Deficits in cognitive and affective theory of mind relate to dissociated lesion patterns in prefrontal and insular cortex

Article *in* Cortex · March 2020

DOI: 10.1016/j.cortex.2020.03.019

CITATIONS
24

READS
305

6 authors, including:

Corrado Corradi-Dell'Acqua
Università degli Studi di Trento
70 PUBLICATIONS **1,731** CITATIONS
SEE PROFILE

Roberta Ronchi
University of Geneva
40 PUBLICATIONS **1,471** CITATIONS
SEE PROFILE

Marine Thomasson
University of Geneva
24 PUBLICATIONS **161** CITATIONS
SEE PROFILE

Arnaud Saj
Université de Montréal
127 PUBLICATIONS **1,896** CITATIONS
SEE PROFILE

Running head: Lesion patterns in cognitive and affective theory of mind

Deficits in cognitive and affective theory of mind relate to dissociated lesion patterns in prefrontal and insular cortex

Corrado Corradi-Dell'Acqua[a,b*], Roberta Ronchi[c,d], Marine Thomasson[c,e], Therese Bernati[d], Arnaud Saj[d,f] & Patrik Vuilleumier[b,c,g]

[a] Theory of Pain Laboratory, Department of Psychology, Faculty of Psychology and Educational Sciences (FPSE), University of Geneva, Geneva, Switzerland.

[b] Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland.

[c] Laboratory of Behavioural Neurology and Imaging of Cognition, Department of Neuroscience, University Medical Center, University of Geneva, Geneva, Switzerland.

[d] Department of Clinical Neurosciences, Geneva University Hospital, Faculty of Medicine.

[e] Clinical and Experimental Neuropsychology Laboratory, Department of Psychology, Faculty of Psychology and Educational Sciences (FPSE), University of Geneva, Geneva, Switzerland. University of Geneva, Switzerland.

[f] Department of Psychology, University of Montréal, Canada.

[g] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland.

*Correspondence should be addressed to: Corrado Corradi-Dell'Acqua, University of Geneva – Campus Biotech, Ch. des Mines 9, CH-1211, Geneva, Switzerland. Tel: +41223790958. E-mail: corrado.corradi@unige.ch

Keywords: "mentalizing"; "perspective taking"; "empathy"; "pain"; "lesion symptom mapping"

Abstract

Neuroimaging studies suggest that understanding emotions in others engages brain regions partially common to those associated with more general cognitive Theory-of-Mind (ToM) functions allowing us to infer people's beliefs or intentions. However, neuropsychological studies on brain-damaged patients reveal dissociations between the ability to understand others' emotions and ToM. This discrepancy might underlie the fact that neuropsychological investigations often correlate behavioural impairments only to the lesion site, without considering the impact that the insult might have on other interconnected brain structures. Here we took a network-based approach, and investigated whether deficits in understanding people's emotional and cognitive states relate to damage to similar or differential structures. By combining information from 40 unilateral stroke damaged patients, with normative connectome data from 92 neurotypical individuals, we estimated lesion-induced dysfunctions across the whole brain, and modeled them in relation to patients' behavior. We found a striking dissociation between networks centered in the insular and prefrontal cortex, whose dysfunctions led to selective impairments in understanding emotions and beliefs respectively. Instead, no evidence was observed for neural structures shared between the two conditions. Overall, our data provide novel evidence of segregation between brain networks subserving social inferential abilities.

