Effective connectivity gateways to the Theory of Mind network in processing communicative intention

Marco Tettamanti[1], Matilde M. Vaghi[2,3], Bruno G. Bara[4,5,6], Stefano F. Cappa[7], Ivan Enrici[5,6,8*], Mauro Adenzato[4,5,6]

1. Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milano, Italy;
2. Faculty of Psychology, Vita-Salute San Raffaele University, Milano, Italy;
3. Department of Psychology, Behavioural and Clinical Neuroscience Institute, University of Cambridge, UK;
4. Department of Psychology, University of Turin, Italy
5. Center for Cognitive Science, University of Torino, Italy;
6. Neuroscience Institute of Turin, Italy,
7. Institute for Advanced Studies IUSS, Pavia, Italy,
8. Department of Philosophy and Educational Sciences, University of Turin, Italy.

*Corresponding author:

Ivan Enrici, Ph.D.

Department of Philosophy and Educational Sciences, University of Turin

via Gaudenzio Ferrari, 9 - 10124 Turin, Italy

Phone: +39.011.670.3177

Email: ivan.enrici@unito.it

Abstract

An Intention Processing Network (IPN), involving the medial prefrontal cortex, precuneus, bilateral posterior superior temporal sulcus, and temporoparietal junctions, plays a fundamental role in comprehending intentions underlying action goals. In a previous fMRI study, we showed that, depending on the linguistic or extralinguistic (gestural) modality used to convey the intention, the IPN is complemented by activation of additional brain areas, reflecting distinct modality-specific input gateways to the IPN. These areas involve, for the linguistic modality, the left inferior frontal gyrus (LIFG), and for the extralinguistic modality, the right inferior frontal gyrus (RIFG). Here, we tested the modality-specific gateway hypothesis, by using DCM to measure inter-regional functional integration dynamics between the IPN and LIFG/RIFG gateways. We found strong evidence of a well-defined effective connectivity architecture mediating the functional integration between the IPN and the inferior frontal cortices. The connectivity dynamics indicate a modality-specific propagation of stimulus information from LIFG to IPN for the linguistic modality, and from RIFG to IPN for the extralinguistic modality. Thus, we suggest a functional model in which the modality-specific gateways mediate the structural and semantic decoding of the stimuli, and allow for the modality-specific communicative information to be integrated in Theory of Mind inferences elaborated through the IPN.

Keywords: Communicative intention, Communicative modality, Dynamic Causal Modeling, Inferior frontal gyrus, Theory of Mind.

Introduction

Human communicative competence is based on the ability to process a specific class of mental states, namely, communicative intention (Bara, 2010). According to the cognitive pragmatics approach, communicative intention is defined as the intention to communicate a meaning to someone else, plus the intention that the former intention should be recognized by the addressee (Grice, 1975). The process involved in understanding this form of intention is independent of the communicative modality (linguistic or gestural) through which it is conveyed, and connects human communication with a more general type of social competence, such as Theory of Mind (ToM), i.e., the ability to explain and predict other people's communicative and non-communicative behavior by attributing independent mental states to them (Baron-Cohen, 1995; Premack & Woodruff, 1978).

In previous studies we proposed the Intention Processing Network (IPN) model, according to which a set of brain areas are differentially involved in comprehending different types of intentions, such as private or social intentions. Whereas a private intention involves the representation of a private goal, i.e. a goal involving only a single actor, a social intention involves the representation of a social goal, i.e. a goal that necessitates at least another person to achieve the goal. In three functional Magnetic Resonance Imaging (fMRI) studies (Ciaramidaro et al., 2007; Walter et al., 2004; Walter et al., 2009), we used a story completion task presented in a comic strip form to show the differential recruitment of the ToM network according to private versus social intentions. The brain areas associated to the IPN include the medial prefrontal cortex (MPFC), the precuneus (PREC), the bilateral posterior superior temporal sulcus (pSTS), and the temporoparietal junctions (TPJ). During the comprehension of a social (communicative) intention, all four areas of the IPN are recruited. In contrast, the comprehension of a private intention involved only the PREC and the right TPJ/pSTS. As a whole, the four IPN brain regions constitute a subset of the ToM system that is specifically recruited when people try to infer the intentions of others. This occurs even in the absence of detailed information on biological motion (Van Overwalle & Baetens, 2009). Thus, the IPN shows no complete anatomo-functional overlap, neither with the mirror system, nor with the brain regions of

the ToM system specifically implicated in inferring other's affective mental states such as emotions (Corradi-Dell'Acqua et al., 2014).

Previous work extensively clarified the specific role of individual brain areas constituting the IPN in communicative intention recognition and comprehension. For example, the anterior (in particular the MPFC) and posterior (in particular the right TPJ) cortices have a key role for verbal irony comprehension (Spotorno et al., 2012), for metaphors comprehension (Prat et al., 2012), and in indirect replies in spoken dialogue (Bašnáková et al., 2014), as shown by studies entailing the comprehension of pragmatic phenomena in which literal and intended meaning dissociate. Meta-analysis studies (Van Overwalle 2009; Van Overwalle & Baetens, 2009) suggested the implication of the PREC for elaboration of contextual information and identification of situational structure. In contrast, the role of the TPJ was generally associated with the identification of end state behaviors. Specifically, according to Van Overwalle (2009), the TPJ along with the PREC and MPFC takes part in the broader process of goal identification in a social context. Strong empirical evidence demonstrates MPFC engagement in social inferences, in particular in understanding social scripts that do not only concern a single actor, but that describe adequate social actions for all of several actors involved in a particular context (for reviews, see Van Overwalle 2009; 2011).

