Brain & Language 212 (2021) 104878

Understanding particularized and generalized conversational implicatures: Is theory-of-mind necessary?

Wangshu Feng [a,b], Hongbo Yu [c], Xiaolin Zhou [b,d,e,f,*]

[a] Research Institute of Foreign Languages, Beijing Foreign Studies University, Beijing 100089, China
[b] School of Psychological and Cognitive Sciences, Peking University, Beijing 1008 1, China
[c] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, C 9 10 9 0, US
[d] Institute of Linguistics, Shanghai International Studies University, Shanghai 0008 , China
[e] Key Laboratory of Behavior and Mental Health, Peking University, Beijing 1008 1, China
[f] U I c overn Institute for Brain Research, Peking University, Beijing 1008 1, China

ARTICLE INFO

Keywords
Conversational implicature
Theory of mind
fMRI
tDCS
dmPFC

ABSTRACT

A speaker's intended meaning can be inferred from an utterance with or without reference to its context for particularized implicature (PI) and/or generalized implicature (GI). Although previous studies have separately revealed the neural correlates of PI and GI comprehension, it remains controversial whether they share theory-of-mind (ToM) related inferential processes. Here we address this issue using functional MRI (fMRI) and transcranial direct current stimulation (tDCS). Participants listened to single-turn dialogues where the reply was indirect with either PI or GI or was direct for control conditions (i.e., PIC and GIC). Results showed that PI and GI comprehension shared the multivariate fMRI patterns of language processing; in contrast, the ToM-related pattern was only elicited by PI comprehension, either at the whole-brain level or within dorsal medial prefrontal cortex (dmPFC). Moreover, stimulating right TPJ exclusively affected PI comprehension. These findings suggest that understanding PI, but not GI, requires ToM-related inferential processes.

1. Introduction

Imagine that Pat asks the hotel's front-desk clerk about where his friend went. The clerk responds by saying: "Some of guests are already leaving". In this conversation, the listener needs not only to decode the context-invariant "sentence meaning", but also to infer the implicated meaning (conversational implicature) beyond the literal expression (Grice, 1989; Hagoort & Levinson, 2014; Noveck & Reboul, 2008)k r which can be further classified into pa convey (Sperber & Wilson, 1986). That is to say, both PI and GI are derived from the same cognitive processes, which take contextual considerations into account from the beginning. Finally, Semantic Minimalism o hin

Table 1
Examples of the dialogue scenarios in the four experimental conditions, translated into English.

Condition	Cover Story	Dialogue
PI	In a movie city, a director is going to finish off the shoot of her first literary film. The following is the dialogue between the director and her friend.	: Will my film be successful at the box office? 我的电影会收获高票房吗? A: It is hard for audiences to really enjoy a literary film. 观众们很难真正欣赏文艺片。
PIC		: Do audiences like literary films? 观众们会喜欢文艺片吗? A: It is hard for audiences to really enjoy a literary film. 观众们很难真正欣赏文艺片。
GI	After completing his performance, the supporting actor is removing makeup in the backstage of the theater. The following is the dialogue between the actor and the director.	: Did everyone like our performance? 每个人都喜欢我的表演吗? A: Some of the audiences enjoyed your performance. 有的观众欣赏你的表演。
GIC		: Did everyone like our performance? 每个人都喜欢我的表演吗? A: Not all of the audiences enjoyed your performance. 不是所有观众欣赏你的表演。

processing share the same or similar neural representation; thus we could hardly distinguish the fMRI multivariate patterns of PI and GI processing. Semantic Minimalism predicts that PI and GI processing share similar language processing systems, but distinguish in inferential processing. Accordingly, we could identify a neural pattern of language processing that responds to both PI and GI generation, and an inference-related pattern that specifically responds to PI processing.

2. fMRI experiment

1 *Methods*

1 1 *Participants*

Twenty-nine university students were recruited for the fMRI experiment. One participant was excluded from data analysis on the basis of binary judgment accuracy (three SDs lower than group average), leaving 28 participants for data analysis (14 females; mean age 21.5, SD = 1.9). All participants were right-handed Chinese native speakers with normal or corrected-to normal vision. None of them suffered from neurological, psychiatric, or hearing disorders. This study was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences at Peking University. Written informed consents were obtained from all the participants.

1 *Design and Materials*

We used single-turn dialogue scenarios as stimulus materials. Each dialogue scenario was comprised of three parts - a cover story, a yes/no question, and an indirect or direct reply to the preceding question (Table 1, see *Supplementary Materials* for pretests). In the critical conditions (i.e., PI and GI), the reply was indirectly related to the question. For the control of PI condition, namely PIC, we used the same sentence as a direct reply to the preceding question. For the control of GI condition, namely GIC, we replaced the weak scalar term (e.g., *some of*) in the reply utterance of GI condition with its implicated meaning (e.g., *not all*), and thus the modified utterance served as a direct reply to the question. Various pairs of scalar items were included in GI pairs to minimize the repetition of certain lexical items, such as, “*some of oude ou ie* in Chinese vs. *all Suoyou uan u*”, “*sometimes* (*oushi oudeShihou oushihou* vs. *always ongshi ong*”, “*sometimes* vs. *often ingchang Changchang Shichang Chang*”, “*occasionally uer uyou* vs. *often*”, “*many times* vs. *all the time everyday*”, “*may eneng e u* vs. *must iding ending*”, “*want try iang asuan uli Changshi to do something* vs. *succeed in doing something*”, “*strive heng u to do something* vs. *promise Bao heng to do something*”, and “*a little adv* vs. *very adv*”. For each scenario, the question was strongly expected to receive a “*yes*” or a “*no*” answer and the reply gave a definite answer. Within each pair of scenarios, both direct and indirect replies were equivalent in giving a definite answer (“*yes*” or “*no*”) to the preceding questions. For the PI pairs, half of the replies answered “*yes*” to the questions while the other half answered “*no*”. However, for the GI pairs, all replies would give negative answers to the questions, rendering interpreting the scalar implicature of a weak term (i.e., the stronger term is not true) necessary for understanding the speaker's meaning of the reply. For example, in Table 1, the utterance “*Some of the audiences enjoyed your performance*” triggered a “no” answer to the question “*Did everyone like our performance*”. In this case, to understand the reply, listeners need to know that the usage of *some of*” warrants a GI “*some but not all*”. But, in the case that the same utterance gives a “*yes*” answer to the question “*Did anyone like our performance*”, it is unnecessary for listeners to notice that “*some of*” has the meaning of “*not all*”.

Apart from the scenarios in the four conditions, we created filler scenarios, which were similar to the critical scenarios in form and content. For each filler scenario, the question included a stronger term. Among these filler scenarios, 20 replies with strong terms were direct answers to the preceding questions, while the other 20 replies with weak terms were indirect. We added these fillers to balance the yes/no judgment of the scenarios, and to balance the yes/no response to replies with strong/weak terms (“*all*”, “*some*”, “*always*”, “*sometimes*” etc.), which made the materials more diversified and prevented the participants from formulating a certain response strategy.