1. Introduction

How we understand emotions felt by other people is a central but still unresolved question in cognitive-affective neuroscience. Many studies used brain imaging techniques to explore the neural foundations of this ability, and suggested that they might rely on a widespread network, involving the anterior insula, the cingulate cortex, supramarginal/postcentral gyrii, etc. (see, Bzdok et al., 2012; Del Casale et al., 2017; Ding et al., 2019; Fan et al., 2011; Jauniaux et al., 2019; Lamm et al., 2011; Timmers et al., 2018, as meta-analyses). More specifically, these regions seem involved when individuals witness and empathize with others in different affective states (e.g., disgust, pain, and even happiness) displayed through different means (facial expressions, abstract cues, written texts etc. – see also, Bruneau et al., 2012, 2013; Corradi-Dell'Acqua et al., 2011, 2014, 2016; Hennenlotter et al., 2005; Jabbi et al., 2007; Jacoby et al., 2016; Silani et al., 2013; Wicker et al., 2003). Overall, this network has been often considered the core mechanism underlying emotion recognition and empathy, according to which individuals understand others' affect by simulating their behavioral/physiological reactions (smiles, tears, shivers, etc.) on one's own body (Bernhardt & Singer, 2012; Goldman & de Vignemont, 2009; Stietz et al., 2019).

The understanding of others' emotions is not based exclusively on processes of affective resonance (in which individuals "share" the state observed in others), but can also rely on a cognitive pathway that underlies abstract, propositional knowledge about how people's behaviour, reactions and thoughts relate to specific emotions (Shamay-Tsoory, 2011; Stietz et al., 2019). This process is often referred to as perspective-taking or cognitive 'theory of mind' (ToM), and corresponds to the ability to infer and represent others' beliefs or goals (Amodio &

Frith, 2006; Saxe et al., 2004). Accordingly, a wealth of studies mapped the neural structures underlying cognitive ToM abilities, and pinpointed a network comprising the temporo-parietal junction, middle temporal gyrus, precuneus, as well as lateral and medial prefrontal cortices (see, Bzdok et al., 2012; Krall et al., 2015; Molenberghs et al., 2016; Schurz et al., 2017; Van Overwalle, 2009; van Veluw & Chance, 2014, as meta-analyses). Importantly, parts of this ToM-network activate, not only when participants judge beliefs and thoughts of others, but also when they evaluate their emotions (Bodden et al., 2013; Corradi-Dell'Acqua et al., 2014; Hynes et al., 2006; Peelen et al., 2010; Schlaffke et al., 2015; Sebastian et al., 2012; Völlm et al., 2006). These observations were interpreted in terms of a mentalistic (or representational, Flavell, 1999; Saxe et al., 2004) interpretation of affect attribution, according to which the emotional experience is not represented exclusively in terms on their bodily manifestations (smiles, tears, shivers, etc.) but also in relation to mental states such beliefs, thoughts and intentions.

Mentalistic interpretations of affect attribution are prevalently motivated by neuroimaging evidence that emotional processing recruits in part similar neural processes than cognitive ToM. However, this overlap between neural structures seems in apparent contrast with other research using neurostimulation, developmental and neuropsychological approaches, who show how impairments in cognitive ToM can dissociate from those in inferring others' affect. For instance, stimulating dorsal (DLPFC) and medial (DMPFC) portions of the prefrontal cortex can impair selectively the appraisal of cognitive (beliefs/intentions) and affective states respectively (Kalbe et al., 2010; Krause et al., 2012). Furthermore, developmental investigations confirm that the proficiency at assessing people's cognitive states declines in elderly individuals proportionally with troubles in executive functioning and

inhibition (Bottiroli et al., 2016; Charlton et al., 2009; German & Hehman, 2006; Z. Wang & Su, 2013), but independently from the inference of others' emotions (Bottiroli et al., 2016; Z. Wang & Su, 2013).