Converging evidence for the role of the IPN in communicative intention processing comes from lesion studies. Deficits in inferring speaker intentions were found in people with MPFC lesions (Lee et al., 2010). Impaired comprehension of non-literal language, such as sarcasm, metaphor, and indirect requests was found in people with brain diseases that affect the functioning of the medial frontal cortex, such as frontotemporal dementia (Shany-Ur et al., 2012), Tourette syndrome (Eddy et al., 2010), and progressive supranuclear palsy (Ghosh et al., 2012), even when controlling for the possible confounding effect of executive function deficits (see however Aboulafia-Brakha et al., 2011, for the complex relationship between executive functions and ToM in patients with acquired neurological pathology). Conversely, extensive damage to the perisylvian fronto-temporal language network resulting in aphasia and characterized by lexical-semantic impairments, does not cause

specific deficits in intention recognition (see Willems & Varley, 2010, for a review), nor does it compromise the ability to express intended communicative meanings *per se.* Indeed, using alternative communicative resources, such as drawing, facial expression, and gesture, these patients are able to convey meaningful messages (Siegal & Varley, 2006; Varley & Siegal, 2006). As shown by Willems et al. (2011), aphasic patients are able to process communicative intention (both comprehension and production) and to exhibit communication strategies comparable to those adopted by the healthy population, when using a novel non-verbal communication paradigm.

In a more recent study by our group (Enrici et al., 2011), we specifically asked whether the verbal versus the non-verbal communication modalities are processed by distinct neural networks, and whether these neural networks do overlap or are rather independent from the IPN network implicated in communicative intention processing. We used a story completion task, whose distinguishing feature was that the stories represented the social communicative intention in either a verbal (linguistic) or a gestural, (extralinguistic) modality. We showed that the IPN was recruited for the comprehension of communicative intention, independently of the linguistic or extralinguistic modality through which it was conveyed. Additional brain areas, outside those involved in intention processing, were specifically engaged according to the particular communicative modality. Specifically, the linguistic modality additionally recruited the peri-sylvian language network, including the pars opercularis of the left inferior frontal gyrus (LIFG). In contrast, the extralinguistic modality additionally recruited a sensorimotor network, including the pars opercularis of the right inferior frontal gyrus (RIFG). Based on these activation results, we hypothesized that the LIFG and RIFG reflect modality-specific input gateways, conveying stimulus and associated high-order information to the IPN.

The importance of the IFG as an interface node to the IPN is suggested by the presence of structural inter-connection pathways. In particular, the frontal aslant white matter tract links the IFG directly to the MPFC and is part of the core neural network underlying communicative intention processing (Catani & Bambini, 2014). In addition, the IFG is a crucial integration hub for

communication comprehension (Kemmerer, 2014), and is thus a likely candidate region to exchange high-order information with the IPN for the purpose of communicative intention decoding. In the context of modality-specific parsing of communicative signals, the LIFG and RIFG present a relative hemispheric specialization for, respectively, sentences and gestures (Straube et al., 2012).

While these observations altogether provide a plausible premise, the precise functional relationship between IPN and the inferior frontal gyri in the two hemispheres has not been investigated yet. In the present study, we tested the modality-specific gateway hypothesis, by focusing on inter-regional functional integration between the IPN and LIFG/RIFG. To this aim, we further analyzed the data collected in the Enrici et al. (2011) study, by measuring effective connectivity with Dynamic Causal Modeling (DCM). More specifically, we employed DCM network discovery (Friston and Penny, 2011; Friston et al., 2011), as an approach that enables one to test the connectivity between a priori specified brain regions, and to discover, over a large number of possible models, the one with the greatest evidence to have generated the observed fMRI data. Based on the body of knowledge reviewed above, we specified our models as including four brain regions of the IPN – i.e., MPFC, left TPJ (LTPJ), right TPJ (RTPJ) and PREC – together with LIFG and RIFG as modality-specific input gateways. We expected that the model with greatest evidence would be consistent with the modality-specific propagation of stimulus information from the LIFG to IPN for the linguistic modality, and from the RIFG to IPN for the extralinguistic modality.

Materials and Methods

A full description of fMRI data acquisition and preprocessing procedures can be found in Enrici et al. (2011). Details relevant for the present study are reported in what follows.

Participants

Twenty-four right-handed Italian native speakers (13 females, mean age 24.45 years, SD 5.71) with no history of neurological or psychiatric diseases participated in the imaging study. The Ethics

Committee of the San Raffaele Scientific Institute approved the study. All participants gave their written informed consent prior to scanning.

Stimuli and task

The experiment conformed to a 2 x 2 factorial design, with factors *Intention* (communicative intention versus non-intentional physical causality) and *Modality* (linguistic versus extralinguistic). The four resulting experimental conditions were: 1) Linguistic Communicative Intention (LCInt); 2) Extralinguistic Communicative Intention (XLCInt); 3) Linguistic Physical Causality (LPhC); 4) Extralinguistic Physical Causality (XLPhC). Examples of comic strips for each condition are available at http://www.psych.unito.it/csc/pers/enrici/pdf/com_int_protocol.pdf and in Enrici et al. (2011).

The task required participants to observe comic strip stories and to choose the most appropriate between two alternative story endings. Each story consisted of three consecutive pictures (development phase), followed by two alternative choice pictures presented simultaneously side by side (response phase). The first and second pictures established a story setting and introduced the characters or the objects involved, while the third picture represented the communicative intention or physical causality events. The third picture also determined the linguistic versus extralinguistic *Modality* factor level. In LCInt and LPhC, the intention or physical events, respectively, were presented in a written form. In XLCInt and XLPhC, they were presented in a pictorial form. The two alternative choice pictures presented, respectively, a plausible and implausible outcome of the communicative scenario.

Sentences used in the linguistic modality stimuli were controlled for number of words and content word frequency. Communicative intentions depicted in the extralinguistic modality consisted of conventional ideational gestures, in particular emblem gestures that convey a meaning even in the absence of speech.