To simulate natural conversations in daily life, all parts of dialogue scenarios were presented aurally. Fourteen Chinese native speakers were recruited to record specified parts of materials. One female and one male speaker were responsible for recording the cover stories, while six other female and six other male speakers recorded the single turn dialogues. For a particular scenario, the dialogue always occurred between a female and a male speaker. Each auditory stimulus was recorded in a sound-proof booth with a microphone (RODE NT1-A), digitized at 11.0 kHz sampling rate in a 16-bit format, and equated for the maximum sound intensity.

1 *Procedures*

For fMRI scanning, participants first performed a listening comprehension task. This task was separated into two sessions, each lasting about twenty minutes. All scenarios were divided into four experimental lists based on a Latin-square design, with each list further separated into two sessions. Each list consisted of 120 scenarios, including 20 scenarios for each experimental condition (i.e., PI, PIC, GI, and GIC) and 40 fillers. Scenarios in each list were sorted pseudorandomly, such that 1) no more than three scenarios in a certain experimental condition showed up consecutively; and 2) no more than four scenarios requiring an identical response showed up consecutively. In each trial, participants experienced the following events. First, a fixation cross was presented in the middle of the screen and remained for a jittered duration ranging from 1.5 to 5.5 s, before a blank screen lasting 0.1 s. Next, participants clearly heard the cover story, the question and the reply sequentially; at the meantime, only a fixation point was shown on the screen. We set up a fixed interval of 1 s after the presentation of the cover story, as well as a jittered interval ranging from 0.5 to 1.5 s between the presentation of the question and the reply. Finally, two option characters (“yes” on the left and “no” on the right) were presented and remained on the screen for 3 s immediately after the presentation of the reply utterance. Participants had to make a forced binary judgment as accurately and quickly as possible as to whether the latter speaker really intended to answer “yes” or “no” to the question. The judgment was indicated by a button press with the index or middle finger of the participants’ right hand. Reaction time (RT) was measured as the latency of his/her response to the presentation of “yes” and “no” choices.

After the listening comprehension task, participants also completed a ToM task in the scanner. Stimulus materials of this task were obtained from the Saxelab website (http://saxelab.mit.edu/localizers; credit David Dodell-Feder, Nicholas Dufour, and Rebecca Saxe), containing 10 false belief” and 10 control stories. We first translated these stories and its corresponding statements into Chinese. Then an English-Chinese bilingual, with English as his native language, translated the Chinese version back to English. This English translation and the original version were almost identical, indicating that the Chinese version was consistent with what the English version intended to convey. For each trial, a story was visually shown for 12 s, followed by a statement about the preceding story for 4 s. Each participant made a binary judgment as to whether the statement was True or False according to the story. A fixation interval of 12 s was presented between the trials.

Prior to fMRI scanning, all participants received written instructions concerning how to complete the tasks and performed a short practice for each task. After scanning, each participant completed a Chinese version of Autism Spectrum Quotient (AQ) questionnaire which is intended to measure individuals’ social skills (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). The subscale scores of this questionnaire reflect the degree of autistic-like social and communication difficulties; that is, the higher the score, the poorer the social or communication skills.

1 *Data acquisition and preprocessing*

Functional images were gathered on a research-dedicated 3-Tesla MRI scanner (GE MR750, General Electric, Fairfield, Connecticut), with a T2*-weighted echo-planar imaging sequence. Each volume contained 35 transversal slices, with repetition time/echo time/flip angle = 2000 ms/30 ms/90°, slice thickness/inter-slice gap = 4 mm/0.75 mm, field of view = 192 × 192 mm^2, resolution within slice = 64 × 64, and voxel size = 3.0 × 3.0 × 4.0 mm^3. Slices of each volume were acquired in an interleaved order. Head movements were minimized using pillows and cushions within the head coil.

The fMRI data preprocessing was conducted using SPM8 (Wellcome Centre of Human Neuroimaging, London; https://www.fil.ion.ucl.ac.uk/spm/). The first five volumes in each session were excluded from data analysis to allow the MR stabilization. Images were time sliced and realigned to the sixth volume to correct for head-motion artifacts. We used a high-pass temporal filter (cutoff period = 128 s) to remove low-frequency drifts in fMRI time series. We spatially normalized all functional images into the standard Montreal neurological institute (MNI) space by matching gray matter, white matter, and cerebrospinal fluid (Ashburner & Friston, 2005) and resampled to 3 × 3 × 3 mm^3 voxel. On this basis, the normalized data was smoothed using a 6-mm full-width half-maximum Gaussian kernel. No participants’ head movements exceeded 3 mm.

1 *Univariate analysis*

Whole-brain analyses were conducted using the generalized linear model (GLM) of SPM8 firstly at the participant level and secondly at the group level. For each session, all regressors were constructed as a boxcar function convolved with the canonical hemodynamic response function (HRF).

For the listening comprehension task, we defined nine/ten regressors in the GLM at the participant-level to model the following events: the auditory presentation of the cover story, the question and the reply, and the participants’ response. More specifically, the reply presentation was separately modeled by six/seven regressors, corresponding to four critical conditions (i.e., PI, PIC, GI, and GIC) and two types of fillers, as well as the misunderstood replies if the participant response was incorrect. The presentation of the cover story and the question, and the participants’ response were modeled by three regressors of no interest, respectively. Six rigid body parameters calculated from the realignment procedure were additionally included to correct for head-motion artifacts. The onset and duration of each regressor were defined as the actual onset and duration of each auditory stimulus. The simple main effect was examined in each experimental condition to identify brain regions significantly activated for each condition. For the group level analysis, a flexible factorial repeated-measures ANOVA was conducted on the participant-level contrast images of each experimental condition. At the group level, we used a cortical mask to exclude the cerebellum and conducted further analyses within this mask. We defined two contrasts for the two types of conversational implicatures, respectively, comparing the PI and GI conditions to their corresponding controls.

For the ToM task, the participant-level models were created by using a GLM with the false belief and control conditions as regressors of interest. The duration of each regressor contained the duration of the story reading (12 s) and the True/False judgment (4 s). At the group level, the two contrast maps corresponding to the two conditions from each participant were fed into a flexible factorial design. We defined one contrast comparing the false belief condition to the control.

Conjunction Analysis. To explore regions that were activated in interpreting both types of conversational implicatures, we further performed an SPM conjunction null’ analysis (Nichols, Brett, Andersson, Wager, & Poline, 2005) with $(PI > PIC) \cap (GI > GIC)$ (Friston, Holmes, Price, Büchel, & Worsley, 1999).

