) and smaller variance across time (Mean standard deviation = 19.35 for ToM engagement; 32.97 for social interactions). Participants were less similar to one another on their ToM engagement ratings, as seen in lower pairwise correlations between participants (Median = .21) than social interactions. Still, the ToM engagement ratings were highly reliable (split-half correlations [.86, .95] in 2000 permutation samples, M = .92, corrected by the Spearman-Brown formula), supporting the validity of applying those ratings to another participant group.

To obtain ratings for ToM demands, four researchers in our group independently annotated the narratives in the unit of sentence parts, the smallest chunk in a sentence that contains one or one sequence of predicates and their associated objects (Nijhof & Willems, 2015). We assigned every word in the narrative into one sentence part. The raters determined whether each sentence part had explicitly mentioned mental states of characters (1 = yes, 0 = no), as the labels for "ToM demands". In the current report, "mental states" included both the cognitive aspects (e.g., beliefs and goals), and the affective aspects (e.g., emotions and desires), and were typically signaled by specific verbs and adjectives (see Figure 1b for an example). Overall, the raters gave moderately similar annotations about ToM demands (pairwise correlations between raters ranged from .51 to .60 (M = .57) and the Fleiss' kappa was .54). Then we grouped all the annotations and assigned a consensus label to each sentence part (conflicts were addressed by discussion).

We compared the ToM demand labels to the medians of ToM engagement ratings under the Text and Audio modalities separately. Correlations between the two measures were not significant (r = .14, p = .22 for Text, and r = .19, p = .09 for Audio; p values estimated from 10000 bootstrapped phase randomization samples). This indicates that ToM demands annotations capture different behavioral profiles than participants' usage of ToM.

Subheader 2. Higher sensitivity of ToM engagement to localize ToM-related neural activity

Given that ToM engagement and ToM demands may be qualitatively different on the behavioral level, we expected to find different neural correlates associated with them. The question is which of the two is better at locating neural activity related to ToM. We answered this by comparing the neural correlates of the two measures both qualitatively (checking which regions were significant) and quantitatively (within brain masks associated with ToM in previous studies). Specifically, we fit two GLMs with the only difference being whether the ToM regressors were related to ToM engagement or ToM demands, and we compared neural results across sensory modalities. For the qualitative comparison, we found that ToM engagement was significantly correlated with bilateral TPJ, STS, PC, dmPFC, and IFG in both modalities, the typical regions associated with ToM (Schurz et al., 2021). By contrast, ToM demands only showed cross-modality activations in a subset of those regions, including bilateral PC and STS, left TPJ, and left IFG, and the areas of significant regions were smaller under the same threshold (Figure 4a).

For the quantitative comparison, we selected brain masks from the “tom” and “mentalizing” term maps from the Neurosynth database (Yarkoni et al., 2011) and the ToM group maps from a false belief task (Dufour et al., 2013). The first comparison was between the brain-behavior associations within masks, operationalized as the average standardized regression coefficients (betas). Standardized betas were calculated as each voxel’s beta divided by the average of the absolute betas in the whole brain, which were at roughly the same scale and comparable across models. The average standardized betas of ToM engagement and ToM demands did not differ in the whole brain ($t(89) = 1.51$, $p = .27$ for Text; $t(89) = 2.04$, $p = .09$ for Audio). By contrast, in almost all selected masks and both modalities, ToM engagement had larger standardized betas than ToM demands (all $t(89)$s > 3, all ps $< .05$), with the only exception in the right STS mask under the Text modality ($t(89) = 1.34$, $p > .05$; results from the Neurosynth masks were shown in Figure 4b, other results were in Figure S4). This indicates that ToM engagement is associated with neural activity in canonical ToM regions more strongly than ToM demands.

The second comparison was on the similarity between the spatial patterns of betas and meta-analytic maps within masks. In each mask created from the Neurosynth database, we extracted each participant’s betas and correlated them with the z values in the meta-analytic map. We then transformed the Pearson’s correlation coefficients to Fisher’s z values to ensure normality, which were then subject to paired-sample t-tests. In both “tom” and “mentalizing” masks and both modalities, betas from ToM engagement had a larger spatial pattern correlation with meta-analytic maps (all $t(89)$s > 5.6, all ps $< .001$; Figure 4c). This analysis could not be performed on the ToM group maps (Dufour et al., 2013) because t values were not available for all voxels. This indicates that the neural activity associated with ToM engagement is more similar to existing

results about ToM. Together, there is strong evidence that ToM engagement has larger statistical power at localizing ToM-related brain regions than ToM demands.

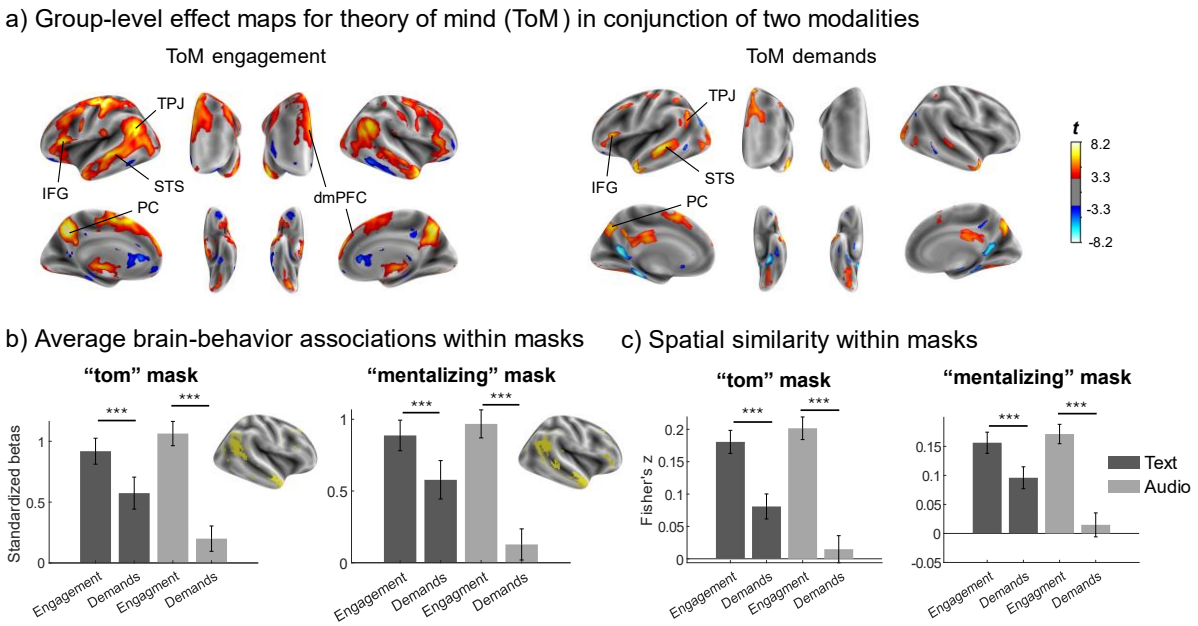


Figure 4. Neural correlates of theory of mind (ToM) engagement and ToM demands. a) Group-level effect maps of ToM as the medians of online participants' ratings of ToM engagement and the offline annotations of ToM demands, thresholded at FDR $q < .01$. We only displayed the conjunction maps across modalities for simplicity. The method for projection from volumetric space to surface space and abbreviations were the same as in Figure 3a. b) Group-level comparisons of the first-level beta estimates for ToM engagement ("Engagement") and ToM demands ("Demands") in two modalities within two brain masks. Bar heights indicate means of standardized betas and error bars indicate 95% confidence intervals. Parts of the masks on the right hemisphere were indicated by the yellowish-green fill color in the surface plots. c) Group-level spatial pattern correlations between beta maps and Neurosynth maps within each mask. Bar heights indicate the means of Fisher *z* values transformed from Pearson's correlation coefficients and error bars indicate 95% confidence intervals. In all graphs, *** indicates $p < .001$ (after Bonferroni correction).

Comparison between social interaction perception and theory of mind

So far, we have shown that the group-level summaries (medians) of online participants' ratings can serve as a reliable and sensitive measure to find the neural correlates of social interaction perception and ToM in the narratives. The medians of the two ratings were correlated to a low to moderate degree ($r = .29$ for Text modality, $r = .37$ for Audio modality, $r = .32$ for all

time points; Figure 2c). This supports the view that people tend to use ToM at the same time as perceiving social interactions during natural experience and there may not be a complete distinction between the two processes on the behavioral level. Based upon that, we sought to investigate the common and distinct neural correlates of social interaction perception and ToM.

Subheader 1. Contributions of both social interactions and ToM to neural activity in overlapping regions

We fitted a GLM with both the medians of self-reported social interactions and the medians of self-reported ToM, with separate regressors for Text and Audio modality. The correlations between social interaction and ToM ratings stayed relatively constant after being convolved with the canonical hemodynamic response function (r = .29 for Text modality; r = .38 for Audio modality). Those correlations did not raise concerns for multicollinearity problems (variance inflation factors < 1.3 for all regressors in the model).

Between the Text and Audio modalities, the effect maps of social interaction perception when controlling for ToM were not as similar as when being modeled alone (r = .50 for unthresholded t maps, Figure 5a left); similar for the effect maps of ToM (r = .51, Figure 5b left). Still, we took the conjunction across modalities (Figure 5a and 5b right) because the common neural correlates across modalities were more invariant to low-level stimulus features and thus more specific to the mental processes to be compared. The brain regions responding to social interactions under both modalities include bilateral PC, STS, IFG, dmPFC, and left TPJ; those responding to ToM include bilateral TPJ, PC, IFG, dmPFC, and left STS, as well as bilateral premotor cortex (PMC), intraparietal sulcus (IPS), and bilateral lateral occipitotemporal cortex (LOTC) which encompasses the middle temporal and medial superior temporal area (MT+). Besides whole-brain analysis, we tested the effects of social interactions and ToM in some selected cortical regions (Glasser et al., 2016; Table 2), namely left TPJ, bilateral dmPFC, right PMC, bilateral MT+ (Figure 5c), bilateral PC, left STS, right TPJ, and bilateral supplementary motor area (SMA) (Figure S5). In some selected regions like left TPJ, both social interactions and ToM had significant effects under both modalities, while in some other regions like right PMC, only ToM had significant effects. These results imply that social interactions and ToM have largely overlapping neural correlates, and both contribute independently to driving the neural activity in those regions. At the same time, ToM appears to recruit brain regions that social interactions do not.