Most critically, the proficiency at appraising people's cognitive and emotional states dissociates also following brain diseases. In particular, a wealth of studies reported deficits in these abilities following neurodegenerative disorders (see Adenzato et al., 2010; Bora et al., 2015, 2016; Henry et al., 2014; Kipps & Hodges, 2006; Kumfor et al., 2017; Poletti et al., 2012, as reviews/meta-analyses) and traumatic brain injury (Balaban et al., 2016; Biervoye et al., 2016; Campanella et al., 2014; Domínguez D et al., 2019; Happé et al., 1999; Leigh et al., 2013; Leopold et al., 2012; Muller et al., 2010; Rowe et al., 2001; Samson et al., 2004, 2005; Shamay-Tsoory et al., 2005, 2006, 2010; Shamay-Tsoory & Aharon-Peretz, 2007; Stuss et al., 2001; Yeh & Tsai, 2014). However, those researches directly comparing the two conditions often reported dissociations, for instance in patients with frontotemporal dementia where impairments in the assessment of cognitive and affective states interact differently with syndrome severity (Torralva et al., 2015) or executive functioning (Freedman et al., 2013; Kipps & Hodges, 2006; Lough et al., 2006). As for the neural structures implicated, selective impairments for understanding others' cognitive states often follow dysfunctions at the level of lateral and medial prefrontal cortex (Bejanin et al., 2017; Shamay-Tsoory & Aharon-Peretz, 2007), whereas selective impairments in understanding others' affect follow dysfunctions to the most ventral orbitofrontal cortex (Shamay-Tsoory et al., 2006, 2010; Shamay-Tsoory & Aharon-Peretz, 2007) and medial temporal pole (Bejanin et al., 2017). These dissociations suggest that cognitive ToM abilities may not be necessary (at least not always) for assessing people's affective states.

Overall, the literature provides a mixed set of results, with some studies supporting common neural structures underlying understanding others' cognitive states and emotions, and others suggesting instead segregated processes. This might reflect the heterogeneity of the human brain, which is able to represent others' affect through independent pathways, only one of which relies on the same processes underlying perspective taking and cognitive ToM (Shamay-Tsoory, 2011; Stietz et al., 2019). However, previous results might also be confounded by the experimental approach adopted: whereas functional associations between cognitive and affective states have been described exclusively in neuroimaging research on neurotypical individuals exploiting full information from brain networks (Corradi-Dell'Acqua et al., 2014; Hynes et al., 2006; Schlaffke et al., 2015; Sebastian et al., 2012; Völlm et al., 2006), dissociations were prevalently reported in studies describing cerebral dysfunction on isolated regions (Bejanin et al., 2017; Kalbe et al., 2010; Shamay-Tsoory et al., 2006, 2010; Shamay-Tsoory & Aharon-Peretz, 2007) without considering the impact that these impairments might have on other brain structures. Indeed, it has been often pointed out that both focal neurostimulations (Ruff et al., 2009) and local brain injury (Vuilleumier et al., 2004, 2008) could affect the functional properties of distant interconnected regions, at the point that scholars often interpret similar symptomatology associated with different brain dysfunctions as an impairment of the same interconnected network (Bartolomeo et al., 2007; Boes et al., 2015; He et al., 2007; Y. Wang & Olson, 2018; Wawrzyniak et al., 2018). In this view, the heterogeneity of the impairments in emotional processing and cognitive ToM might be only apparent, as different lesion loci could underlie a broader dysfunction to a common network. To the best of our knowledge, it is still unclear whether deficits in the appraisal of others' cognitive and

affective states underlie dissociations, not only in isolated regions, but also at the network - level.

Here, we engaged 40 unilateral brain damaged patients (and 24 neurotypical controls) in a study in which they read brief stories and subsequently answered questions about the protagonists' emotional state. Patients' ability at solving this emotional task [*E*] was compared with that of: a false belief task, requiring the judgment of the protagonists' beliefs [*B*]; a pain judgment task [*Pa*], requiring the judgment of the protagonists' aching sensations; and a false photograph control task [*Ph*], involving the assessment of stories about physical events without human protagonists (which serves as high-level control for the same mix of linguistic, memory and attentional abilities held to impact also the main conditions). By employing this same task in a previous neuroimaging investigation on neurotypical individuals, we found that inference of emotion and beliefs triggered shared activity patterns at the level of temporo-parietal structures, but dissociated responses at the level of the dorsal prefrontal cortex (Corradi-Dell'Acqua et al., 2014). Here, by repeating the paradigm on a clinical population, we planned to map neural correlates of lesion-induced deficits. More specifically, by combining lesional information, with normative connectome data from 92 matched neurotypical individuals, we estimated the neural structures most frequently connected with the lesion site, which could most likely exhibit dysfunctional responses due to the brain insult. This allowed us to assess whether deficits in appraising both emotions and beliefs, underlie the same or different brain networks.