Parametric Analysis. To further reveal the functions of dmPFC during the comprehension of PI and GI, we conducted group-level parametric analyses using small volume correction within a dmPFC region-of-interest (ROI) to explore whether the dmPFC activation in PI/GI processing correlated with individual differences in social skills. The dmPFC ROI was defined by the co-activation of the contrasts PI > PIC and GI > GIC in the conjunction analysis at a relatively liberal threshold of voxel-level $p < 0.01$ uncorrected (1038 voxels in total). At the group-level, we used the measure of social skills (a subscale of AQ questionnaire) as a between-participant covariate and activations in the contrasts PI > PIC and GI > GIC recorded from the participant-level analyses as dependent variables, constructing two regression models, respectively. We next defined a sphere of 6-mm radius centered on the group peak coordinates identified by the parametric analysis (MNI coordinates: 9, 32, 49; see *Results*), and extracted the parameter estimates from this sphere in the contrast map PI > PIC and GI > GIC, respectively. Pearson correlation coefficients were computed between the scores of social skills and the dmPFC activation in the contrasts PI > PIC and GI > GIC, respectively. We then performed a statistical comparison of correlation to formally test whether the correlation coefficients were significantly different through Fisher’s Z-transform method and Zou’s confidence interval (CI) method (Zou, 2007). Both methods were performed using the cocor 1.1–3 R package (http://comparingcorrelations.org/; Diedenhofen & Musch, 2015).

Psychophysiological interaction (PPI) analysis. Given that dmPFC was found to be involved in generating both PI and GI (see *Results*), our interest lied in whether the functional interplay between dmPFC and other regions was modulated by the type of conversational implicature. For this purpose, we conducted a PPI analysis (Friston et al., 1997) with dmPFC revealed in the abovementioned conjunction analysis as the seed region, and calculated a PPI map corresponding to the contrast between PI and GI. The regression model contained three regressors and six head motion parameters. The first regressor, called physiological regressor, was the fMRI signals from a 6 mm-radium sphere centered on the group peak coordinates in the co-activated dmPFC (MNI coordinates: −9, 38, 43; see Table S1 in *Supplementary Materials*); the second, called psychological regressor, was the design vector (PI vs. GI); the third was calculated as the interaction between the physiological and psychological regressor.

All results were thresholded at $p < 0.001$ uncorrected at voxel-level and $p < 0.05$ family-wise error (FWE) corrected for multiple comparisons at cluster-level (whole-brain or within the dmPFC ROI using small-volume-correction; Chen, Jimura, White, Maddox, & Poldrack, 2015).

1 Multivariate pattern analysis

To identify the distributed neural representations of PI and GI pro-

To identify brain regions activated by ToM processing, we examined the false belief > control contrast at the whole-brain level. This contrast evoked clusters of activation in bilateral TPJ extending inferiorly to anterior temporal gyrus, mPFC, precuneus extending to post cingulum cortex, bilateral IFG and MFG. These results are highly consistent with the ToM network identified in previous studies (Dodell-Feder et al., 2011; Lee & McCarthy, 2016). As shown in Fig. 1E, PI-specific activations (in blue) were almost completely embedded in ToM processing network identified in this study (in red).

hole rain ultivariate attern analysis

To test the hypothesis that PI and GI processing have shared neural representations, we first trained and tested multivariate patterns at the whole-brain level. Multivariate fMRI pattern classifier trained to dissociate PI vs. PIC could discriminate PI from its control with 96 accuracy (95 confident interval (CI): 90–100 , < 0.001). When this classifier was applied to discriminate GI and its control, an accuracy approaching 100 (95 CI: 100–100 , < 0.001) was obtained. Similarly, the classifier trained to dissociate GI vs. GIC could discriminate GI condition from its control with 96 accuracy (95 CI: 90–100 , < 0.001), and could be generalized to discriminate PI vs. PIC with an accuracy of 96 (95 CI: 90–100 , < 0.001). These findings provided evidence for the existence of functionally shared neural representations for PI and GI. In addition, we found that the classifier trained to dissociate PI vs. GI could discriminate PI condition from GI condition with 96 accuracy (95 CI: 91–100 , < 0.001). Although such between-item comparison is informal, this finding may offer the possibility of distinction between these two processes.

Fig. 2A displays the thresholded whole-brain weight maps of the classifiers that discriminate PI (vs. PIC) and GI (vs. GIC), respectively (bootstrap tests with 10,000 iterations, a threshold of < 0.001 uncorrected for illustration purpose only). PI vs. PIC was predicted by activations in bilateral IFG, left anterior temporal lobe, right anterior MTG, bilateral TPJ and mPFC, while GI vs. GIC was predicted by increased activity in bilateral IFG, left posterior MTG and mPFC.

As a quantitative method, the neural similarity analyses with Neurosynth found that the PI classifier was positively correlated with prototypical brain patterns associated with language or ToM processing (with the terms *language*, *semantic*, *theory ind*, *intention*), while the GI classifier was only correlated with prototypical brain patterns associated with language processing (*language*, *semantic*; Fig. 2B).

We further examined the extent to which PI and GI engage language and ToM processing, by showing how activation patterns of these two processes classify neural representations of PI and GI. "Language" and ToM" prototypical brain patterns (Fig. 2D), defined by the meta-analytic database (term "language" and "theory mind" respectively), were used to discriminate PI and GI from their own controls (see Fig. 2C). The "Language" pattern performed significantly above chance in discriminating both PI vs. PIC (82 , 95 CI: 68–93 , < 0.001) and GI vs. GIC (79 , 95 CI: 64–91 , = 0.004), but performed at chance level in discriminating PI vs. GI (64 , 95 CI: 50–78 , = 0.18), suggesting that PI and GI comprehension engaged essentially the same neural pattern of language processing. In contrast, the "ToM" pattern could discriminate both PI vs. PIC (93 , 95 CI: 84–100 , < 0.001) and PI vs. GI (89 , 95 pe.