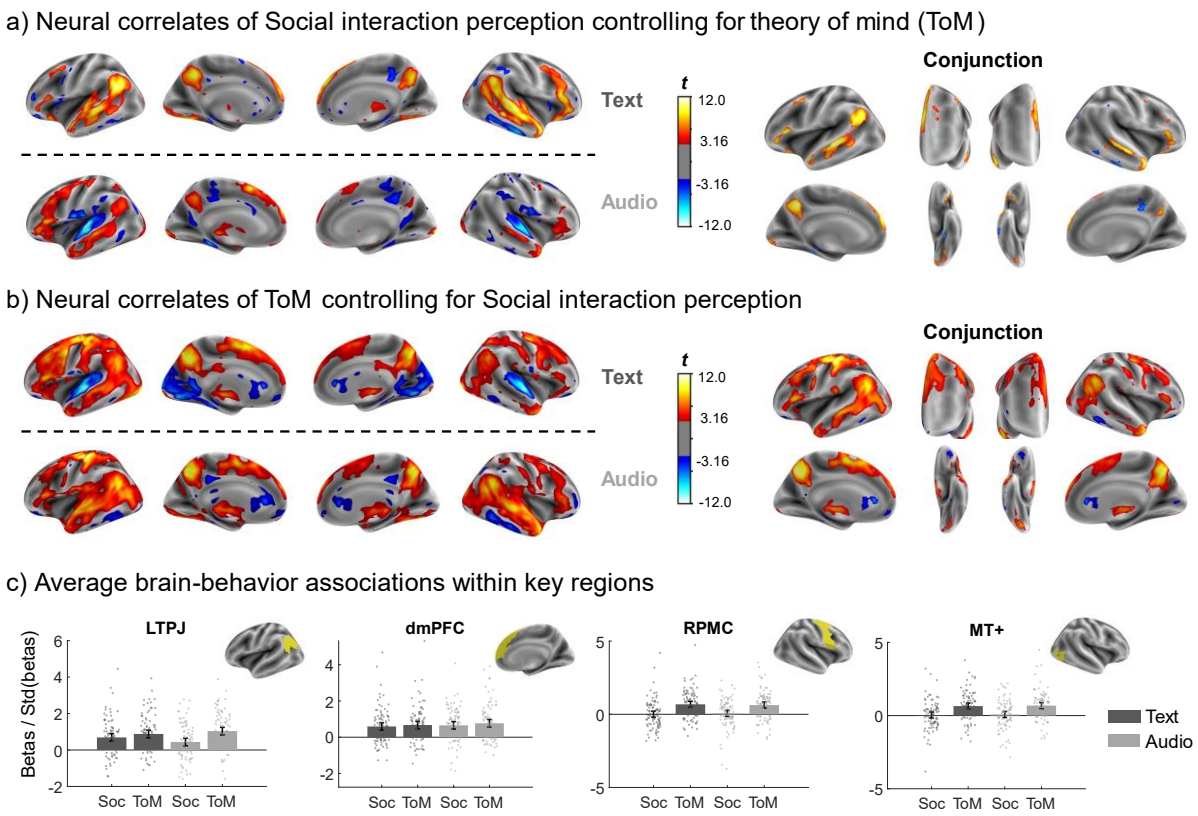


Figure 5. Neural activity associated with social interaction perception and theory of mind (ToM). a) and b) Group-level effect maps of medians of social interaction and ToM ratings after controlling for each other, thresholded at FDR $q < .01$. The method for projection from volumetric space to surface space was the same as Figures 3a and 4a. Surfaces on the left show neural correlates under Text and Audio modality separately, while surfaces on the right show the conjunction across modalities. The color bar of *t* values applies to all surfaces. c) Bar plots of the brain-behavior associations in four regions of interest. Bar heights indicate the means of first-level betas divided by their standard deviation, error bars indicate 95% confidence intervals, and shaded scatters indicate estimates for each participant. Abbreviations: std – standard deviation, LTPJ – left temporoparietal junction, dmPFC – dorsomedial prefrontal cortex, RPMC – right premotor cortex, MT+ – middle temporal and medial superior temporal area, Soc – social interactions.

Subheader 2. Cross-modal neural correlates of social interaction perception as a subset of those of ToM

We quantitatively investigated which voxels and regions showed significant correlations with both social interactions and ToM, or only one of them but not the other. To do so, we calculated Bayes Factors for each voxel and region of interest to quantify the evidence for the

alternative and null hypothesis, or presence and absence of an effect. For one voxel or region, if we had significant evidence for the alternative hypothesis of the effects of both ToM and social interactions, we would classify it as “Both”; if we had significant evidence for the alternative hypothesis of ToM and the null hypothesis of social interactions, we would classify it as “ToM only”; in the opposite case, we would classify it as “Social interactions only”. We chose 5 as a threshold for the alternative hypothesis and 1/5 for the null hypothesis to obtain moderate to strong evidence given the current sample size (Rouder et al., 2009).

As seen in Figure 6a, across two modalities, both social interactions and ToM correlated with neural activity in bilateral PC, dmPFC, STS, IFG, and left TPJ. In Text modality, social interactions showed unique correlations with activity in right STS, left early visual cortex, and bilateral ITC; in Audio modality, right early visual cortex and left PMC. However, social interactions showed rarely any unique correlations that were common across modalities. By contrast, ToM showed cross-modal unique correlations with neural activity in left anterior IPS, right PMC, bilateral MT+, and left primary somatosensory cortex. We calculated the mean betas and Bayes Factors in the same eight selected regions as for the beta maps and found further evidence that there were regions responding to “Both” and regions responding to “ToM only” consistently across modality, but not for “Social interactions only” (Figure 6b and Figure S5). Across two modalities, activity in about 6.46% of all voxels was correlated with both social interactions and ToM, about 2.84% with ToM only, and a negligible proportion (< .01%) with social interactions only (Figure 6c). The results provide strong evidence that the common neural correlates of social interaction perception across modalities are a subset of those of ToM during naturalistic experience. Besides, several brain regions are only associated with ToM but not social interactions.

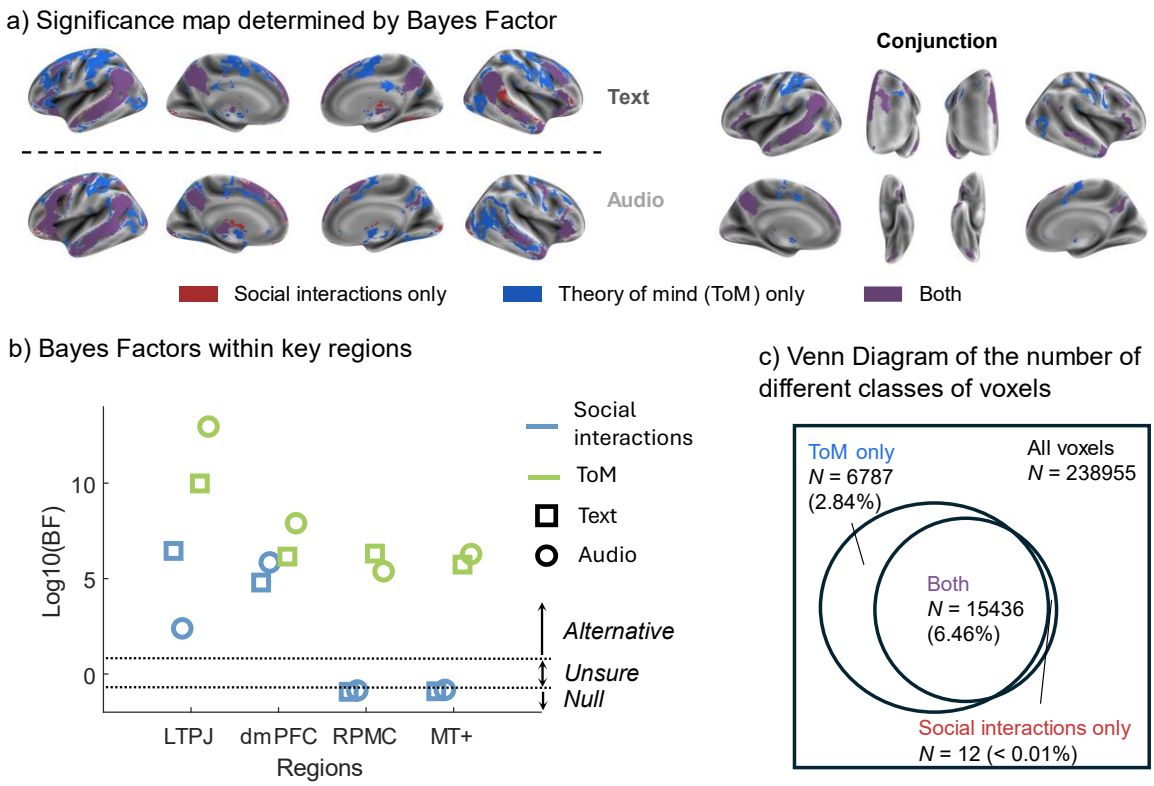


Figure 6. Classifications of voxels/brain regions by Bayes Factors. a) Surface projection of the class of each voxel (Social interactions only, ToM only, or Both). Surfaces on the left show results for Text and Audio modality separately, and surfaces on the right show the conjunction across modalities. The surface space was the same as those in previous figures but the projection method was nearest neighbor here. b) Group-level Bayes Factors for the same four regions shown in Figure 5c (also with the same abbreviations). The thresholds applied to Bayes Factors were shown by horizontal black dashed line, and the corresponding conclusions in each of the three Bayes Factor value ranges were named by italic notes on the right. For example, if one data point has a Bayes Factor larger than 5 (Log10(BF) > ~0.7), there is sufficient evidence supporting the alternative hypothesis that an effect is significantly above zero. c) A Venn diagram demonstrating the proportions of voxels that belonged to each class in the conjunction map as shown in Panel a. The percentages in parentheses refer to the ratios of the number of voxels in one class over the number of all voxels in the brain.

Discussion

Using a combination of normative subjective ratings and naturalistic neuroimaging during the experience of written narratives, we demonstrated that in closer-to-real-life experiences, perceiving social interactions and using theory of mind (ToM) co-occurred to a small to moderate

extent. They had both common and distinct neural correlates across sensory modalities. In a side analysis, we showed that self-report measurement was a valid and potentially better way to study neural correlates of ToM because it was more sensitive than the annotation method widely adopted by previous studies (e.g., Lee Masson & Isik, 2021).