The stimuli were presented in a randomized order by means of Presentation 11.0 (Neurobehavioral Systems, Albany, CA, USA), and viewed via a back-projection screen located in front of the scanner and a mirror placed on the head coil. Behavioral responses were collected via a fiber-optic response box.

MRI data acquisition

fMRI scans were acquired on a 3T Intera Philips body scanner (Philips Medical Systems, Best, NL) using an 8 channels-sense head coil (sense reduction factor = 2). Whole-brain functional images were obtained with a T2*-weighted gradient-echo, echo-planar sequence, using blood-oxygenation-level-dependent contrast. Each functional image comprised 30 contiguous axial slices (4 mm thick), acquired in interleaved mode, and with a repetition time of 2000 ms (echo time: 30 ms; field of view: 240 mm x 240 mm; matrix size: 128 x 128). Each participant underwent four functional scanning sessions (each lasting 155 scans, preceded by 5 dummy scans). A fieldmap to be used for the unwarping of echo-planar image spatial distortions was acquired for each subject prior to functional scanning.

fMRI data preprocessing

Statistical Parametric Mapping 5 (SPM5, Wellcome Department of Imaging Neuroscience, London, UK) was used for fMRI data preprocessing, including image realignment and unwarping, unified segmentation with normalization to the Montreal Neurological Institute (MNI) standard space, and smoothing by a 8 mm FWHM Gaussian kernel.

DCM network discovery analysis

Based on evidence previously obtained from these data regarding fMRI activation (Enrici et al., 2011), we tested a specific hypothesis of effective connectivity in a restricted brain network using DCM, an approach to understand distributed neuronal architectures underlying observed brain responses (Friston et al., 2003).

Specifically, we employed DCM network discovery, based on post-hoc Bayesian model selection (Friston and Penny, 2011). The network discovery approach enables one to discover the optimum model over a given model-space (Friston et al., 2011). The post-hoc optimization routine searches among a large number of possible reduced model of a full model of connections, and uses post-hoc model selection to select the best model (i.e. the one fitting the observed data with the best balance between accuracy and complexity).

The specified dynamic causal model comprised the following six brain regions (Table 1): LIFG, RIFG, MPFC, LTPJ, RTPJ, and the PREC. These six brain regions were identified based on the random-effects group analysis of functional localization, as reported in Tables 1A and 2 of our previous paper (Enrici et al., 2011). The use of the significant functional localization effects to test a hypothesis of effective connectivity on the same data does not entail a problem of circularity, since the functional localization and effective connectivity analyses are aimed at answering different questions (Stephan et al., 2010).

As a preparatory step for DCM network discovery, we used SPM8 to define for the data of each participant two General Linear Models (GLM) that were specifically designed to encompass the requirements of the intended DCM analysis. One GLM served to extract the first eigenvariate of BOLD signal from the six regions of the brain network model (voi-GLM), whereas the other GLM served as input during DCM model specification (dcm-GLM). ~~In such a way, we avoided the issue of collinearity that would have arisen by including all the required explanatory regressors in one single GLM, and which would have interfered with the definition and extraction of volumes of interest.~~ In both GLMs, the four functional scanning sessions were concatenated as one single session, and the concatenated time series were high-pass filtered at 128 s and pre-whitened by means of an

autoregressive model AR(1). No global normalization was performed. Hemodynamic evoked responses were modeled as canonical hemodynamic response functions, time-locked to the presentation of the first picture of each story and an epoch duration covering both the development and the response phases.

The voi-GLM included one stimulus-onset regressor for each experimental condition (LCInt, XLCInt, LPhC, XLPhC), and additional constant regressors to account for mean between-sessions variability. Within the voi-GLM model of each participant, we computed two t-Student contrasts defining the main effect of Intention [(LCInt + XLCInt) - (LPhC + XLPhC)], and the main effect of Modality [(LCInt + LPhC) - (XLCInt + XLPhC)], respectively. The former contrast was used to identify subject-specific volumes of interest in MPFC, LTPJ, RTPJ, and PREC, whereas the latter contrast in LIFG and RIFG. Subject-specific volumes of interest were defined through a small volume correction procedure. Based on the respective contrast, we defined spherical volumes (radius = 8 mm) around the group-level coordinates (Table 1), and extracted the maximum activation peak for each subject. We also checked that the subject-specific coordinates identified through this procedure actually corresponded to the same anatomical location represented by the group-level coordinates. We extracted the first eigenvariate of BOLD signal from spherical volumes of interest of 8 mm radius centered on the identified subject-specific coordinates. The first eigenvariates were corrected for the effects of interest (omnibus F-test), such that the signal not biased toward any particular experimental conditions.

The dcm-GLM included only one regressor modeling the stimulus onsets of both LCInt and XLCInt conditions and an associated parametric regressor modeling the LCInt versus XLCInt difference contrast (weights +1 for LCInt, and weights -1 for XLCInt).

The DCM network discovery analysis was carried out in SPM12 (revision code 4750), following a two-stage approach, with a first, single-subject level, and a second, group analysis level. At the first level, based on dcm-GLM, we specified for the data of each participant a fully connected dynamic causal model (intrinsic parameters), in which the LCInt versus XLCInt parametric regressor

provided direct input to LIFG and RIFG (direct input parameters), and modulated (modulatory parameters) all the inter-regional connections in the model (Figure 1A).

At the second level, we applied the DCM "optimize" function featuring the post-hoc Bayesian model selection algorithm, to identify the reduced model best fitting the observed functional data. The output of the post-hoc selection optimize routine is an optimized DCM that contains reduced conditional parameter estimates, representing group fixed-effects. We calculated the Bayes Factor (BF) to assess the significance of the optimized model versus the other (less optimal) models in the optimization ranking. The BF is the ratio of the model evidence of one model over another (significance cut-off: BF > 20, corresponding to strong evidence, see Kass and Raftery, 1995). This corresponds to a posterior probability of 95% that one model is better than the next best model in the comparison.