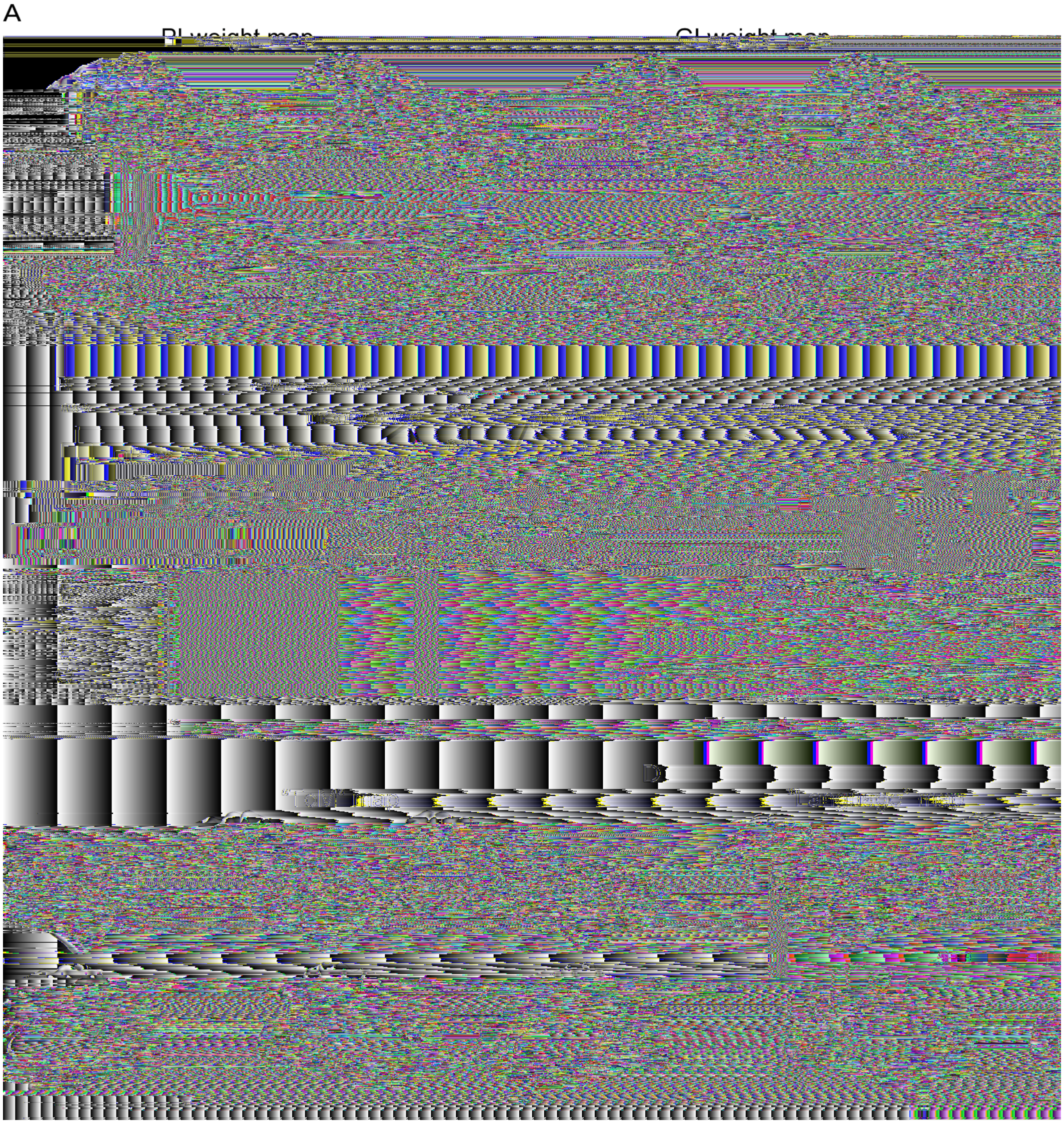

Fig. 2. Results of the whole-brain MVPA. (A) The whole-brain weight maps show voxels whose activity reliably classify PI vs. PIC conditions (i.e., PI weight map) or GI vs. GIC condition (i.e., GI weight map). Positive (warm color) and negative (cool color) weights indicate that more PI/GI processing was predicted by increased and reduced activity, respectively. (B) shows the results of neural similarity analysis using Neurosynth Image Decoder. (C) shows the accuracy of the "Language" map (left three bars) and the ToM map (right three bars) classifying PI vs. PIC, GI vs. GIC, and PI vs. PIC. Error bars represent SEs. ** < 0.01, *** < 0.001, n.s. not significant. (D) shows the prototypic language and ToM maps derived from Neurosynth database.

activated dmPFC, 58.6 voxels were also significantly activated by ToM task (see Fig. 3C). Given these seemingly contradictory findings, we further investigated whether dmPFC played an identical role in PI and GI processing.

We first hypothesized that if a "ToM" neural classifier within dmPFC could discriminate PI vs. PIC, but not GI vs. GIC, then it is reasonable for us to believe that PI and GI employed distinct neural representations in dmPFC. To test this hypothesis, we trained a "ToM" multivariate pattern within *a priori* dmPFC ROI to discriminate the false belief condition and its control in the ToM task. This dmPFC ROI was obtained from the univariate conjunction analysis of the contrast PI > PIC and GI > GIC. The cross-validation test showed that this "ToM" classifier could discriminate the false belief condition from its control with 100 accuracy (95 CI: 100–100 , < 0.001). When applied to discriminate the four experimental conditions (Fig. 3A), this "ToM" classifier performed significantly above chance in discriminating both PI vs. PIC (89 , 95 CI: 79–97 , < 0.001) and PI vs. GI (86 , 95 CI: 73–96 , < 0.001). However, this classifier performed at chance level in discriminating GI vs. GIC (61 , 95 CI: 45–76 , $= 0.34$), consistent with the whole-brain MVPA classification. These findings provided support to the hypothesis that interpreting PI and GI has distinct neural representations within dmPFC. Specifically, the representation of PI, but not GI, may involve a ToM-related inferential component.

Secondly, we carried out univariate parametric analyses for activation in dmPFC ROI. We added the participants' social skills (as measured by A questionnaire; see *Supplementary Materials* for details) as group-level covariates for the PI > PIC and GI > GIC contrasts in two separate models. As shown in Fig. 3B, the magnitude of activation in dmPFC (peak coordinates: 9, 32, 49 ; cluster size = 12; $_{\mathrm{FWE}} = 0.041$, small-volume corrected) negatively correlated with the social skills scores during PI processing ($r = -0.60$, $= 0.001$), but not during GI processing ($r = 0.10$, $= 0.61$). A direct comparison confirmed that the two correlation coefficients differed significantly, $= -3.22$, $= 0.001$, with 95 CI being $-1.05, -0.29$. These findings indicated that individuals' social skills modulated dmPFC activation during PI processing, but had no effect on GI processing.