In the remainder of the Discussion, we will compare our results with previous findings and seek a new interpretation beyond the processing-stage theoretical framework. Specifically, we argue that social interaction perception should be construed as a combination of perceptual and conceptual processes, that both social interaction perception and ToM involve pre-reflective inferential processes, and that ToM additionally involves deliberative inferences. We will unpack these points in more detail and consider the limitations of the current report and potential future work.

Subheader 1. Social Interaction Perception Involves Both Perceptual and Conceptual Processing

There is rich evidence about the neural correlates of visual social interaction perception and specific visual features embedded in social interactions, such as action directions (Wurm & Caramazza, 2019) and motion information (Landsiedel et al., 2022). Only recently have researchers noted the contribution of multimodal processing to neural activity underlying social interaction perception (Landsiedel & Koldewyn, 2023; Lee Masson et al., 2024b). In the current study, we extended this line of work by showing that verbal descriptions of social interactions engaged similar brain regions to visual and audiovisual presentations. The neural correlates of social interaction perception were also highly similar across the text and audio modalities, providing strong evidence that social interactions in narratives are processed in a modality-general way.

The neural correlates of social interaction perception shared across modalities included some largely visual regions such as LOC, ITC, and early visual cortex (Glasser et al., 2016). This held true even under the audio modality when there were no visual stimuli at all (Figure 3a), which suggests that perceiving social interactions in narratives partly relies on lower-level perceptual processing. Additionally, the neural correlates included many multimodal brain regions such as TPJ, STS, PC, dmPFC, IFG, and medial frontal gyrus, which are key regions involved in mentalizing (Schurz et al., 2021). They also constitute the default mode network (DMN) (Smallwood et al., 2021; Yeo et al., 2011) that has been shown to encode higher-level conceptual information during narrative experiences (e.g., Chen et al., 2017; Simony et al., 2016; Zadbood et al., 2022). This indicates that social interactions in narratives are processed in part as high-level amodal, conceptual information, and they might be an important content that people encode

during narratives (Lee Masson et al., 2024b). Given the modality-general findings about its neural correlates, social interaction perception should be more than just an early perceptual process but also involve later conceptual processing.

Subheader 2. Social Interaction Perception and ToM Both Involve Pre-Reflective Inferences

There has been a popular view that perceiving social interactions naturally invites the engagement of ToM and that the presence of social interactions improves people's ability to mentalize (e.g., Cross et al., 2016; Dziobek et al., 2006; Walter et al., 2009). However, few studies have directly tested the covariance between social interactions and ToM. By adopting the self-report method, we showed that people's tendency to use ToM positively correlated with their judgments of social interactions to a small to moderate extent. Additionally, we found that almost all brain regions responding to social interactions across modalities also responded to ToM. Because we included both social interactions and ToM in the same model and they were behaviorally separable, it is unlikely that the neural correlates of social interaction perception could be fully explained by ToM. Instead, there should be computations shared between social interaction perception and ToM that are implemented in their overlapping neural correlates.

The common neural correlates of social interaction perception and ToM included TPJ and dmPFC, which are implicated in pre-reflective, coherent, effortless experience (Lieberman, 2022). They also included other regions in the DMN whose homologous regions in rhesus monkeys have selective responses to social interactions (Sliwa & Freiwald, 2017). Given this cognitive automaticity and evolutionary primacy, the shared neural correlates may function to serve automatic, pre-reflective social inferences. The goal of these inferences is to simulate the social world around us (Mitchell, 2009) and prepare ourselves to act in it. The inferences could be engaged whenever we need to make sense of an agent outside the self, here, and now, whether for understanding social interactions and others' minds, disambiguating animations (Nguyen et al., 2019), or recollecting autobiographical memories (Rabin et al., 2010). This view argues against the processing-stage framework, indicating that ToM may involve more automatic processing than previously thought and that social interaction perception and ToM involve computations on the same processing level.

Subheader 3. ToM Involves Deliberative Inferences Limited by Cognitive Resources

Beyond the canonical ToM regions shared with social interaction perception, ToM engagement ratings in the current study were associated with neural activity in other regions, including the right anterior part of IPS, left PMC, and bilateral LOTC or MT+, which were less commonly reported in existing studies (Schurz et al., 2021). We interpret these results as implying

that ToM involves deliberative operation and manipulation over mental-state-related information. Specifically, the anterior part of IPS and PMC are important in understanding the goals and intentions of actions, a key feature of the “mirror” system (Zaki & Ochsner, 2009). The mirror system can transmit information to mPFC and lateral temporal cortex to serve higher-order mentalizing, especially when people deliberate about mental states from simpler perceptions (e.g., de Waal & Preston, 2017; Van Overwalle & Baetens, 2009). Similarly, LOTC is sensitive to action verbs (Taylor et al., 2017) and possible actions (affordances) towards a visually presented tool (Wu et al., 2020). Thus, the association between neural activity in PMC, anterior IPS, and LOTC and ToM engagement may reflect voluntary efforts for mentalizing from the action cues in narratives.

Importantly, the controlled mental processing unique to ToM should be different from pre-reflective inferences and gated by limited cognitive resources. Two observations support this view: First, the IPS and PMC are also important for executive functions, such as multitasking, which entail deliberation and reflection (Worringer et al., 2019). Second, ToM might have “recycled” existing neural infrastructure for social interaction understanding (especially the DMN) during evolution to support its automatic, core functioning (Sliwa & Freiwald, 2017). Considering that humans are capable of abstract inferences based on formal propositions but nonhuman primates are not (Deen et al., 2023), the unique ToM neural correlates outside the DMN should be supporting more cognitively demanding inferences. Thus, our results imply that ToM includes both pre-reflective and controlled inferences, with the former engaged by perceiving social interactions but not the latter.

Subheader 4. Limitations and Future Directions

Some limitations of the current study are worth mentioning. First, most of our results were based on group-level statistical inferences. Although this provided insights into the “on average” effects, there must be important individual differences omitted, especially considering the idiosyncratic nature of ToM. More work is needed in developing new ways to measure ToM engagement and relate it to neural activity on the individual level. Second, given the limited spatial resolution of fMRI, we may have missed some brain regions that only respond to social interactions across modalities. It will be helpful to investigate the dose-response profiles of the regions related to social interactions and ToM to get a finer picture of how they process the two types of information. Third, our results and interpretations do not consider action understanding as a separate psychological construct on its own, although it has shared and unique neural correlates from social interaction perception and ToM (Arioli & Canessa, 2019). Some of our results are thus likely to be attributed to action understanding. Future work should model social

interactions and ToM together with actions in the stimuli to get a finer picture of their neural correlates.

Conclusions

The current study shows that the modality-invariant processing of social interactions relies on a subset of the brain regions responding to ToM in narratives, while ToM has its unique neural correlates. This implies that both perceiving social interactions and ToM involve pre-reflective inferences, with social interaction perception having its own perceptual components and ToM having its own deliberative inference components. Future research is needed to test the reliability and generalizability of these findings and to refine the theoretical accounts for them.

Methods

Behavioral ratings of social interactions and theory of mind

Subheader 1. Participants

A total of 268 participants took part in our online experiment from Prolific.com. Of these, 32 participants failed more than half of the attention check questions, two participants provided the same ratings throughout the experiment, and three participants did not have their data saved. The remaining 231 participants with valid data (104 females, 127 males) are all healthy adults (19-45 years old, M = 33.65, SD = 6.51) whose native language is English and who have no literacy difficulties. All participants have an approval rate (proportion of valid data from studies they have taken part in) of over 95% on Prolific.

Subheader 2. Stimuli

The main stimuli are eight short narratives (stories) in the third-person perspective. Each narrative includes a single protagonist who moves along a single storyline composed of nine different situations, which are sampled from Polti's 36 dramatic situations (Polti, 1917) and characterized by interpersonal relationships and actions. During the experiment, each situation or "part" was presented in one experimental trial. The first four narratives were presented as audios, where an AI-generated adult male voice read out the narratives. The second four were presented as white texts on a black full-screen background at a constant speed of three words per second; the number of words on each screen was 10 or the number of remaining words in a part, whichever was smaller (Figure 1a). Table 1 below provides a more detailed description of the duration and length of each narrative, the presenting modality, and the duration in the unit of volumes (TRs) in the neuroimaging study.

Importantly, the variety of situations and transitions between them make each narrative contain rich descriptions of both the activity of a single character and the joint activity of multiple characters, allowing for a contrast over social interactions. Similarly, some parts of the narratives contain dense descriptions of characters' mental states (e.g., "she was furious that she did not know what was going on") while others do not (e.g., "Amy and Linda continued to be best friends"), making it possible to induce different levels of ToM engagement at different times during the narratives.

Subheader 3. Experimental Procedures

Potential participants open the study link from Prolific.com and will be directed to the webpage-based experiment hosted on Pavlovia.org. There they learn more about the experiment and provide their consent of participation before starting the experiment.

The experiment consists of four runs. Between two runs, participants could take a short break of no more than one minute. In each run, two narratives are played as texts or audios (Table 1). Because each narrative is presented in nine experimental trials, there are 18 trials in each run. At the start of the first and third runs, there is an additional practice trial for participants to get familiar with the experimental interface and ratings. One similar but irrelevant narrative part is used in the practice trials.

Each trial starts with a fixation period with pseudo-randomized durations between 2 and 8 seconds. Then a narrative part is presented, while participants need to give their ratings moment by moment by moving their mouse on top of a continuous visual analogue scale with two anchors at the ends and one anchor at the midpoint. The rating question is randomly chosen for each participant. About half of the participants have answered the social interaction question: "*Are there social interactions? (How much do you think there are social interactions in the narrative at this moment?)*", with three anchors being "*Not at all*", "*Partially*", and "*Definitely yes*". The other half of the participants have answered the theory of mind question: "*Thinking about mental states? (How much are you thinking about the mental states of characters in the narrative at this moment?)*", with three anchors being "*Not at all*", "*Half the time*", and "*All the time*". Questions in the bracket are only presented in practice trials. At the end of each trial, participants are asked to give an "intermittent" rating about the previous trial: "*On average, how much [were you thinking about the mental states of characters]/[do you think there were social interactions] during the last part of the narrative?*" The anchors are the same as the previous ratings (Figure S1). The intermittent ratings have not been analyzed or reported in this article.

The sequence of the two narratives within each run is counterbalanced across participants. To check whether participants were paying attention, in one random trial per run, a

surprise question would follow the intermittent rating and ask participants to click on a specific location of the scale.