Having identified the optimal model structure at the group level, we next wished to make inferences about the parameters (connection strengths), in such a manner that would generalize to the wider population. We therefore applied classical inference using the typical summary statistic approach, based on taking each subject's estimated connection strengths to the group level (n = 24 participants). In this instance, we simply tested the null hypothesis of a departure of any effect from its prior expectation of zero. As in the standard summary statistic approach in random effects analysis, the only source of variation was between subjects. Therefore, these results might be generalized to the wider population from which we sampled our subjects. ~~After DCM optimization, we tested the random-effects group-level (n = 24 participants) significance of the intrinsic, modulatory, and direct input parameters in the optimized connectivity model, by~~ Inferential statistical analyses were carried out using R (R Foundation for Statistical Computing, 2010). First, we checked the normality of the distribution of the values pertaining to each parameter by Shapiro-Wilk test. Second, we tested for each parameter, the alternative hypothesis of a significant difference from zero. In case of a parameter with normal value distribution, we applied parametric, two-sided, one-sample t-Student tests of means. In case of a connection with non-normal value distribution, we instead applied non-

parametric, two-sided, Wilcoxon signed-rank tests of means. To account for multiple comparisons (tests on 2 direct input, 35 intrinsic, and 27 modulatory parameters), we calculated False Discovery Rate (Benjamini and Hochberg, 1995) corrected P values, and declared each test to be significant with a corrected $P < 0.05$.

Results

The Bayesian model selection algorithm yielded clear cut posterior evidence in favor of a single optimum model that was superior to a large number of possible reduced models. The post-hoc model evidence provided strong confidence that the observed fMRI activation was generated by the selected optimum model with a posterior probability of 96.49 % (Bayes factors all > 28) (Figure 1C).

The optimum model featured a connectivity architecture that was equivalent to the fully connected model that served as a departure for model optimization, with the exception of a few parameters that were pruned by the optimization algorithm in converging to an optimum model. The pruned parameters were the intrinsic connection from PREC to MPFC, and the modulatory connections from PREC to LIFG, from RTPJ to LIFG, and from RTPJ to MPFC (Figure 1D).

While the reduced parameter estimates of the optimum model represent fixed-effects that have validity limited to the collected data sample, we also wanted to assess the validity of the connectivity parameters at the general population level. To this aim, we performed a random-effects group-level analysis on the direct input, intrinsic, and modulatory parameters in the optimized model. We found significantly different from zero estimates for both direct input parameters (Table 2A): the mean input effect to LIFG indicated a stronger activation of LIFG induced by the LCInt versus XLCInt modality; in turn, the mean input effect to RIFG indicated a stronger activation of RIFG induced by the XLCInt versus LCInt modality. With respect to intrinsic connectivity parameters, we found condition-independent significantly different from zero estimates in the inhibitory self-connections of all six brain regions comprised in the model. Additionally, we found significant estimates in three connections originating from MPFC (interestingly, all but the connections to LIFG and RIFG), in two

connections originating from RTPJ, in one connection originating from LTPJ, one connection from PREC, and one connection from LIFG (Table 2B). Finally, with respect to modulatory parameters, we found significantly different from zero estimates in three connections originating from LIFG, in all five connections originating from RIFG, and in the connection from LTPJ to MPFC (Table 2C). The three modulatory effects originating from LIFG were all positive in sign, indicating a stronger modulation of these connections by LCInt than by XLCInt; in turn, the modulatory effects originating from RIFG were all negative, indicating a stronger modulation of these connections by XLCInt than by LCInt. Interestingly, the connection from RIFG to LIFG was included among the connections that were more strongly modulated by XLCInt than by LCInt, whereas the respective connection from LIFG to RIFG was not significantly modulated.

The significant random-effects optimized model connectivity architecture is summarized in Figure 1E.

Discussion

We used DCM post-hoc model optimization to determine the best model fit in terms of effective connectivity architecture that accounted for the different spread of activation induced by linguistic and extralinguistic intentional communication within the IPN network. The first striking observation is that there is one and only one connectivity architecture that accounts for the regional activations measured in our fMRI study, in that the optimum model turned out to be superior to a large number of possible model configurations. The optimum connectivity architecture is largely equivalent to a fully connected model, with just one intrinsic and three modulatory connections eliminated, and is thus suggestive of an overall strong functional integration between the six brain regions included in the network. Furthermore, the superiority of this particular connectivity architecture indicates that the activation propagation within the network, its direction, and modality-specificity, are strictly regulated, and not variable or random.

A second observation is that, at the random-effects group-level, the functional region-specific activation effects, which in the context of DCM are represented by the direct input parameters, were entirely consistent with our previously reported findings stemming from the same fMRI data (Enrici et al., 2011). The direct input to LIFG was stronger for LCInt than XLCInt, whereas the direct input to RIFG was stronger for XLCInt than LCInt. This hemispheric lateralization asymmetry replicates the one that we have observed and reported before (Enrici et al., 2011). The findings corroborate the hypothesis formulated in the present study, namely that the LIFG and RIFG represent the modality-specific gateways allowing linguistic and extralinguistic stimulus information, respectively, to be propagated to the IPN.