Finally, we conducted a PPI analysis by using the *a priori* dmPFC (peak coordinates: $-9, 38, 43$) as seed region. We found that dmPFC

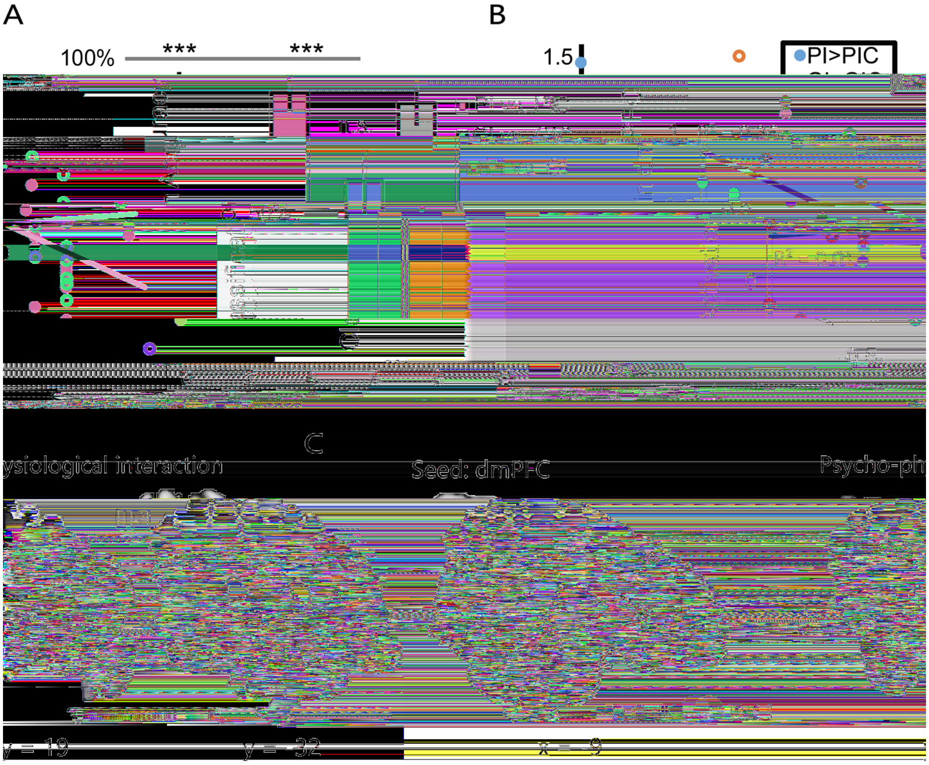

Fig. 3. Results of the ROI analyses. (A) shows the cross-validated accuracy of the ToM map within the dmPFC ROI classifying PI vs. PIC, GI vs. GIC, and PI vs. PIC. (B) shows the results of ROI-based parametric analyses. The parameter estimates corresponding to four experimental conditions were extracted from the dmPFC (based on the parametric analysis). (C) shows the overlapping area (shown in yellow) within dmPFC between the ToM network (shown in red) and the co-activation of PI and GI (shown in green), and the results of PPI analysis.

showed significantly stronger functional interplay with several brain regions, including precentral gyrus, left inferior parietal lobule (IPL), right IFG pars opercularis and pars orbitalis (extending to right anterior insula), and pre-SMA during PI processing, relatively to GI processing (Fig. 3C and Table S2).

Discussion

In the current experiment, we investigated the neural representations of PI and GI comprehension. Results from both univariate and multivariate fMRI data analyses consistently demonstrate that comprehension of PI and GI share a language processing component but differ in that PI but not GI comprehension further relies on ToM-related inferential processes.

In the effect of PI comprehension (comprising indirect replies with PI vs. direct replies), there were activations in bilateral IFG, MTG, mPFC (extending to pre-SMA), TPJ, precuneus, and MFG, which essentially replicated previous findings on PI comprehension (Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2014; Feng et al., 2017; Shetreet, Chierchia, & Gaab, 2014; Shibata, Abe, Itoh, Shimada, & Umeda, 2011; van Ackeren, Smaragdi, & Rueschemeyer, 2016). IFG and MTG, as core regions of language network, have been implicated in recovering literal content (Ferstl & von Cramon, 2001; Xu, Kemeny, Park, Frattali, & Braun, 2005), while mPFC, TPJ and precuneus constitute a “ToM network” that is typical for tasks involving higher-order, ToM-related inferential processes (Koster-Hale & Saxe, 2013; Van Overwalle & Baetens, 2009). Moreover, instead of creating a pragmatically mismatch context (e.g., the sentence verification or picture-sentence verification paradigm), the current listening comprehension task revealed the neural substrates of GI processing beyond the scope of previous studies. By contrasting indirect replies with GI against direct replies, we showed that interpreting GI also more reliably activates bilateral IFG, left MTG, and mPFC (extending to pre-SMA) than understanding direct replies. Thus, PI and GI processing may engage common neural substrates from the perspective of overlapping fMRI activations.

The brain regions in the “ToM network” can support multiple cognitive functions other than the classic ToM-related processes (i.e., inferring the mental states of other people, such as false belief reasoning). For example, some studies proposed that the dmPFC activation observed in discourse comprehension may reflect some general functions shared by ToM and discourse comprehension (Mason & Just, 2011); TPJ region supports cognitive control or attention (Carter & Huettel, 2013; Lee & McCarthy, 2016); and ToM network supports working memory of social information (Meyer, Spunt, Berkman, Taylor, & Lieberman, 2012). Therefore, the activation of “ToM” network does not necessarily imply the involvement of the typical ToM processing. To avoid the informal reverse inference of a cognitive process from activation in a certain brain region or system on the basis of a biased literature review (Aguirre, 2003; Poldrack, 2006), we performed more fine-grained analyses, combining the MVPA approach with independent neural representations drawn from large-scale meta-analysis, to clarify whether the overlapping activations arise from the same or different neural representations (Peelen & Downing, 2007) and to identify what cognitive processes were engaged among all likelihood (Poldrack, 2011). On the one hand, our results of whole-brain multivariate pattern decoding provide considerable evidence for the argument that the PI and GI comprehension engage the same neural representation of language processing, which may well be recruited by constructing and maintaining a coherent representation of utterances in the discourse (Menenti, Petersson, Scheeringa, & Hagoort, 2009; Rapp, Mutschler, & Erb, 2012). On the other hand, our results demonstrate that the neural representation engaged in performing ToM-like inferential processes is merely observable during PI comprehension, not during GI comprehension. Combined with the results of univariate analysis, these findings suggest that the comprehender's ToM-related network is selectively recruited to infer speaker's aims and intentions by recovering the meaning bound up with specific context.

The dmPFC was involved in both PI and GI comprehension. In the ROI analyses, however, we found that there were differences in the common activation of dmPFC during interpreting PI and GI. First, ROI-based MVPA showed that PI processing activated a ToM-related fMRI pattern within dmPFC, but GI processing did not. It means that interpreting GI engages only weakly ToM-like inferential processing at best. Second, activation in dmPFC strongly correlated with individuals' social skills during PI processing, but not during GI processing. Third, dmPFC showed significantly stronger functional connectivity with SMA, premotor cortex, right IFG and left IPL during PI processing, relatively to GI processing. The latter pattern of frontal and parietal activity is associated with domain-general cognitive/executive control (Duncan, 2010; Ye & Zhou, 2009a, 2009b). Given that PI comprehension is generally more difficult than GI comprehension, it is reasonable to predict that PI may require additional cognitive processing to monitor and resolve the conflicts between sentential representations in discourse. Thus, the increased functional connectivity may reflect how the cognitive control system was involved in pragmatic inference during PI comprehension. Thus, a related idea is that this region is engaged in strategic inferential processing to establish the relation between utterances in discourse (Ferstl, Neumann, Bogler, & von Cramon, 2008; Ferstl & von Cramon, 2002; uperberg, Lakshmanan, Caplan, & Holcomb, 2006). The activation in dmPFC during GI comprehension could reflect a more general and encapsulated inferential process (Ferstl & von Cramon, 2001, 2002; Mason & Just, 2011), such as the one underlying the logical reasoning of specific terms (e.g., *some* = *not all*).