Subheader 4. Data collection

All ratings were recorded as a float number between 0 and 100, linearly mapping onto the positions of the visual analogue scale. The moment-by-moment ratings were sampled every 83ms or every cycle of the participants' monitor refreshing, whichever was longer. Because we anticipated a delay in participants' ratings as compared to the narrative flow (e.g., a "yes" rating to social interactions might happen hundreds of milliseconds after social interactions appeared), we continued collecting ratings for another 1.5 seconds after each narrative part ended.

As stated in the Results section, researchers annotated ToM demands in the narratives offline. Because researchers did not experience the narratives continuously as participants did, only "yes/no" (1/0) labels were used for annotations, instead of continuous values. When there was a conflict of annotations, that is when two raters provided the same label while the other two opposed, the consensus label was generated after further discussion. 24 out of 172 (14%) sentence parts had a conflict of annotations.

Subheader 5. Quantification and statistical analysis

To ensure data quality, we excluded ratings of a participant from further analysis if they failed more than two out of the four attention check questions (n = 32) or kept providing the same rating throughout the experiment (n = 2). During the first few seconds of each trial, participants might not have started engaging in the narrative and providing meaningful ratings. Thus, we set as missing values all the moment-by-moment ratings sampled before the first move of the mouse in each trial. Next, to aggregate and compare moment-by-moment ratings across participants, we down-sampled them to 230ms per sample (chosen as half of the resolution of the BOLD data in the neuroimaging study) by linear interpolation. In generating a group-level summary statistic that was later applied to neuroimaging data, we excluded data from participants whose social interaction ratings had a correlation with the medians of all other participants' ratings lower than 0.3 (n = 7), as well as those whose ToM ratings had a correlation lower than 0.2 (n = 14).

To estimate the extent to which online participants had similar ratings for social interactions and ToM, two metrics were calculated. First, we randomly divided participants into two equal-size groups and correlated the mean ratings of the two groups. This split-half correlation value was corrected by the Spearman-Brown formula (Brown, 1910; Spearman, 1910). We repeated the sampling and calculation 2000 times randomly to estimate the mean and confidence interval of the correlation coefficients. Second, we calculated the pairwise correlation between

participants' ratings and found the range, mean, and median of the correlations. The latter metric is a more conservative estimate of how similar participants were in their ratings.

We also investigated the extent to which researchers' annotations of ToM demands agreed with online participants' ratings of ToM engagement. To do that, we resampled the annotations to 230ms per sample by linear interpolation. We then extracted the medians of online participants' ratings as a group-level summary and calculated its correlation with researchers' annotations. The statistical significance of the correlation was estimated by 10000 bootstrap samples with phase randomization (Lancaster et al., 2018).

Neuroimaging study on social interaction perception and theory of mind

Subheader 1. Participants

Ninety-seven participants took part in an fMRI experiment. Data from one participant was missing due to technical issues, and data from another six participants were excluded from further analysis because of failures in fieldmap collections, which are essential for the susceptibility distortion correction step in fMRI preprocessing. The remaining 90 participants with valid data (55 females, 34 males, 1 other) are healthy adults (18-45 years old, M = 24.72, SD = 5.57) without any current physical or mental disorders. All participants are native English speakers or have comparable fluency in English, and have normal or corrected-to-normal vision.

Subheader 2. Stimuli and Experimental Procedures

The neuroimaging study is part of a larger project about the neural correlates of various cognitive and affective processes (Jung et al., 2024). The project was conducted at the Dartmouth Brain Imaging Center and required participants to pay four separate visits on different days. Data analyzed here came from the "narratives" task, which was performed on the second visit of every participant. At the beginning of their second visit, participants provided their written consent to continuing participation in the study. After that, they practiced every task they would perform that day on a laptop under experimenters' instructions. All instructions were previously scripted and read out loud to each participant. When participants had no further questions about the tasks, they would lie into the scanner and begin the formal tasks.

The run structure and narratives played in each run were the same as in the behavioral study (Table 1). The main difference between the two studies was in the trial structure: fMRI participants did not provide any ratings when the narratives were presented; after one narrative part finished, they were asked to provide two ratings about their feelings and expectations. The ratings were not analyzed or reported in the current article (Figure S1).

Subheader 3. Data collection

Structural and functional MRI data were acquired on a 3T Siemens MAGNETOM Prisma MRI scanner with 32-channel parallel imaging. Structural images were acquired using high-resolution T1 spoiled gradient recall images. Functional images were acquired with a multiband EPI sequence (repetition time (TR) = 460ms, echo time (TE) = 27.2ms, field of view = 220mm, multiband acceleration factor = 8, flip angle = 44°, 64x64 matrix, 2.7x2.7x2.7mm voxels, 56 interleaved ascending slices, phase encoding posterior to anterior). For more details about the scan, see Jung et al. (2024). Stimuli were programmed and presented by Psychtoolbox (MATLAB, MathWorks) on a Linux laptop.

Subheader 4. Quantification and statistical analysis

MRI data preprocessing

Preprocessing was performed using fMRIPrep 21.0.2 (Esteban et al., 2019) based on Nipype 1.6.1 (Gorgolewski et al., 2011). To correct for magnetic field inhomogeneity, a B0-nonuniformity map (fieldmap) was estimated based on two echo-planar imaging (EPI) references with `topup` (Andersson et al., 2003, FSL 6.0.5.1:57b01774).

For each participant, one T1-weighted (T1w) image was corrected for intensity non-uniformity with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a `Nipype` implementation of the `antsBrainExtraction.sh` workflow using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (Zhang et al., 2001, FSL 6.0.5.1:57b01774, RRID:SCR_002823). Brain surfaces were reconstructed by `recon-all` (Dale et al., 1999, FreeSurfer 6.0.1,RRID:SCR_001847). The brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017, RRID:SCR_002438). Volume-based spatial normalization was performed through nonlinear registration with `antsRegistration` (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. Two templates were selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* (MNI152NLin2009cAsym, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym) and *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* (RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym).

For each of the four BOLD runs per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated by aligning and averaging

1 single-band reference (SBRefs) using a custom method of fMRIPrep. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (Jenkinson et al., 2002, FSL 6.0.5.1:57b01774). The estimated fieldmap was then aligned with rigid registration to the target EPI reference run. The field coefficients were mapped on to the reference EPI using the transform. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009), generating a preprocessed BOLD run in MNI152NLin2009cAsym space. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Several confounding time series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS were calculated for each functional run both using their implementations in `Nipype` following the definitions by Power and colleagues (Power et al., 2014). The three global signals were extracted within the CSF, the WM, and the whole-brain masks. The six head-motion estimates were expanded with the inclusion of temporal derivatives and quadratic terms for each (Lund et al., 2006). Gridded (volumetric) resampling was performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Lastly, the functional data were spatially smoothed with a 6 mm full-width-at-half-maximum (FWHM) Gaussian kernel.

Voxel-wise general linear model (GLM)

To find the neural correlates of variables of interest (social interactions and ToM), we applied a hierarchical random effect model approach using two-stage summary statistics. Specifically, in each voxel, we fit a GLM for each participant using the ordinary least squares method. We then aggregated the regression coefficients in each voxel across participants and performed one-sample t-tests and robust regression (Holland & Welsch, 1977) on single coefficients or planned contrasts over several coefficients. Results from t-tests and robust regression were almost identical and we only reported those from t-tests. To find labels for the significant brain regions on the cortex, we overlaid the thresholded *t* maps on a multi-modal parcellation of the human cerebral cortex (Glasser et al., 2016) and calculated the percentage of significant voxels within each parcel. Parcels with more than 25% of significant voxels were treated as significant, and adjacent parcels were grouped into larger regions with more commonly used names (such as TPJ instead of PGi and PGs).

Variables of interest came from two different sources: online participants' moment-by-moment ratings and researchers' offline annotations. The medians of online participants' ratings

were linearly projected from the range [0, 100] to [-1, 1] before being entered as regressors in the GLM. To ensure the medians were valid and meaningful, for each trial, we excluded data from participants who had less than 40% of non-missing moment-by-moment ratings. We also excluded the time points when less than 50% of the participants had non-missing rating values by adding indicator functions to the GLM whose values were 1 at the excluded time points and 0 at all other time points. Researchers' annotations of ToM demands were coded as two separate binary regressors with values 1 or 0. One regressor modeled the sentence parts where there were ToM demands and the other modeled where there were no ToM demands; for a sentence part with ToM demands, the first regressor had a value of 1 and the second a value of 0. Thus, a contrast between the estimates of the two regressors revealed the effects of ToM demands.

To control for the effect of answering questions on neural activity, we added a binary regressor whose values were 1 when participants were giving ratings at the end of each trial and 0 otherwise. This rating regressor, together with all variables of interest, was convolved with the canonical hemodynamic response function (Glover, 1999). This made the fixation periods the only implicit baseline of the BOLD signals.

Some further data denoising was performed in the GLM. First, we high-pass-filtered the functional data at 1/128 Hz by adding a series of discrete cosine functions to each model. Second, we added 24 head-motion-related parameters to account for the effects of head motions, including six head motion parameters (x, y, and z translations, and pitch, yaw, and roll), their temporal derivatives, their squares, and the squares of the derivatives. Third, we added the mean signal in the cerebral-spinal fluids as an approximation of physiological artifacts. Lastly, we identified outlier brain volumes by searching for the volumes whose average signals or framewise displacements were outside three standard deviations from the mean of each run. We excluded those volumes by entering each of them as an indicator function whose values were 1 at the time point of the outlier and 0 at all other time points.

Because the functional signals at the start of each run may be unstable and deviate from the rest of the run, we discarded the first 6 volumes in each run when there were no stimuli on the screen. The rest of each run's functional data were rescaled to have a grand mean of 100 (across all volumes and voxels) to account for differences in signal baselines across runs and participants (Chen et al., 2017). Functional data were concatenated across all four runs, as were the design matrices. Variables of interest, as well as the rating regressor, were shared across runs, such that only one beta value was estimated for each variable of interest in each voxel. All other variables, including the intercept, were modeled separately for each run.