2. Methods

We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedure/analysis has been pre-registered prior the research being conducted.

2.1 Participants

40 patients (16 females, median time post stroke = 195.57 days [interquartile range: 36.75-350.40]; see Table 1) with unilateral brain damage due to a stroke (25 in the right hemisphere) were continuously admitted to the study. Patients had no evidence or history of previous neurological and psychiatric history. They were screened with the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) for the assessment of the global cognitive efficiency (score = 26 [26-27], with 8 patients scoring below 26 due to neglect/language problems). They were tested in the Department of Neurology of the Geneva University Hospital. In addition, 24 neurologically unimpaired individuals (12 females) were recruited as control population (see Table 1 for statistical comparisons with patients). The inclusion criteria of this sample were changed following the suggestion of an anonymous reviewer to insure that controls were matched for both age and education to the patients. All subjects (healthy and patients) gave informed written consent according to the rules of Geneva University Hospital ethics committee.

Table 1. Demographic information.

	Patients	***Controls***	***RS dataset***	*Pat. vs. Beh.*	*Pat. vs. Rest.*
N	40	24	92	–	–
Gender	40% F	50% F	45% F	$X^2 = 2.50^{\dagger}$	$X^2 = 0.88^{\dagger}$
Age	61.5 [53-72]	67.5 [51-75.5]	67.5 [62.5-72.5]	$t_{(62)} = 0.21^{\dagger}$	$t_{(62)} = 1.55^{\dagger}$
Education	12 [10-14]	14 [12-16]	–	$t_{(62)} = 1.82^{\ddagger}$	–

Demographic information of 40 patients, compared with that of neurotypical control group, as well as that of a cohort who underwent a neuroimaging experiment in which resting state data was collected (see more details in 2.4.2 subsection). Each population is described in terms of size, gender (percentage to the overall size), Age and Education (median years with interquartile range). Statistical comparisons between Patients and the other two groups are displayed, with gender differences assessed through X^2*-test*, and age/education assessed through independent-sample *t-test*. $^{\dagger}p > 0.10$; $^{\ddagger}p \geq 0.068$.

2.2 Cognitive and Affective Theory of Mind Task

Participants underwent a modified version of the Theory of Mind paradigm, as implemented in Corradi-Dell'Acqua et al. (2014). In this task, each participant was exposed to 36 short French-written narratives describing a person engaged in various situations, each of which was followed by a question probing for their awareness of the protagonist's beliefs (*B*), emotions (*E*), or somatic aching states (pain, *Pa*). As such task relies heavily on linguistic proficiency, attention, and working memory, a high-level control condition was also included, characterized by 12 additional stories with no human protagonist but referring to an outdated physical representation on a map or photograph (photos, *Ph*). All 48 narratives (12 per condition) were presented to participants in randomized order, across two separate experimental sessions of about 10 minutes each. Please see Corradi-Dell'Acqua et al. (2014) for the full list of narratives.

For each short narrative, participants were prompted to read the text from a computer screen, and to press a key once they were finished. Subsequently the question appeared,

together with two possible answers, each located on a different side of the screen. Participants made responses by pressing one of two possible keys, placed at each hand's reach. They had to press the key corresponding to the same side as the answer they believed to be correct (i.e. press the right hand button when they felt that the correct response was on the right side of the screen). The position of the correct response on the screen was counterbalanced across all narratives. Overall, the experiment was self-paced, with potentially no time-limit for reading the story or choosing the appropriate response. When necessary, patients were further assisted by an experimenter who read out loud the text, insured good understanding of the content, and manually delivered the response.