Nevertheless, our findings are fully congruent with the idea that dmPFC contains multiple, different neural populations that encode distinct mental states. The dmPFC is involved in a variety of high-order cognitive functions. Although dmPFC is one of the central regions in ToM processing (Van Overwalle, 2009), dmPFC is also recruited by ToM-unrelated inductive reasoning (Ferstl & von Cramon, 2002; Siebörger et al., 2007). We suggest that the dmPFC activity in the current study is probably more related to the activation of social information and situational context during generating PI, compared with GI. Specifically, in understanding indirect replies with PI, dmPFC is co-activated with other ToM-related regions, including TPJ and precuneus, and the dmPFC activity supports ToM-like inferential processing in order to infer the current mental state of the speaker in a particular context.

3. Brain stimulation (HD-tDCS) experiments

Given that ToM-related inferential processes may play a critical role in generating PI, but not GI, we performed two independent brain stimulation experiments, using HD-tDCS to test the causal role of a ToM-related brain region (right TPJ) in processing the two types of conversational implicature. In our fMRI experiment, right TPJ was specifically activated during interpreting PI, but not during interpreting GI. This region is generally considered as a critical brain region of the ToM network (rall et al., 2015; Lee & McCarthy, 2016; Mar 2011; Saxe & Powell, 2006), responsible for extracting and integrating social information from the bulk of information (Carter & Huettel, 2013; Schaafsma, Pfaff, Spunt, & Adolphs, 2015). Moreover, previous studies have revealed that the anodal brain stimulation to right TPJ could improve ToM-related processing in social interaction (Santiesteban, Banissy, Catmur, & Bird, 2012; Sowden, Wright, Banissy, Catmur, & Bird, 2015), and the cathodal stimulation to right TPJ could reduce such function (Leloup, Miletich, Andriet, Vandermeeren, & Samson, 2016; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Therefore, we selected right TPJ region to deliver tDCS.

1 *Methods*

1 1 *Participants*

Sixty-seven (37 females; mean age = 21.3, SD = 2.4, range 18–28 years) and eighty-eight (56 females; mean age = 20.7, SD = 2.0) university students, who did not take part in either the pretests or the fMRI experiment, participated in one anodal and one cathodal tDCS experiments, respectively. For the anodal experiment, a sub-group of the participants (n = 34, 22 females) received anodal tDCS over right TPJ, whereas the other sub-group (n = 33, 15 females) received sham stimulation over the same area. For the cathodal experiment, 46 participants (26 females) received cathodal tDCS over right TPJ, whereas 42 participants (30 females) received sham stimulation over the same area. Five additional participants were excluded from the anodal experiment and seven from the cathodal experiment, due to incomplete data collection or their poor task performance (three SDs longer than average in reaction times or lower in task accuracy).

All the participants were right-handed Chinese native speakers with normal or corrected-to normal vision. None of them suffered from neurological, psychiatric, or hearing disorders. This study was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences at Peking University, and written informed consents were obtained from all the participants.

1 *Procedure*

Two independent tDCS experiments were completed. Both experiments were double-blind; that is, neither the participants nor the experimenter who gave instructions to the participants was aware of the assigned type of brain stimulation. HD-tDCS was delivered using a multichannel stimulation adapter (SoterixMedical, 4 × 1-C3) connected to the constant current stimulator (SoterixMedical, Model 1300-A). Five Ag-AgCl sintered ring electrodes were embedded in an EEG cap and connected to the scalp with electrode gel. To deliver stimulation over right TPJ, one central electrode was placed on CP6, and four return electrodes surrounding it were placed sham) × 2 (inference type: belief vs. control) repeated measures ANOVAs on participants' task accuracy. For the anodal experiment (Fig. 4A left panel), a marginally significant interaction between the two factors was revealed, $F(1,65) = 3.48$, $= 0.067$, $\eta^2 = 0.05$. Simple effect analysis revealed that for the sham group, the accuracy rate was lower in false belief condition (70.6 ± 2.7) than in the control condition (81.2 ± 2.2 ; < 0.001, $\eta^2 = 0.21$); for the anodal group, there was no significant difference in accuracy between false belief condition (80.0 ± 2.6) and control condition (83.8 ± 2.1 ; $= 0.14$, $\eta^2 = 0.03$). For the cathodal experiment (Fig. 4A right panel), the analysis also showed a marginally significant interaction, $F(1,86) = 3.81$, $= 0.054$, $\eta^2 = 0.04$. Simple effect analysis revealed that for the sham group, the accuracy rate was lower in false belief condition (71.7 ± 2.7) than in control condition (81.7 ± 1.9 ; $= 0.001$, $\eta^2 = 0.12$). This effect was larger for the cathodal group (false belief, 65.4 ± 2.6 vs. control, 83.3 ± 1.8 ; < 0.001, $\eta^2 = 0.33$). These findings confirmed that enhancing or disrupting right TPJ functions through tDCS facilitates or hinders ToM-related inferential processes.

We then analyzed behavioral data in the listening comprehension task. For each experimental condition, participants correctly responded to more than 95 of all trials. For the anodal experiment (Fig. 4B left panel), a 2 (tDCS type: anodal vs. sham) × 2 (scenario pair: PI pair vs. GI pair) × 2 (implicature: critical condition vs. control condition) repeated measures ANOVA on participants' RTs revealed a significant three-way interaction between tDCS type, scenario pair and implicature, $F(1, 65) = 4.30$, $= 0.042$, $\eta^2 = 0.06$. Separate ANOVAs on the tDCS effect were carried out for the PI and GI scenario pairs, respectively. For the PI pair, there was a significant interaction between tDCS type and implicature, $F(1, 65) = 4.12$, $= 0.046$, $\eta^2 = 0.06$. Tests for simple effects showed that for the sham group, the RTs were longer in the PI condition (765 ± 49 ms) than in the PIC condition (583 ± 40 ms; < 0.001, $\eta^2 = 0.41$), while this effect was much larger for the anodal group (PI, 827 ± 48 ms vs. PIC, 566 ± 40 ms; < 0.001, $\eta^2 = 0.59$), suggesting that the anodal stimulation over right TPJ causally slowed down responses to the indirect replies with PI. For the GI pair, there was neither a main effect of tDCS type, nor an interaction between tDCS type and implicature (Fs < 1), indicating that the anodal brain stimulation over right TPJ could not affect GI comprehension.