Mask-based analysis

To understand whether our fMRI results were congruent with existing findings, we performed analysis within several brain masks defined by results from previous studies. The first way to find brain masks was to use term-based association maps from Neurosynth.org (Yarkoni et al., 2011), whose z-scores from a two-way ANOVA quantify the possibility of each voxel being positively activated in a paper that includes a specific term. The original maps were thresholded at FDR $q < .01$, and we further thresholded them by $z >= 3$ to reduce false positive voxels and make the masks more specific. The terms used in the current study include: “social interaction”, “social cognition”, “psts”, “tom”, and “mentalizing”.

The second way to find brain masks was to use the ToM group maps from a false belief task (Dufour et al., 2013). The masks include bilateral TPJ (combining the original “LTPJ” and “RTPJ” maps), RSTS (the original “RSTS” map), bilateral PC (the original “PC” map), and bilateral dmPFC (combining the original “DMPFC” and “MMPFC” maps). We did not use the “VMPFC” map because no GLM results in the current study showed significant effects in the ventral medial prefrontal cortex.

The third way was to define region masks based on a multimodal parcellation of the cerebral cortex (Glasser et al., 2016). Each mask was a union of several parcels based on their anatomical positions and functional topography. For a full list of masks used in the current study and how we defined them, see Table 2. Within each mask defined in any of the three ways, we extracted individual-level regression coefficients from the GLM analysis and compared the group mean to zero to investigate whether a variable of interest had consistent neural activations across participants in that mask.

Bayes Factor analysis

To identify voxels associated with both social interactions and ToM, and those associated with one but not the other, we took the Bayes Factor (BF) approach to quantify evidence for both alternative and null hypotheses. Specifically, the BF value of a given voxel is proportional to the ratio of the likelihoods under the alternative and null hypotheses (e.g., the neural activity in that voxel is correlated with social interactions in narratives or not). We estimated BFs from *t* values of one-sample tests, assuming a Jeffreys-Zellner-Siow (JZS) prior, following the formula provided by Rouder et al. (2009):

$$BF_{10} = \frac{\int_0^\infty (1+Ngr^2)^{-1/2}(1+\frac{t^2}{(1+Ngr^2)(N-1)})^{-N/2}(2\pi)^{-1/2}g^{-3/2}e^{-1/(2g)}dg}{(1+\frac{t^2}{N-1})^{-N/2}}$$

In the formula, *t* represents the *t*-statistic, *N* the sample size, and *r* the scale factor. We set the scale factor to $\sqrt{2}/2$, assuming a moderate effect a priori (Rouder et al., 2009), to increase the

support for alternative hypotheses given the frequently smaller effect size in fMRI univariate analysis (see also Bo et al., 2024).

BFs calculated in this manner are asymmetric for alternative and null hypotheses, because evidence for null hypotheses is bound by $t = 0$. Given the current sample size ($N = 90$), the BF cannot be smaller than .12 but can be infinitely large. Thus, we chose a heuristic threshold of 5 for the alternative hypothesis and 1/5 for the null hypothesis to attain moderate to strong evidence in favor of both hypotheses.

After calculating the BF for a voxel under one modality, we classified the voxel into one or none of the three categories (Figure 6). Because the BF threshold for alternative hypotheses corresponds to a t value of 2.84, which is lower than the critical t value (3.16) for FDR $q < .01$, we further bound the alternative results by the significance in the FDR corrected analysis. Besides, we only considered positive effects (positive t statistics) in this analysis. The specific criteria are as follows:

- “Both” – BF > 5 and $t > 3.16$ for both Social interactions and ToM.
- “Social interactions only” – BF > 5 and $t > 3.16$ for Social interactions, and BF < 1/5 for ToM.
- “ToM only” – BF > 5 and $t > 3.16$ for ToM, and BF < 1/5 for Social interactions.

To create the conjunction classification map across modalities (Figure 6a), we assigned a label to one voxel in the conjunction map if and only if it had the same label under the two modalities. This provides a more conservative and specific estimate of the modality-general neural system responding to social interactions only, ToM only, or both.

References

Ainsworth, M., Sallet, J., Joly, O., Kyriazis, D., Kriegeskorte, N., Duncan, J., Schüffelgen, U., Rushworth, M. F. & Bell, A. H. (2021). Viewing ambiguous social interactions increases functional connectivity between frontal and temporal nodes of the social brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *41*(28), 6070–6086. https://doi.org/10.1523/JNEUROSCI.0870-20.2021

Andersson, J. L. R., Skare, S. & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7

Arioli, M. & Canessa, N. (2019). Neural processing of social interaction: Coordinate-based meta-analytic evidence from human neuroimaging studies. *Human Brain Mapping*, *40*(13), 3712–3737. https://doi.org/10.1002/hbm.24627

Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*(1), 289–300. http://www.jstor.org/stable/2346101

Bo, K., Kraynak, T. E., Kwon, M., Sun, M., Gianaros, P. J. & Wager, T. D. (2024). A systems identification approach using Bayes factors to deconstruct the brain bases of emotion regulation. *Nature Neuroscience*, *27*(5), 975–987. https://doi.org/10.1038/s41593-024-01605-7

Bohl, V. & van den Bos, W. (2012). Toward an integrative account of social cognition: marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, *6*, 274. https://doi.org/10.3389/fnhum.2012.00274

Brown, W. (1910). Some experimental results in the correlation of mental abilities1. *British Journal of Psychology* , *3*, 296–322. https://doi.org/10.1111/J.2044-8295.1910.TB00207.X

Chen, G., Taylor, P. A. & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage*, *147*, 952–959. https://doi.org/10.1016/j.neuroimage.2016.09.066

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A. & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W. & de C Hamilton, A. F. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the*

Royal Society of London. Series B, Biological Sciences, *371*(1686), 20150075. https://doi.org/10.1098/rstb.2015.0075

Dale, A. M., Fischl, B. & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

de Waal, F. B. M. & Preston, S. D. (2017). Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews. Neuroscience*, *18*(8), 498–509. https://doi.org/10.1038/nrn.2017.72

Deen, B., Schwiedrzik, C. M., Sliwa, J. & Freiwald, W. A. (2023). Specialized Networks for Social Cognition in the Primate Brain. *Annual Review of Neuroscience*, *46*, 381–401. https://doi.org/10.1146/annurev-neuro-102522-121410

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

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T. & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A. & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Frith, C. & Frith, U. (2005). Theory of mind. *Current Biology: CB*, *15*(17), R644-6. https://doi.org/10.1016/j.cub.2005.08.041

Georgescu, A. L., Kuzmanovic, B., Santos, N. S., Tepest, R., Bente, G., Tittgemeyer, M. & Vogeley, K. (2014). Perceiving nonverbal behavior: neural correlates of processing movement fluency and contingency in dyadic interactions. *Human Brain Mapping*, *35*(4), 1362–1378. https://doi.org/10.1002/hbm.22259

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M. & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, *9*(4), 416–429. https://doi.org/10.1006/nimg.1998.0419

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L. & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Greve, D. N. & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Heider, F. & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Holland, P. W. & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics: Theory and Methods*, *6*(9), 813–827. https://doi.org/10.1080/03610927708827533

Isik, L., Koldewyn, K., Beeler, D. & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), E9145–E9152. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2015.11.025

Jenkinson, M., Bannister, P., Brady, M. & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8

Jung, H., Amini, M., Hunt, B. J., Murphy, E. I., Sadil, P., Halchenko, Y. O., Petre, B., Miao, Z., Kragel, P. A., Han, X., Heilicher, M. O., Sun, M., Collins, O. G., Lindquist, M. A. & Wager, T. D. (2024). A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks. *bioRxiv.Org: The Preprint Server for Biology*, 2024.06.21.599974. https://doi.org/10.1101/2024.06.21.599974

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E. & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, *13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R. & Nummenmaa, L. (2012). Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in Human Neuroscience*, *6*, 233. https://doi.org/10.3389/fnhum.2012.00233

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V. & Stefanovska, A. (2018). Surrogate data for hypothesis testing of physical systems. *Physics Reports*, *748*, 1–60. https://doi.org/10.1016/j.physrep.2018.06.001

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76–85. https://doi.org/10.1137/0701007

Landsiedel, J., Daughters, K., Downing, P. E. & Koldewyn, K. (2022). The role of motion in the neural representation of social interactions in the posterior temporal cortex. *NeuroImage*, *262*, 119533. https://doi.org/10.1016/j.neuroimage.2022.119533

Landsiedel, J. & Koldewyn, K. (2023). Auditory dyadic interactions through the "eye" of the social brain: How visual is the posterior STS interaction region? *Imaging Neuroscience (Cambridge, Mass.)*, *1*, 1–20. https://doi.org/10.1162/imag_a_00003

Lee Masson, H., Chang, L. & Isik, L. (2024a). Multidimensional neural representations of social features during movie viewing. *Social Cognitive and Affective Neuroscience*, *19*(1), nsae030. https://doi.org/10.1093/scan/nsae030

Lee Masson, H., Chen, J. & Isik, L. (2024b). A shared neural code for perceiving and remembering social interactions in the human superior temporal sulcus. *Neuropsychologia*, *196*(108823), 108823. https://doi.org/10.1016/j.neuropsychologia.2024.108823

Lee Masson, H. & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, *245*, 118741. https://doi.org/10.1016/j.neuroimage.2021.118741

Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, *58*(1), 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lieberman, M. D. (2022). Seeing minds, matter, and meaning: The CEEing model of pre-reflective subjective construal. *Psychological Review*, *129*(4), 830–872. https://doi.org/10.1037/rev0000362

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L. & Nichols, T. E. (2006). Non-white noise in fMRI: does modelling have an impact? *NeuroImage*, *29*(1), 54–66. https://doi.org/10.1016/j.neuroimage.2005.07.005

Mak, M., Faber, M. & Willems, R. M. (2023). Different kinds of simulation during literary reading: Insights from a combined fMRI and eye-tracking study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *162*, 115–135. https://doi.org/10.1016/j.cortex.2023.01.014

McMahon, E. & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.09.001

McMahon, E., Bonner, M. F. & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology: CB*, *33*(23), 5035-5047.e8. https://doi.org/10.1016/j.cub.2023.10.015

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1309–1316. https://doi.org/10.1098/rstb.2008.0318

Moessnang, C., Schäfer, A., Bilek, E., Roux, P., Otto, K., Baumeister, S., Hohmann, S., Poustka, L., Brandeis, D., Banaschewski, T., Meyer-Lindenberg, A. & Tost, H. (2016). Specificity, reliability and sensitivity of social brain responses during spontaneous mentalizing. *Social Cognitive and Affective Neuroscience*, *11*(11), 1687–1697. https://doi.org/10.1093/scan/nsw098