2.3 Lesion Mapping

We used the brain images collected as part of the routine clinical investigation after the patient was admitted to the Geneva University Hospital due to acute onset of stroke symptoms. 37 patients were investigated with magnetic resonance imaging (MRI), and the lesion was most clearly demarcated in the diffusion-weighted (18 cases), T2-weighted (18), or T1 brain scans (1). The 3 remaining patients were investigated with spiral computed tomography (CT) covering the whole brain. In all cases, the lesion was mapped with the Clusterize-toolbox (https://www.medizin.uni-tuebingen.de/kinder/en/research/neuroimaging/software/), through first automated identification of the local lesion clusters on each image slice based on its intensity, followed by subsequent manual validation and potential freehand correction (Clas et al., 2012; de Haan et al., 2015). The resulting lesion-map was then normalized to the Montreal Neurological Institute (MNI) single subject template with the aid of SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/). We applied to each map a deformation field estimated

from a registered T1 (13 cases), T2 (24) or CT brain scan (3). The collection of radiological images occurred always in the acute stroke phase (median: 2.5 days [1-5.25], with two patients in sub-acute phase: 75 and 122 days respectively). Supplementary Figure 1 displays the overlay of the lesion maps from our sample population.

2.4 Data Analysis

2.4.1 Single Case Analysis.

To measure performance in the task, we calculated the percentage of correct responses over the 12 narratives of each condition. For each of these conditions, the single case Crawford *t*-test (Crawford & Howell, 1998) was implemented to assess significant decreases in accuracy of each individual patient with respect to the control group under a one-directional hypothesis (lesion-induced impairments). Furthermore, to insure that observed deficits were indeed condition-specific, we employed interaction analysis through the Revised Standardized Difference Test (Crawford & Garthwaite, 2005), which assesses whether the differential performance of individual patients across two conditions diverges significantly from that of the control sample. Effect sizes (and 95% Confidence Intervals) of single case analyses are reported in terms of Z_{CC} (for single case *t*-tests) and Z_{DCC} (for Revised Standardized Difference Test) scores, following the methodology described in Crawford et al. (2010).

***2.4.2 Network-based Lesion-Symptom Mapping** (NLSM).*

We investigated the neural networks most predictive of patients' impairments through a network-based lesion mapping. In this technique the lesion maps are combined with normative connectome data from a matched population, to obtain an estimate of the brain regions

functionally-connected with the lesion site, and which could also exhibit dysfunctional properties following brain damage. Subsequently the relationship between the network functionally connected to the lesion and continuous behavioural measures (the accuracy scores for each task) is established through parametrical mapping. Differently from traditional voxel-based approaches, NLSM is ideal for mapping symptoms which are not uniquely localized in one brain area, but could occur following impairment of different parts of a same network.

For the purpose of the present study, we took a dataset of 92 neurotypical individuals matched for age and gender with our patients group (see Table 1), who were part of a larger dataset of resting state data available at OpenfMRI database (accession number *ds000221*)[1]. Each participant underwent from one up to five a resting state sessions of 15 minutes each in the 3T Verio whole-body MRI Scanner (Siemens, Tarrytown, NY). Functional images were acquired using a 64-channel head-and-neck coil, and a multiband imaging sequence with time to recovery = 1400 ms, time to echo (TE) = 39 ms, flip angle = 69°, 64 interleaved slices, 88 x 88 in-plane resolution, 2.3 x 2.3 x 2.3 mm voxel size, and no inter-slice gap. The multiband acceleration factor was 4. The functional images of each subject were preprocessed using standard pipeline from SPM12, involving realignment to account for head movements, unwrapping using a field map image to correct for geometrical distortions due to the magnetic field inhomogeneity, artifact detection (through the Artifact Detection Tools [ART], https://www.nitrc.org/projects/artifact_detect/), normalization to the MNI single subject template (with a voxel size of 2 x 2 x 2 mm^3), and smoothing by convolution with an 8 mm full-width at half-maximum isotropic Gaussian kernel. Additionally, a group-based Independent

[1] Please note that, as in this dataset age was provided in 5 years bin (e.g. "70-75"), all comparisons with the patient population from the present research (Table 1) were obtained by approximating the precise age to the center of each bin range ("70-75" becomes "72.5").