The same pattern of results was obtained in the cathodal experiment (Fig. 4B right panel). The ANOVA on RT showed a significant three-way interaction, $F(1, 86) = 4.28$, $= 0.042$, $\eta^2 = 0.05$. Separate ANOVAs on the tDCS effect were carried out for the PI and GI scenario pairs. For the PI pair, there was a significant interaction between tDCS type and implicature, $F(1, 86) = 4.97$, $= 0.028$, $\eta^2 = 0.06$. Tests for simple effects showed that for the sham group, the RT was longer in the PI condition (690 ± 34 ms) than in the PIC condition (514 ± 27 ms; < 0.001, $\eta^2 = 0.33$), and this effect was much larger for the cathodal group (PI, 793 ± 33 ms vs. PIC, 534 ± 26 ms; < 0.001, $\eta^2 = 0.54$), indicating that the cathodal stimulation over right TPJ causally showed down responses to the indirect replies with PI. For the GI pair, there was neither a main effect of tDCS type, nor an interaction between tDCS type and implicature (Fs < 1.5), indicating that the cathodal brain stimulation over right TPJ could not affect GI comprehension.

To further explore the relationship between brain stimulation over right TPJ and behavioral performance on PI, we examined the indirect pathway from tDCS stimulation via ToM ability (the accuracy difference between false belief and control conditions) to PI comprehension. Results showed that the association between brain stimulation over right TPJ and PI comprehension could be mediated by ToM ability, for both anodal (the indirect effect estimate ± SE = 22.97 ± 15.77, 95 CI = 0.59, 65.25) and cathodal (16.84 ± 13.19, 95 CI = 0.41, 57.03) experiments (Fig. 4C). Similar analyses could not be conducted for GI comprehension, as the brain stimulation over right TPJ exhibited no effect on it.

iscussion

Previous studies have consistently showed that the brain stimulation over right TPJ could causally affect ToM processing (Leloup et al., 2016; Santiesteban et al., 2012; Sowden et al., 2015; Young et al., 2010). Here, to further clarify the functions of ToM network in PI and GI comprehension by distinguishing its causal roles, we selected right TPJ region to

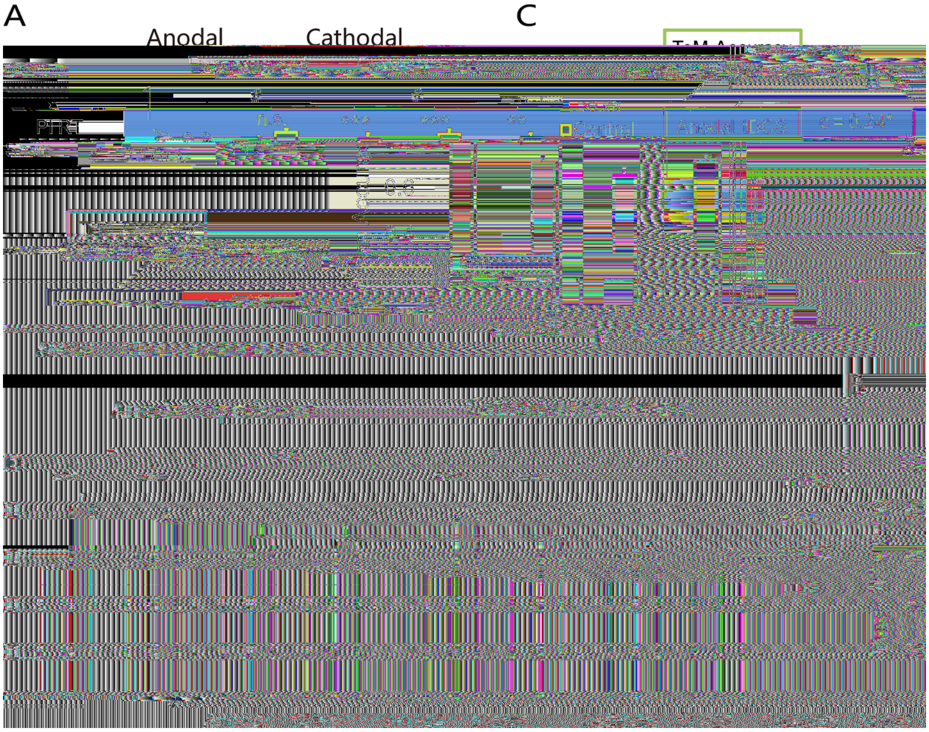

Fig. 4. tDCS results for the ToM task (A) and the listening comprehension task (B). (C) The indirect pathway from the brain stimulation over right TPJ, via ToM ability, to PI comprehension. Error bars represent between-subject SEs. < 0.07, * < 0.05, ** < 0.01, *** < 0.001, n.s. not significant.

deliver tDCS.

First of all, results of the ToM task verified the validity of tDCS manipulation by showing that enhancing or disrupting right TPJ functions through tDCS did facilitate or hinder ToM-related inferential processes. More importantly, both anodal and cathodal stimulation causally engendered slower responses to the indirect replies with PI, and the individual's ToM ability mediates the influence of tDCS on PI comprehension. But, neither anodal nor cathodal stimulation over right TPJ impacted responses to the indirect replies with GI. According to previous studies, ToM ability is tightly related to pragmatic language processing (Cummings, 2017), such as irony comprehension (Martin & McDonald, 2004; Monetta, Grindrod, & Pell, 2009), proverb comprehension (Brüne & Bodenstein, 2005), and the interpretation of indirect speech (Cuerva et al., 2001; Muller et al., 2010). In interpreting an utterance, a comprehender is always to infer and identify the speaker's intentions in a certain linguistic expression. When the speaker's meaning of an utterance relies highly on particular context (as in PI condition), the complexity of such inferential processing increases. Hence during PI processing, the comprehender's communicative-pragmatic performance would be sensitive to his/her ToM ability in discerning the speaker's current intentions. Right TPJ, as a core region of ToM network, is selectively necessary for individuals' PI processing.

Surprisingly, similar to cathodal stimulation, anodal stimulation over right TPJ disrupted PI processing, whereas it did improve the individuals' theory-of-mind ability. This finding is incongruent with our prediction, although it does not invalidate our conclusion that changing the neural activity in right TPJ would affect PI processing but not GI processing. To this finding, one possible explanation is that the individuals' increased ToM ability by anodal stimulation may go well beyond what might be needed □ x not in GI comprehension, whether at the whole-brain level or within the co-activated dmPFC region. Secondly, the tDCS experiments revealed that the brain stimulation over right TPJ could causally affect PI comprehension through its impacts upon the ToM ability, but it does not affect GI comprehension. These findings consistently indicated that the cognitive processes underlying PI and GI generation are distinct, supporting the intuitive distinction between PI and GI by Grice (1975). Thus, these findings are compatible with the accounts of either Default Theory or Semantic Minimalism. Overall, the evidence from this study suggests that compared to Default Theory and Relevance Theory, Semantic Minimalism provides more felicitous theoretical description of the cognitive processes underlying PI and GI generation and the relationship between these two types of implicatures.