Nguyen, M., Vanderwal, T. & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, *184*, 161–170. https://doi.org/10.1016/j.neuroimage.2018.09.010

Nijhof, A. D. & Willems, R. M. (2015). Simulating fiction: individual differences in literature comprehension revealed with FMRI. *PloS One*, *10*(2), e0116492. https://doi.org/10.1371/journal.pone.0116492

Pitcher, D. & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Polti, G. (1917). *The thirty-six dramatic situations*. Editor Company.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Rabin, J. S., Gilboa, A., Stuss, D. T., Mar, R. A. & Rosenbaum, R. S. (2010). Common and unique neural correlates of autobiographical memory and theory of mind. *Journal of Cognitive Neuroscience*, *22*(6), 1095–1111. https://doi.org/10.1162/jocn.2009.21344

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1027. https://doi.org/10.1038/s41467-018-03399-2

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

Saxe, R. & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind.” *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/s1053-8119(03)00230-1

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R. & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730. https://doi.org/10.1016/j.neuropsychologia.2005.07.017

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J. & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A. & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*, 12141. https://doi.org/10.1038/ncomms12141

Sliwa, J. & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science*, *356*(6339), 745–749. https://doi.org/10.1126/science.aam6383

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E. & Margulies, D. S. (2021). The default mode network in cognition: a topographical perspective. *Nature Reviews. Neuroscience*, *22*(8), 503–513. https://doi.org/10.1038/s41583-021-00474-4

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology* , *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Spunt, R. P. & Adolphs, R. (2014). Validating the Why/How contrast for functional MRI studies of Theory of Mind. *NeuroImage*, *99*, 301–311. https://doi.org/10.1016/j.neuroimage.2014.05.023

Taylor, L. J., Evans, C., Greer, J., Senior, C., Coventry, K. R. & Ietswaart, M. (2017). Dissociation between Semantic Representations for Motion and Action Verbs: Evidence from Patients with Left Hemisphere Lesions. *Frontiers in Human Neuroscience*, *11*, 35. https://doi.org/10.3389/fnhum.2017.00035

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A. & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Van Overwalle, F. & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *NeuroImage*, *48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Varrier, R. S. & Finn, E. S. (2022). Seeing social: A neural signature for conscious perception of social interactions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0859-22.2022

Wagner, D. D., Kelley, W. M., Haxby, J. V. & Heatherton, T. F. (2016). The Dorsal Medial Prefrontal Cortex Responds Preferentially to Social Interactions during Natural Viewing. *The Journal of*

Neuroscience: The Official Journal of the Society for Neuroscience, *36*(26), 6917–6925. https://doi.org/10.1523/JNEUROSCI.4220-15.2016

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

Walter, H., Ciaramidaro, A., Adenzato, M., Vasic, N., Ardito, R. B., Erk, S. & Bara, B. G. (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. *Social Cognitive and Affective Neuroscience*, *4*(2), 166–176. https://doi.org/10.1093/scan/nsn047

Wolf, D., Mittelberg, I., Rekittke, L.-M., Bhavsar, S., Zvyagintsev, M., Haeck, A., Cong, F., Klasen, M. & Mathiak, K. (2018). Interpretation of Social Interactions: Functional Imaging of Cognitive-Semiotic Categories During Naturalistic Viewing. *Frontiers in Human Neuroscience*, *12*, 296. https://doi.org/10.3389/fnhum.2018.00296

Worringer, B., Langner, R., Koch, I., Eickhoff, S. B., Eickhoff, C. R. & Binkofski, F. C. (2019). Common and distinct neural correlates of dual-tasking and task-switching: a meta-analytic review and a neuro-cognitive processing model of human multitasking. *Brain Structure & Function*, *224*(5), 1845–1869. https://doi.org/10.1007/s00429-019-01870-4

Wu, W., Wang, X., Wei, T., He, C. & Bi, Y. (2020). Object parsing in the left lateral occipitotemporal cortex: Whole shape, part shape, and graspability. *Neuropsychologia*, *138*, 107340. https://doi.org/10.1016/j.neuropsychologia.2020.107340

Wurm, M. F. & Caramazza, A. (2019). Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *NeuroImage*, *202*, 116153. https://doi.org/10.1016/j.neuroimage.2019.116153

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H. & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Zadbood, A., Nastase, S., Chen, J., Norman, K. A. & Hasson, U. (2022). Neural representations of naturalistic events are updated as our understanding of the past changes. *eLife*, *11*. https://doi.org/10.7554/eLife.79045

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

Zhang, Y., Brady, J. M. & Smith, S. M. (2001). An hmrf-em algorithm for partial volume segmentation of brain mri fmrib technical report tr01yz1. *Brain: A Journal of Neurology*. https://www.fmrib.ox.ac.uk/datasets/techrep/tr01yz1/tr01yz1.pdf

Table 1

Properties of the narrative stimuli

Run #	Narrative #	Modality	Duration (mm:ss)	Number of TRs	Number of words
1	7	audio	2:01	263	429
1	8	audio	1:36	209	314
2	5	audio	2:30	326	511
2	6	audio	2:04	269	447
3	3	text	3:11	415	573
3	4	text	3:29	454	626
4	1	text	2:43	355	490
4	2	text	2:40	348	480

Note. The table is sorted by run number followed by narrative number. The sequence of the two narratives within a run was counterbalanced across participants.

Table 2

Masks defined on parcellations from Glasser et al. (2016)

Mask	Parcels included	Hemisphere
LTPJ	PGi, PGs	Left
dmPFC	d32, 8BM, 9m	Bilateral
RPMC	FEF, PEF, 6d, 6v, 6r, 6a	Right
MT+	V4t, FST	Bilateral
PC	7m, v23ab, d23ab, 31pv, 31pd	Bilateral
LSTS	STSda, STSdp, STSva, STSvp, TPOJ1, TPOJ2	Left
RTPJ	PGi, PGs	Right
SMA	SCEF, 6ma, 6mp	Bilateral

Note. A “Bilateral” entry in the Hemisphere column denotes that the parcels on both hemispheres were concatenated to form a single brain mask for analysis. “Left” and “Right” denote that parcels on only one hemisphere were included.

Figure Captions

Figure 1. Study overview. a) The stimuli used in this study were eight narratives, with the first four presented as audios and the last four presented as texts. One group of participants (*N* = 90) experienced the narratives during functional magnetic resonance imaging (fMRI) scan. Another group of participants (*N* = 231) experienced the narratives online on their personal computers and simultaneously provided moment-by-moment ratings about social interactions in the narratives (*N* = 114) or their use of theory of mind (ToM) (*N* = 117). The black rectangles illustrate the screen that participants saw when narratives were presented. One online participant only saw one of the two questions shown in this figure throughout the study. We concatenated blood-oxygen-level-dependent (BOLD) signals in fMRI scanning (one fake BOLD time series was plotted for illustration only) and the medians of online ratings across all trials (actual data were plotted). Then we located the neural correlates of social interaction perception and ToM by two-stage voxel-wise general linear models. Shaded areas around the lines of online ratings indicate confidence intervals calculated by median absolute difference. b) An example narrative excerpt extracted from the eighth trial of Narrative #1 (see Supplementary Materials for all narrative texts). The medians of social interaction and ToM ratings during this excerpt, as well as researchers' annotations of ToM demands, were displayed as the highlight color of each word. The approximate time window of this excerpt was shown by the horizontal gray bar in Panel a.

Figure 2. Ratings of online participants about social interactions and theory of mind (ToM). a) and b) In the line plots on the left, the moment-by-moment ratings of one participant were plotted as a thin quasi-transparent line. Thicker/darker parts indicate that more participants shared ratings at those time points. In the histograms on the right, all ratings were binned and plotted. c) In the line plots on the left, medians of ratings were shown by solid lines, and confidence intervals of the medians were shown by shaded areas around the line. We divided all ratings into two modalities and showed them in two plots. Note: the lines have many discontinuity points because a few data points at the start of each trial were removed from further analysis due to inadequate valid ratings (see Methods for details) and we show the version after removal in this figure. In the scatter plot on the right, each dot corresponds to a time point in the line plots. The cyan line is the best fit line estimated by ordinary least squares regression (r = .323).

Figure 3. Neural correlates of social interaction perception in narratives. a) Group-level effect maps of social interactions as the medians of online participants' ratings, thresholded at FDR $q < .01$. Maps on the left showed separate patterns for Text and Audio modality, and maps on the right showed the conjunction across modalities. We did analyses in the volumetric space and projected results onto the Freesurfer average surface maps (Fischl, 2012) by linear interpolation for visualization only. The colormap of *t* values applied to all maps. We annotated several key regions that were significant in both hemispheres and both modalities. Abbreviations: TPJ – temporoparietal junction; STS – superior temporal sulcus; PC – precuneus; dmPFC – dorsomedial prefrontal cortex; IFG – inferior frontal gyrus; LOC – lateral occipital cortex; ITC – inferior temporal cortex. b) Group-level effects of social interactions in two modalities within three brain masks. Parts of each mask were illustrated by the yellowish-green fill color in the surface plots of the right hemisphere; the whole masks were used for analysis. Bar heights indicate the means of regression coefficients (betas) from the first-level models within each mask across participants, which reflect the strength of associations between social interaction ratings and neural activity within each mask. Error bars indicate 95% confidence intervals. *** indicates $p < .001$ in one-sample *t*-tests against 0 (after Bonferroni correction).

Figure 4. Neural correlates of theory of mind (ToM) engagement and ToM demands. a) Group-level effect maps of ToM as the medians of online participants' ratings of ToM engagement and the offline annotations of ToM demands, thresholded at FDR $q < .01$. We only displayed the conjunction maps across modalities for simplicity. The method for projection from volumetric space to surface space and abbreviations were the same as in Figure 3a. b) Group-level comparisons of the first-level beta estimates for ToM engagement ("Engagement") and ToM demands ("Demands") in two modalities within two brain masks. Bar heights indicate means of standardized betas and error bars indicate 95% confidence intervals. Parts of the masks on the right hemisphere were indicated by the yellowish-green fill color in the surface plots. c) Group-level spatial pattern correlations between beta maps and Neurosynth maps within each mask. Bar heights indicate the means of Fisher *z* values transformed from Pearson's correlation coefficients and error bars indicate 95% confidence intervals. In all graphs, *** indicates $p < .001$ (after Bonferroni correction).