A third fundamental observation concerns the intrinsic connectivity architecture of the optimum model, that is the connectivity parameters representing condition-independent signal propagation in the network, occurring in a comparable manner for communicative intention processing in the linguistic and extralinguistic modalities[1]. We found significant random-effects group-level parameters in eight inter-regional connections. Seven out of these eight connections originated from IPN brain regions. Importantly, the MPFC was the brain region from which the greatest number of connections originated, and all three connections departing from the MPFC were directed to the other three IPN brain regions, with no connections reaching the input gateways, namely the LIFG and RIFG. This indicates that the MPFC has a prominent orchestration role within the IPN, possibly propagating the modality-independent activation information in a top-down mode. Three other significant intrinsic connections were serially organized, representing a putative information flow, from the RTPJ and LTPJ back to the MPFC, via the LIFG. This could represent a recirculation of information, from MPFC to the other IPN regions and backward to MPFC. Information looping is required in the context of the present communicative intention processing task, which involves the integration of perceptual and social interaction information over a prolonged interval of several seconds for each trial. This finding may also suggest a role of the LIFG as a functional node that allows for a continuous re-update of stimulus information to be fed into the IPN. This intrinsic effect

is modality independent, suggesting that the LIFG re-update mode is equally implicated in both the linguistic and extralinguistic modalities. The presence of a significant modulatory connection from RIFG to LIFG (Figure 1D), which was stronger for XLCInt than LCInt, further speaks in favour of the LIFG involvement not only for the linguistic but also for the extralinguistic condition, therefore of its modality-independent re-update mode. It must be noted, however, that our fMRI data, and the size of volumes of interest as defined for the DCM analysis, may lack sufficient spatial resolution to detect possible modality-specific functional sub-divisions within the LIFG. Future studies endowed with finer spatial resolution or using different techniques may better clarify this issue.

Finally, two significant intrinsic connections originating from IPN brain regions were directed to the RIFG (from PREC to RIFG, and from RTPJ to RIFG). This result is more difficult to explain, since there are no other intrinsic connections that depart from the RIFG and allow the modality-independent information to propagate further to other regions of the dynamic causal network. One speculative possibility is that these intrinsic connections mediate the flow of feedback information from the IPN to the extralinguistic input gateway. The presence of a symmetric intrinsic connection in the left hemisphere (from LTPJ to LIFG) may suggest that the same type of feedback signaling from the IPN also occurs for the linguistic modality (note that in this view, the connection from LTPJ to LIFG would have a dual function, as it is involved both in the modality-independent information looping and in feedback signaling).

The most compelling observation in the present study is the presence of significant modality-specific propagation effects. We found that the significant modulatory connection effects originating from LIFG displayed a stronger modulation by LCInt than by XLCInt, whereas the connections originating from RIFG displayed a stronger modulation by XLCInt than by LCInt. This pattern of results is entirely compatible with our a priori hypothesis that the LIFG acts as a linguistic modality-specific gateway of stimulus information to the IPN, whereas the RIFG represents the extralinguistic modality-specific gateway.

Although to date no studies have investigated the relation between input gateways and the ToM network specifically associated to communicative intention processing, two studies analyzed functional and effective connectivity of brain regions associated to ToM processing (Atique et al., 2011; Hillebrandt et al., 2013). Atique and colleagues (2011) used a story completion task in a comic strip form similar to our task, and analyzed task-specific connectivity of ToM brain regions during private and social (affective) ToM: private ToM cartoons depicted a *single* character in a situation that required an action whereas social ToM depicted *two or more* characters in an emotional situation. It is interesting to note that, in the social interaction condition, i.e., affective ToM vs. cognitive ToM, the authors found an overall increase in functional connectivity covariance among IPN brain regions (MPFC, PREC, RTPJ, LTPJ). Hillebrandt and colleagues (2013) used DCM to investigate effective connectivity between MPFC and posterior brain areas, such as the medial temporal gyrus, a region close to TPJ, and the superior occipital gyrus. Using a perspective taking communicative task that requires participants to take into account another person's perspective following auditory instructions of a fictional director character, the study manipulated both the social nature of the stimuli (director present or absent) and executive task demands (perspective taking congruent or incongruent from one's own). The findings showed that the presence of a social cue, but not the executive task demand, increased the strength of the backward connections originating from the MPFC. In turn, forward connections from the posterior regions, as well as backward connections from medial temporal to superior occipital gyrus were not as strongly modulated. These results are in line with the prominent orchestration role of the MPFC we found in the present study, in particular in propagating forward modality-independent activation information.

An interesting domain of investigation is the temporal course of ToM-related brain activity. Although several studies have elucidated the anatomical bases of ToM ability, few studies analyzed the integration between the temporal dynamics and the spatial localization of this process (Liu et al., 2004; Mossad et al., 2016; Vistoli et al., 2011). Early stages of social processes were investigated by Vistoli and colleagues (2011), using magnetoencephalography (MEG) with an intention attribution

task similar to ours that depicted one or two characters performing intentional actions. Main significant activations of the IPN brain areas were reported between 100 and 700 ms, with an intention processing effect starting at 240 ms post stimulus. Results showed earlier onset of activation in the right hemisphere compared to the left hemisphere: in particular, during a 390-440 ms time-window the RTPJ and LTPJ showed modulation in intention processing in relation to different aspects. Namely, the RTPJ reflected the predominant role in attribution of intentions rather than in the detection of social cues per se, whereas the LTPJ predominantly responded simply on the presence of a character. Interestingly, in these early stages, the MPFC involvement was not associated to intention processing but, like LTPJ, responded to the presence of a character. The inferential processes associated to MPFC only occurred in a later time-window, that is after 700 ms. In agreement with these findings, Liu et al. (2004) analyzed late stages of social processes using electroencephalography (EEG) with a false belief task using cartoon animations. The late involvement of the MPFC in inferential social processes emerged as an enhanced EEG component around 800 ms post-stimulus in left frontal electrodes when participants thought about the mental states of a character. More recently, Mossad et al. (2016) used MEG during a false belief task with cartoon drawings and found activations of the whole IPN as well as of the RIFG. In particular, they found a specific right lateralized onset of ToM processing at 100 ms, with strong activation in the RTPJ from 150 ms to 225 ms, in the right PREC from 275 ms to 375 ms, in the RIFG from 200 ms to 300 ms, and in the MPFC from 300 ms to 400 ms. According to the authors, the RTPJ has a role in early orienting processes for belief inference. This is then followed by RIFG activation, underlying the inhibition of one's own beliefs, and finally by MPFC activation, underlying the integration of competing mental representations involved in social inferences.