Considering that we used the verbal false belief task to investigate the neural representation associated with ToM processing in the fMRI experiment and to measure the individuals' ToM ability in the tDCS experiments, one thing is noteworthy. In this ToM task, the false belief condition contains short discourses describing false beliefs, while the control condition contains discourses describing outdated photographs and maps (Dodell-Feder et al., 2011). Although this design roughly matched the domain-general inferences about outdated representations, the stimuli used in the target

de

Eckardt, R. (2007). Licensing 'or'. In U. Sauerland, & P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics* (pp. 34–70). Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language, 1 1*, 52–61.

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping, 9*(5), 581–593.

Ferstl, E. C., & von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research, 11*(3), 325–340.

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: Coherence or theory of mind? *NeuroImage, 1* (3), 1599.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage,* (3), 218–229.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., & Worsley, K. J. (1999). Multi-subject fMRI studies and conjunction analyses. *NeuroImage, 10*(4), 385–396.

Gavilán, J. M., & García-Albea, J. E. (2011). Theory of mind and language comprehension in schizophrenia: Poor mindreading affects figurative language comprehension beyond intelligence deficits. *Journal of Neurolinguistics,* (1), 54–69.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and Semantics III: Speech acts* (pp. 41–58). New York: Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard: Harvard University Press.

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology,* (416), 1–13.

Hagoort, P., & Levinson, S. C. (2014). Neuropragmatics. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 667–674). Cambridge, MA: MIT Press.

Horn, L. R. (2004). Implicature. In L. R. Horn, & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 2–28). Oxford: Blackwell.

Jang, G., Yoon, S. A., Lee, S. E., Park, H., Kim, J., & Ko, J. H. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage, 81*(6), 61–72.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 9*(5), 836–848.

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., … Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function, 0*(2), 587–604.

Kuperberg, G. R., Lakshmanan, B. M., Caplan, D. N., & Holcomb, P. J. (2006). Making sense of discourse: An fMRI study of causal inferencing across sentences. *NeuroImage,* (1), 343–361.

Lee, S. M., & McCarthy, G. (2016). Functional heterogeneity and convergence in the right temporoparietal junction. *Cerebral Cortex,* (3), 1108–1116.

Leloup, L., Miletich, D. D., Andriet, G., Vandermeeren, Y., & Samson, D. (2016). Cathodal transcranial direct current stimulation on the right temporo-parietal junction modulates the use of mitigating circumstances during moral judgments. *Frontiers in Human Neuroscience, 10*, 355.

Levinson, S. C. (2000). *Presumptive Meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lin, N., Yang, X., Li, J., Wang, S., Hua, H., Ma, Y., & Li, X. (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 18*(2), 273–283.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology,* (1), 103–134.

Martin, I., & McDonald, S. (2003). Weak coherence, no theory of mind, or executive dysfunction? Solving the puzzle of pragmatic language disorders. *Brain and Language, 8* (3), 451–466.

Martin, I., & McDonald, S. (2004). An exploration of causes of non-literal language problems in individuals with Asperger syndrome. *Journal of Autism and Developmental Disorders,* (3), 311–328.

Mason, R. A., & Just, M. A. (2011). Differentiable cortical networks for inferences concerning people's intentions versus physical causality. *Human Brain Mapping,* (2), 313–329.

Menenti, L., Petersson, K. M., Scheeringa, R., & Hagoort, P. (2009). When elephants fly: Differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *Journal of Cognitive Neuroscience, 1*(12), 2358–2368.

Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E., & Lieberman, M. D. (2012). Evidence for social working memory from a parametric functional MRI study. *Proceedings of the National Academy of Sciences, 109*(6), 1883–1888.

Monetta, L., Grindrod, C. M., & Pell, M. D. (2009). Irony comprehension and theory of mind deficits in patients with Parkinson's disease. *Cortex,* (8), 972–981.

Muller, F., Simion, A., Reviriego, E., Galera, C., Mazaux, J.-M., Barat, M., & Joseph, P.-A. (2010). Exploring theory of mind after severe traumatic brain injury. *Cortex,* (9), 1088–1099.

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage,* (3), 653–660.

Noveck, I., & Reboul, A. (2008). Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences, 1* (11), 425–431.

Peelen, M. V., & Downing, P. E. (2007). Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends in Cognitive Sciences, 11*(1), 4–5.

Piovan, C., Gava, L., & Campeol, M. (2016). Theory of Mind and social functioning in schizophrenia: Correlation with figurative language abnormalities, clinical symptoms and general intelligence. *Rivista di Psichiatria, 1*(1), 20–29.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron,* (5), 692–697.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 0*(3), 879–891.

Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *Journal of Neuroscience,* (13), 3829–3838.

Rapp, A. M., Mutschler, D. E., & Erb, M. (2012). Where in the brain is nonliteral language? a coordinate-based meta-analysis of functional magnetic resonance imaging studies. *NeuroImage,* (1), 600–610.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology,* (23), 2274–2277.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 1* (8), 692–699.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews,* , 9–34.

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When some is not every: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping,* (4), 1503–1514.

Shibata, M., Abe, J. I., Itoh, H., Shimada, K., & Umeda, S. (2011). Neural processing associated with comprehension of an indirect reply during a scenario reading task. *Neuropsychologia, 9*(13), 3542–3550.

Siebörger, F. T., Ferstl, E. C., & von Cramon, D. Y. (2007). Making sense of nonsense: An fmri study of task induced inference processes during discourse comprehension. *Brain Research, 11* (1), 77–91.

Sowden, S., Wright, G. R. T., Banissy, M. J., Catmur, C., & Bird, G. (2015). Transcranial current stimulation of the temporoparietal junction improves lie detection. *Current Biology,* (18), 2447–2451.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition* (Vol. 1). Harvard: Harvard University Press.

Talbert, B. (2017). Overthinking and other minds: the analysis paralysis. *Social Epistemology, 1*(6), 545–556.

Tettamanti, M., Vaghi, M. M., Bara, B. G., Cappa, S. F., Enrici, I., & Adenzato, M. (2017). Effective connectivity gateways to the Theory of Mind network in processing communicative intention. *NeuroImage, 1* , 169–176.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan, & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 200–239). Hillsdale, N.J.: Erlbaum.

van Ackeren, M. J., Smaragdi, A., & Rueschemeyer, S. A. (2016). Neuronal interactions between mentalising and action systems during indirect request processing. *Social Cognitive and Affective Neuroscience, 11*