Figure 5. Neural activity associated with social interaction perception and theory of mind (ToM). a) and b) Group-level effect maps of medians of social interaction and ToM ratings after controlling for each other, thresholded at FDR $q < .01$. The method for projection from volumetric space to surface space was the same as Figures 3a and 4a. Surfaces on the left show neural correlates under Text and Audio modality separately, while surfaces on the right show the conjunction across modalities. The color bar of *t* values applies to all surfaces. c) Bar plots of the brain-behavior associations in four regions of interest. Bar heights indicate the means of first-level betas divided by their standard deviation, error bars indicate 95% confidence intervals, and shaded scatters indicate estimates for each participant. Abbreviations: std – standard deviation, LTPJ – left temporoparietal junction, dmPFC – dorsomedial prefrontal cortex, RPMC – right premotor cortex, MT+ – middle temporal and medial superior temporal area, Soc – social interactions.

Figure 6. Classifications of voxels/brain regions by Bayes Factors. a) Surface projection of the class of each voxel (Social interactions only, ToM only, or Both). Surfaces on the left show results for Text and Audio modality separately, and surfaces on the right show the conjunction across modalities. The surface space was the same as those in previous figures but the projection method was nearest neighbor here. b) Group-level Bayes Factors for the same four regions shown in Figure 5c (also with the same abbreviations). The thresholds applied to Bayes Factors were shown by horizontal black dashed line, and the corresponding conclusions in each of the three Bayes Factor value ranges were named by italic notes on the right. For example, if one data point has a Bayes Factor larger than 5 (Log10(BF) > ~0.7), there is sufficient evidence supporting the alternative hypothesis that an effect is significantly above zero. c) A Venn diagram demonstrating the proportions of voxels that belonged to each class in the conjunction map as shown in Panel a. The percentages in parentheses refer to the ratios of the number of voxels in one class over the number of all voxels in the brain.

Figure 1.

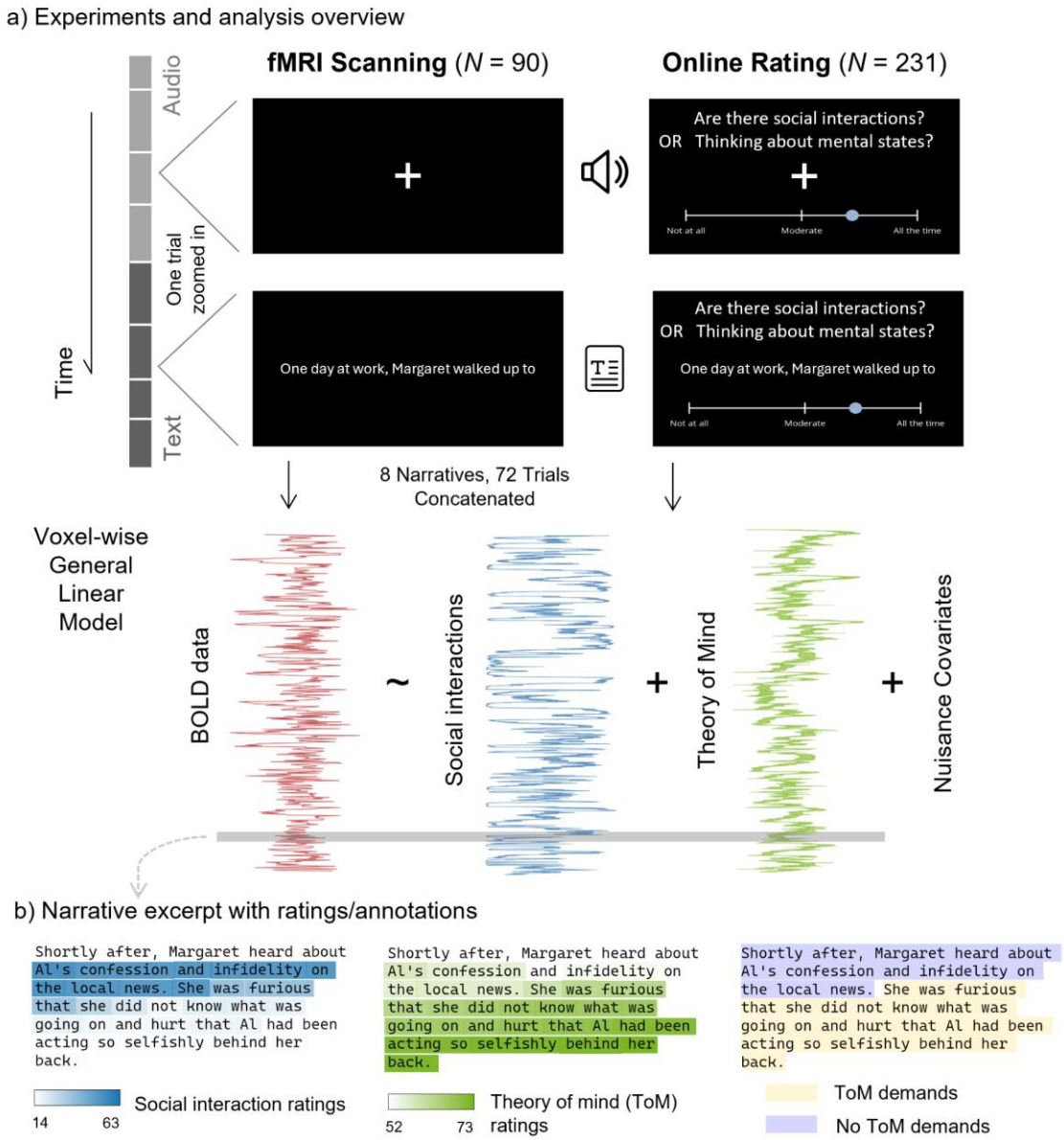

a) Experiments and analysis overview
fMRI Scanning (N = 90)
Online Rating (N = 231)
Audio
One trial zoomed in
Text
Time
Are there social interactions?
OR Thinking about mental states?
Not at all
Moderate
All the time
One day at work, Margaret walked up to
8 Narratives, 72 Trials Concatenated
Voxel-wise General Linear Model
BOLD data
~
Social interactions
+
Theory of Mind
+
Nuisance Covariates
b) Narrative excerpt with ratings/annotations
Shortly after, Margaret heard about Al's confession and infidelity on the local news. She was furious that she did not know what was going on and hurt that Al had been acting so selfishly behind her back.
Social interaction ratings
14
63
Theory of mind (ToM) ratings
52
73
ToM demands
No ToM demands

Figure 2.

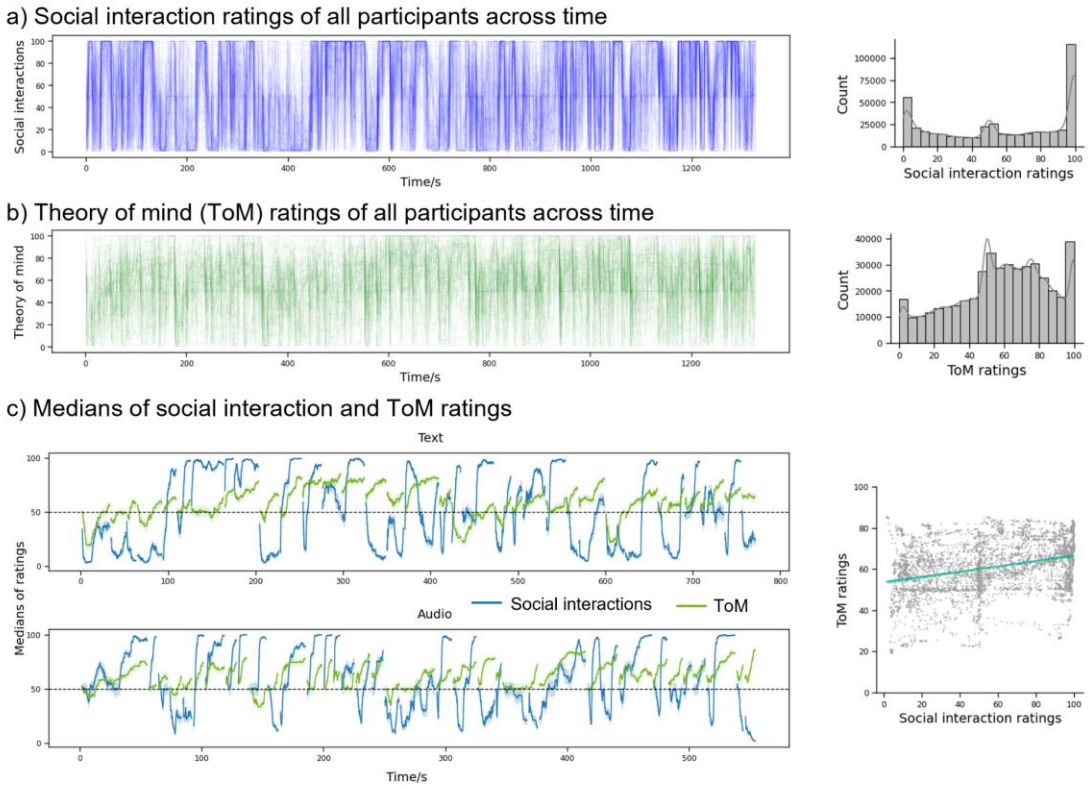

a) Social interaction ratings of all participants across time
Social interactions
Time/s
Count
Social interaction ratings
b) Theory of mind (ToM) ratings of all participants across time
Theory of mind
Time/s
Count
ToM ratings
c) Medians of social interaction and ToM ratings
Text
Audio
Medians of ratings
Social interactions
ToM
Time/s
ToM ratings
Social interaction ratings

Figure 3.