Due to the coarse temporal resolution of fMRI, it is not straightforward to integrate the findings of the present fMRI study with those just reviewed, that analyzed the fine temporal course of activation in IPN and associated brain regions. The most intriguing challenge pertains to the apparently different MPFC role that the results of two methodologies reveal. Namely, in high-

temporal resolution studies, the activation of the MPFC consistently kicks in at a relatively late stage, preceded by other IPN regions such as TPJ and PREC, such that the signal within the IPN seems to spread from posterior regions to the MPFC. In the present fMRI-DCM study, the main direction of condition-independent signal propagation within the IPN appears to take place in the opposite direction, i.e. from the MPFC to posterior regions. However, fMRI is not sensitive enough to the network dynamics occurring within the first 100 to 800 ms after stimulus detection, but rather reflects integration processes over several seconds. Although speculative, a possible reconciliation model accounting for this apparent discrepancy may therefore contemplate an early temporal phase, not detected by fMRI-DCM, in which posterior IPN regions detect a social or private intentional situation, and a later phase, in which the MPFC takes over the integration of this complex information, particularly affective and social aspects. Accordingly, posterior IPN region may first drive the intervention of MPFC (forward signal propagation), and subsequently the MPFC may orchestrate information processing within the entire IPN (backward signal propagation).

In addition, previous fine temporal course studies did not specifically focus on communicative intention, but rather more generally on ToM inferential processes that do not necessarily entail a communicative act. Thus, the role of the inferior frontal cortices, when found activated (RIFG in Mossad et al., 2016), cannot be ascribed to communication processing. When, in turn, the intentional situation involves a communicative act, such as in the task used in the present study, the inferior frontal cortices (LIFG or RIFG, depending on, respectively, the linguistic or extralinguistic communication format) specifically activate and feed information within the entire IPN. Since, in our story completion task, the communicative act was only depicted after introducing the characters and the situation, it is plausible that the IPN dynamics discussed above (reconciliation model) occur first and reach a steady-state, and are then subsequently perturbed and modified by the communicative intention information entering the IFG input gateways.

It will be important for future studies on communicative intention to challenge this putative signal propagation model by means of high temporal resolution techniques, such as MEG, combined with effective connectivity analysis.

Conclusions

The present fMRI study employing DCM network discovery provided strong Bayesian posterior evidence for the existence of a well-defined effective connectivity architecture mediating the functional integration between the IPN and the inferior frontal cortices. The LIFG and RIFG thus most likely represent modality-specific gateways that allow, respectively, linguistic and extralinguistic communicative information to be integrated in the agential situation that is being the object of ToM inferences.

Note

[1] The inhibitory self-connections are an essential Bayesian prior in dynamic causal models (Friston et al., 2003) but are not particularly meaningful in the context of the hypotheses for the present study focusing on network-level interactions. Therefore, they will not be discussed here any further.

Acknowledgments

Ivan Enrici was supported by University of Turin grants ("Ricerca scientifica finanziata dall'Università 2013-2016" Linea Generale and Linea Giovani). Mauro Adenzato was supported by MIUR of Italy (FIRB 2012-2017, RBFR12FOBD_001) and by the University of Turin (Ricerca scientifica finanziata dall'Università "Cognizione sociale e attaccamento in popolazioni cliniche e non cliniche").

References

Aboulafia-Brakha, T., Christe, B., Martory, M.D., Annoni, J.M. (2011). Theory of mind tasks and executive functions: a systematic review of group studies in neurology. *Journal of Neuropsychology*, *5*(Pt 1), 39-55.

Atique, B., Erb, M., Gharabaghi, A., Grodd, W., Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing, *NeuroImage*, *55*, 1899–1911.

Bara, B. G. (2010). *Cognitive pragmatics*. Cambridge, MA: MIT Press.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex*. *24* (10): 2572-2578.

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.

Catani, M., & Bambini, V. (2014). A model for Social Communication And Language Evolution and Development (SCALED). *Current Opinion in Neurobiology*, *28*, 165–171.

Ciaramidaro A., Adenzato M., Enrici I., Erk S., Pia L., Bara B.G., et al. (2007). The intentional network: How the brain reads varieties of intentions. *Neuropsychologia*, *45*, 3105–3113.

Corradi-Dell'Acqua C., Hofstetter C., & Vuilleumier P. (2014). Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *9*, 1175-1184.

Eddy, C. M., Mitchell, I. J., Beck, S. R., Cavanna, A. E., & Rickards, H. E. (2010). Impaired comprehension of nonliteral language in Tourette syndrome. *Cognitive and Behavioral Neurology*, *23*, 178-184.

Enrici, I., Adenzato, M., Cappa, S., Bara, B. G., & Tettamanti, M. (2011). Intention Processing in Communication: A Common Brain Network for Language and Gestures. *Journal of Cognitive Neuroscience*, *23*(9), 2415–2431.

Friston, K., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Friston, K. J., Li, B., Daunizeau, J., & Stephan, K. E. (2011). Network discovery with DCM. *NeuroImage*, *56*(3), 1202–1221.

Friston, K., & Penny, W. (2011). Post hoc Bayesian model selection. *NeuroImage*, *56*(4), 2089–2099.

Ghosh, B. C. P., Calder, A. J., Peers, P. V., Lawrence, A. D., Acosta-Cabronero, J., Pereira, J. M., et al. (2012). Social cognitive deficits and their neural correlates in progressive supranuclear palsy. *Brain*, *135*, 2089-2102.

Hillebrandt, H., Dumontheil, I., Blakemore, S.J., Roiser, J.P. (2013). Dynamic causal modelling of effective connectivity during perspective taking in a communicative task, *NeuroImage, 76*, 116–124.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kemmerer, D. (2014). Sentence Comprehension. In *Cognitive Neuroscience of Language* (pp. 419–470). Psychology Press.

Lee, T. M. C., Ip, A. K. Y., Wang, K., Xi, C.-H., Hu, P.-P., Mak, H. K., et al. (2010). Faux pas deficits in people with medial frontal lesions as related to impaired understanding of a speaker's mental state. *Neuropsychologia*, *48*, 1670-1676.

Liu, D., Sabbagh, M.A., Gehring,W.J.,Wellman, H.M., (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *Neuroreport*, *15*, 991–995.

Mossad, S.I., AuCoin-Power, M., Urbain, C., Smith, M.L., Pang, E.W., Taylor, M.J. (2016). Thinking about the thoughts of others; temporal and spatial neural activation during false belief reasoning. *NeuroImage*, *134*, 320–327.

Prat, C. S., Mason, R. A., & Just, M. A. (2012). An fMRI investigation of analogical mapping in metaphor comprehension: The influence of context and individual cognitive capacities on processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 282-294.

Shany-Ur, T., Poorzand, P., Grossman, S. N., Growdon, M. E., Jang, J. Y., Ketelle, R. S., et al. (2012). Comprehension of insincere communications in neurodegenerative diseases: Lies, sarcasm and Theory of Mind. *Cortex*, *48*, 1329-1341.

Siegal, M., & Varley, R. (2006). Aphasia, language, and Theory of Mind. *Social Neuroscience*, *1*, 167-174.

Spotorno, N., Koun, E., Prado, J., van der Henst, J.-B., & Noveck,I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, *63*, 25-39.

Stephan, K. E., Penny, W. D., Moran, R. J., den Ouden, H. E. M., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, *49*(4), 3099–3109.

Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: An fMRI study. *PLoS ONE*, *7*(11), e51207.

Van Overwalle, F. (2009). Social cognition and the brain: A metaanalysis. *Human Brain Mapping*, *30*, 829-858.

Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *NeuroImage*, *54*(2), 1589-99.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*, 564-584.

Varley, R., & Siegal, M. (2006). Evidence for cognition without grammar from causal reasoning and 'Theory of Mind' in an agrammatic aphasic patient. *Current Biology*, *10*, 723-726.

Vistoli, D., Brunet-Gouet, E., Baup-Bobin, E., Hardy-Bayle, M.C., Passerieux, C., (2011). Anatomical and temporal architecture of theory of mind: a MEG insight into the early stages. *NeuroImage*, *54*, 1406–1414.

Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B.G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience*, *16*, 1854-1863.

Walter, H., Ciaramidaro, A., Adenzato, M., Vasic, N., Ardito, R.B., Erk, S., et al. (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: An fMRI study. *Social Cognitive and Affective Neuroscience, 4,* 166-176.

Willems, R. M., Benn, Y., Hagoort, P., Toni, I., & Varley, R. (2011). Communicating without a functioning language system: Implications for the role of language in mentalizing. *Neuropsychologia*, *49,* 3130-3135.

Willems, R. M., & Varley, R. (2010). Neural insights into the relation between language and communication. *Frontiers in Human Neuroscience*, *4,* 203.

Table 1.

Group-level, random-effects activation MNI coordinates (as reported in Enrici et al., 2011) of the brain regions included in the dynamic causal model

Brain region	**x**	**y**	**z**
Left inferior frontal gyrus (LIFG)	-42	12	24
Right inferior frontal gyrus (RIFG)	50	8	20
Superior medial frontal gyrus (MPFC)	-6	54	32
Left middle temporal gyrus (LTPJ)	-52	-64	20
Right superior temporal gyrus (RTPJ)	56	-46	20
Precuneus (PREC)	2	-56	40

Table 2. **Group-level, random-effects tests of significance of the optimized dynamic causal model estimates**

Table 2A. **Direct input parameter estimates**

Input region	**Mean strength (Hz)**	**SD**	**FDR corrected P value**
LIFG	0.07	0.13	0.0048 §
RIFG	-0.07	0.10	0.0006 §

Table 2B. **Intrinsic connection parameter estimates**

Connection	**Mean strength (Hz)**	**SD**	**FDR corrected P value**
LIFG → LIFG (self-connection)	-0.25	0.21	0.0001
RIFG → RIFG (self-connection)	-0.34	0.25	< 0.0001
MPFC → MPFC (self-connection)	-0.21	0.28	0.0054
PREC → PREC (self-connection)	-0.30	0.18	< 0.0001
LTPJ → LTPJ (self-connection)	-0.36	0.17	< 0.0001
RTPJ → RTPJ (self-connection)	-0.28	0.19	0.0001 §
LIFG → MPFC	-0.18	0.20	0.0010
MPFC → LTPJ	0.26	0.40	0.0149
MPFC → RTPJ	0.50	0.75	0.0112
MPFC → PREC	0.45	0.74	0.0190
PREC → RIFG	-0.13	0.24	0.0429
LTPJ → LIFG	-0.28	0.53	0.0194 §
RTPJ → RIFG	0.11	0.18	0.0048 §
RTPJ → LTPJ	0.18	0.24	0.0069

Table 2C. **Modulatory parameter estimates**

Connection	**Mean strength (Hz)**	**SD**	**FDR corrected P value**
LIFG → MPFC	0.50	0.69	0.0013 §
LIFG → LTPJ	0.60	0.81	0.0066

LIFG → RTPJ	1.08	0.95	0.0003
RIFG → LIFG	-0.38	0.88	0.0022 §
RIFG → MPFC	-0.40	1.00	0.0267 §
RIFG → LTPJ	-0.71	0.91	0.0048
RIFG → RTPJ	-1.12	1.19	0.0013
RIFG → PREC	-0.59	1.05	0.0347
LTPJ → MPFC	-0.22	0.45	0.0097 §

§ In these cases, we applied non-parametric, two-sided, Wilcoxon signed-rank tests of means, due to a non-normal value distribution. In all other cases, we applied parametric, two-sided, one-sample t-Student tests of means.