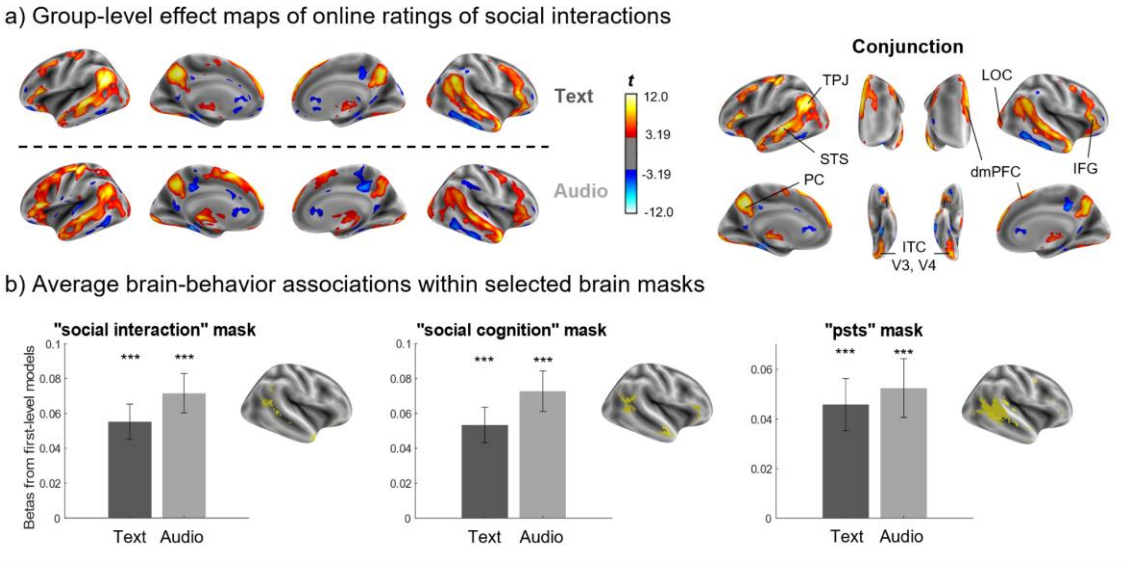

a) Group-level effect maps of online ratings of social interactions
Text
Audio
t
12.0
3.19
-3.19
-12.0
Conjunction
TPJ
LOC
STS
PC
dmPFC
IFG
ITC
V3, V4
b) Average brain-behavior associations within selected brain masks
"social interaction" mask
"social cognition" mask
"psts" mask
Betas from first-level models
***
Text
Audio

Figure 4.

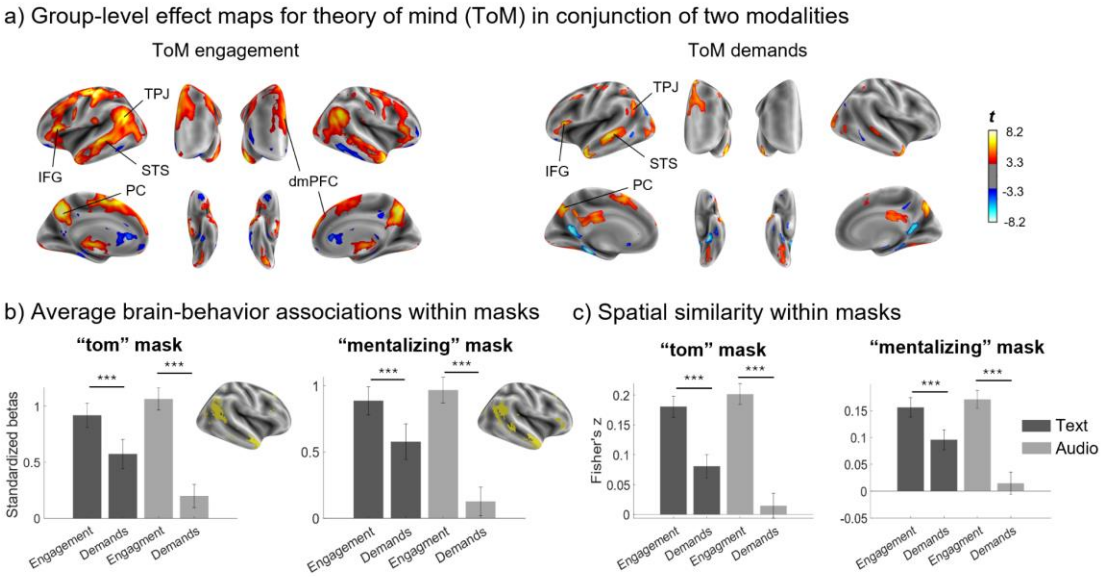

a) Group-level effect maps for theory of mind (ToM) in conjunction of two modalities
ToM engagement
ToM demands
TPJ
IFG
STS
PC
dmPFC
t
8.2
3.3
-3.3
-8.2
b) Average brain-behavior associations within masks
c) Spatial similarity within masks
"tom" mask
"mentalizing" mask
Standardized betas
Fisher's z
Text
Audio
Engagement
Demands
Engagment
***

Figure 5.

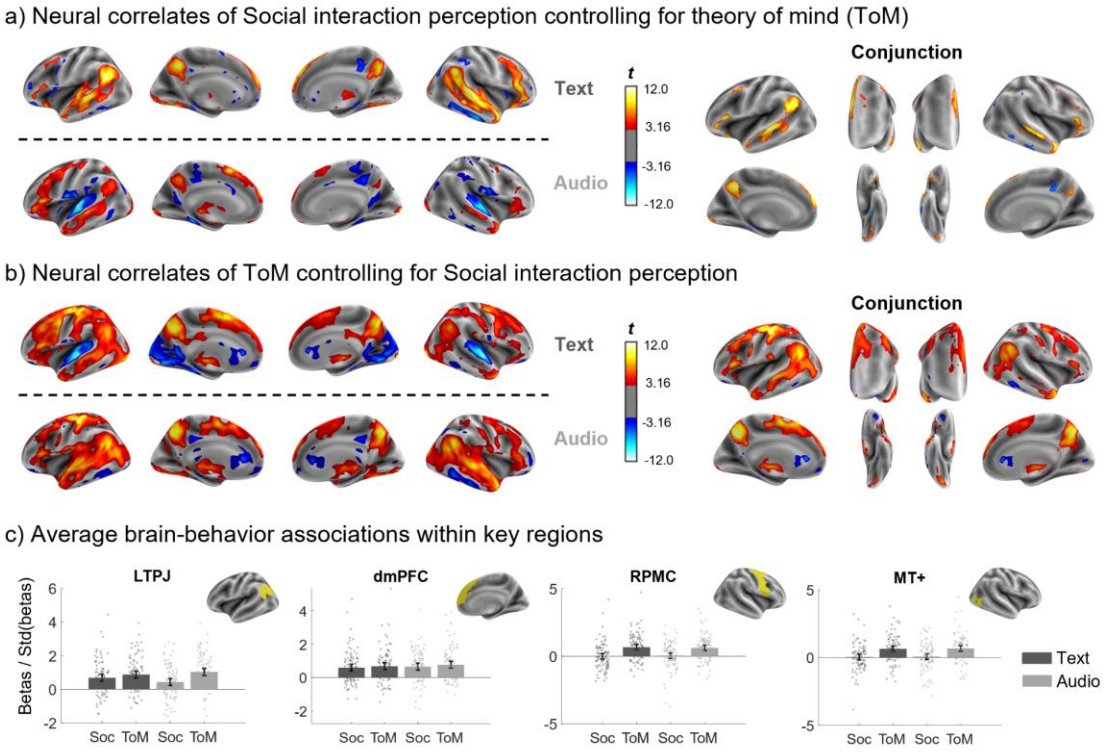
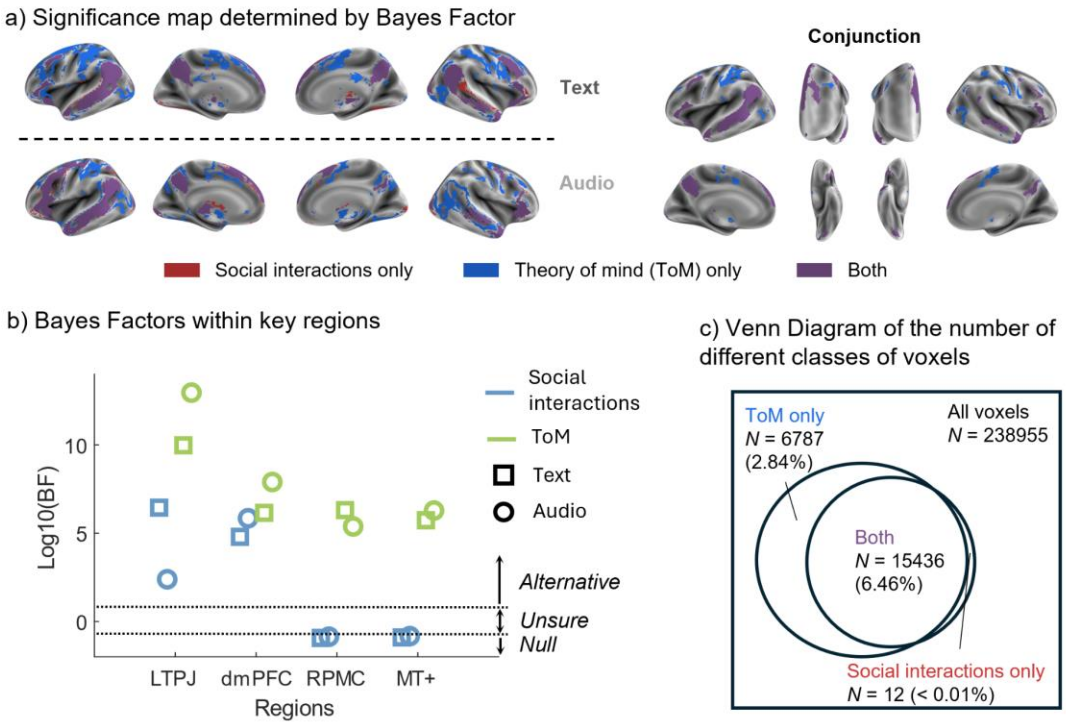

a) Neural correlates of Social interaction perception controlling for theory of mind (ToM)
Conjunction
Text
Audio
t
12.0
3.16
-3.16
-12.0
b) Neural correlates of ToM controlling for Social interaction perception
Conjunction
Text
Audio
t
12.0
3.16
-3.16
-12.0
c) Average brain-behavior associations within key regions
LTPJ
dmPFC
RPMC
MT+
Betas / Std(betas)
Soc ToM Soc ToM
Text
Audio

Figure 6.


a) Significance map determined by Bayes Factor
Conjunction
Text
Audio
Social interactions only
Theory of mind (ToM) only
Both
b) Bayes Factors within key regions
Log10(BF)
10
5
0
LTPJ
dmPFC
RPMC
MT+
Regions
Social interactions
ToM
Text
Audio
Alternative
Unsure
Null
c) Venn Diagram of the number of different classes of voxels
ToM only
N = 6787
(2.84%)
All voxels
N = 238955
Both
N = 15436
(6.46%)
Social interactions only
N = 12 (< 0.01%)