Figure 1

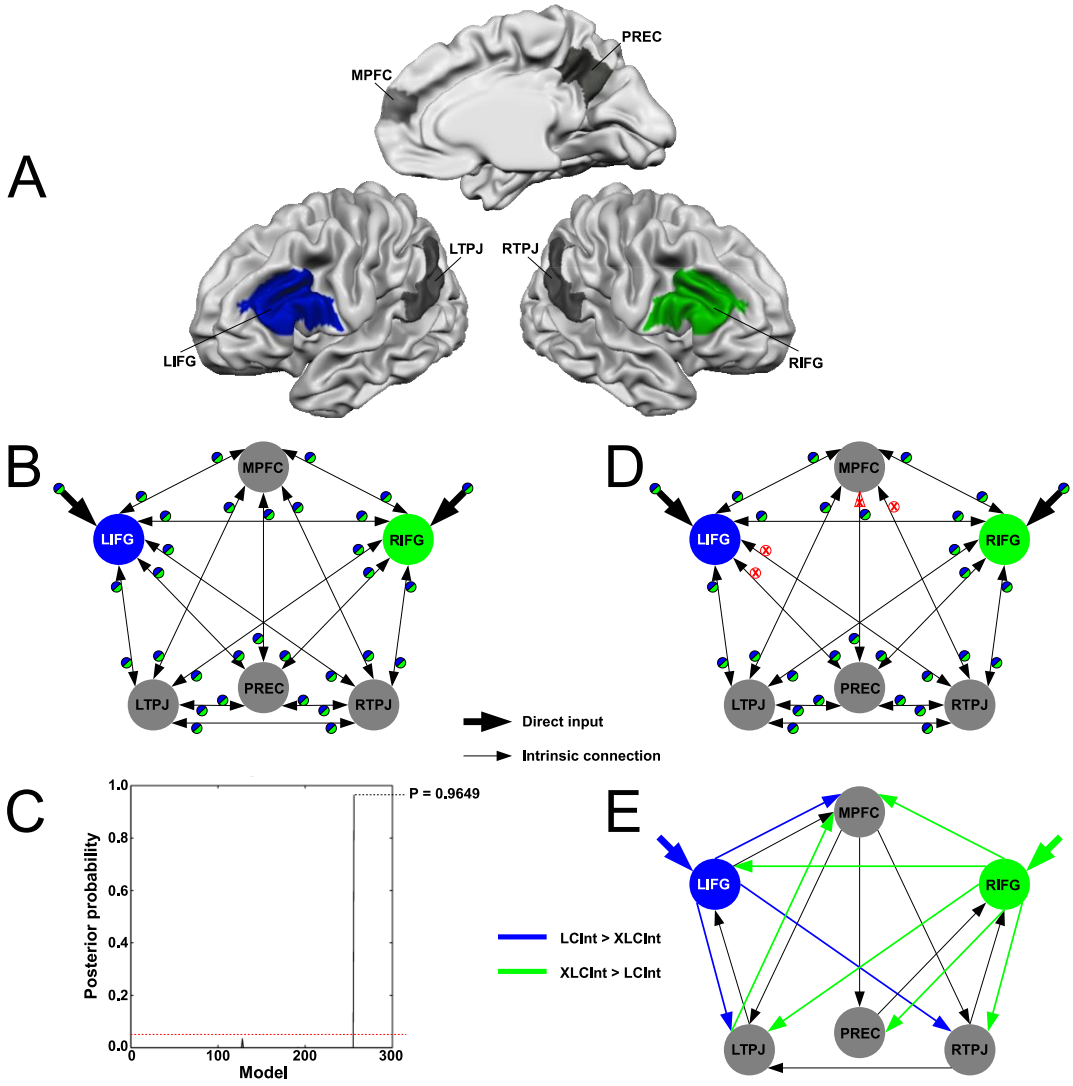
A
PREC
MPFC
LTPJ
RTPJ
LIFG
RIFG
B
MPFC
LIFG
RIFG
PREC
LTPJ
RTPJ
D
MPFC
LIFG
RIFG
PREC
LTPJ
RTPJ
Direct input
Intrinsic connection
C
P = 0.9649
Posterior probability
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
Model
E
MPFC
LIFG
RIFG
PREC
LTPJ
RTPJ
LCInt > XLCInt
XLCInt > LCInt

Figure Captions

Figure 1. **Dynamic causal model post-hoc optimization.** A) Schematic view of the brain regions included in the effective connectivity models, including the four IPN brain regions (MPFC, LTPJ, RTPJ, PREC, all in dark gray), and LIFG (blue) and RIFG (green) as, respectively, the linguistic and extralinguistic input gateways. B) Connectivity architecture of the fully connected model that served as a departure for model optimization. Blue-green circles represent the LCInt versus XLCInt parametric regressor that provided direct psychological input to the model and modulated the inter-regional connections. C) Graph showing the posterior probability of all models generated by the post-hoc optimization. The probability of the optimum model is indicated by a black dashed line. The red dashed line indicates the Bayes Factor significance upper cut-off, corresponding to strong evidence in favor of the optimum model versus the other models. D) Connectivity architecture of the optimum model. The red triangle indicates the only one intrinsic connection pruned by the reduction algorithm, whereas red circles indicate the pruned modulatory connections. E) Schematic connectivity architecture with the significant random-effects parameters of the optimum model. Blue lines indicate stronger direct input (thick arrows) or modulatory (thin arrows) effects induced by LCInt versus XLCInt. Green lines indicate stronger effects induced by XLCInt versus LCInt. Please note that inhibitory self-connections are nowhere represented in this figure.