Component Analysis was carried out to remove acquisition/reconstruction artefacts in the signal (see Supplementary Information for more details). Among the preprocessed data, we selected one functional run for each of the 92 subjects. For those participants for which multiple sessions were acquired, we selected the run associated with the least artefactual scans (as estimated in the ART-toolbox).

Following previous studies employing NLSM, we adopted seed-based connectivity analysis, with the lesion mask of each neurological patient as seed (Boes et al., 2015; Darby et al., 2017; Laganiere et al., 2016; Wawrzyniak et al., 2018). More specifically, for each neurological patient, and for each resting-state subject, a first level General Linear Model (GLM) was carried out, where functional images were modeled against the average time-course extracted from the patient's lesion site. Lesion masks were restricted to coordinates located in the grey matter (Wawrzyniak et al., 2018). To account for movement-related and global signal artifacts, we also included as nuisance covariates: the 6 realignment parameters, dummy predictors of artefactual scans (as estimated in the ART-toolbox) and the average time-courses extracted from anatomical masks of grey matter, white matter, and cerebrospinal fluid. Low frequency signal drifts were filtered using a cutoff period of 128 seconds. Serial correlations in the neural signal were accounted through exponential covariance structures, as implemented in the 'FAST' option of SPM12. This led to an overall of 3680 first-level GLMs (40 patients x 92 resting state subjects). For each neurological patient, the 92 parameter estimates of the associated GLMs were then fed in a second-level one-sample t-test using random-effect analysis. Regions exceeding a threshold corresponding to $t_{(91)} = 8.00$ (Wawrzyniak et al., 2018) were used to create patient specific binary network-mask. This led to 40 masks, one for each

patient. Supplementary Figure 1 displays the overlay of the network maps from our sample population.

The network masks were modeled against task accuracy according to standard pipeline for lesion-symptom mapping. More specifically, we focused the analysis to those voxels which were damaged in at least 10% of patients (N = 4), corresponding to a search area of 55861 voxels (corresponding to 446888 mm^3). For each coordinate within the mask, the accuracy of each condition of interest (*B*, *E*, *Pa*) was fitted against lesion presence through a linear model. To account for potential confounds unrelated to social inferential abilities, the linear regression included nuisance variables descriptive of the lesioned hemisphere, the overall lesion size, patients' gender, age, education level, MoCA score, time post-stroke and performance in the control *Ph* conditions (used as a control to quantify non-specific effects of any linguistic, memory and attentional difficulties). We used permutation techniques to apply to our data a family-wise correction for multiple comparisons at the cluster level (with an underlying height threshold corresponding to $p < 0.001$ uncorrected). Specifically, we randomly reassigned patients' behavioral scores 5000 times and, for each permutated data set, we refit the general linear model and recorded the largest cluster in the whole search area. Clusters in the original (unpermuted) data set were considered as significant only if they exceeded the 95^{th} percentile of the "largest cluster" distribution collected under permutation. Such analysis ensures that, if the null hypothesis is true (and, therefore, no consistent relation exists between brain region and behaviour), the probability of such an extended lesion would be < 5% (Nichols & Holmes, 2002). This approach, previously used in both neuroimaging (e.g., Corradi-Dell'Acqua et al., 2011, 2014, 2016; Qiao-Tasserit et al., 2018) and lesion data literature (Binder et al., 2016;

Mirman et al., 2015; Pillay et al., 2014, 2017), fits well the hybrid nature of our network masks which combines lesion masks with resting state fMRI activity. The analysis was carried out using the VLSM package (https://aphasialab.org/vlsm/) for MATLAB R2013b (Mathworks, Natick, MA) software.

We then borrowed the method from Lorca-Puls et al. (2018) who estimated the variance explained by the impairment of a brain region on behavioral performance. More specifically, for each implicated cluster (*i*) we extracted the signal from network maps from all constituent voxels; (*ii*) we averaged the signal across voxels, leading to a single value per patient ranging from 0 (region entirely spared) to 1 (region entirely dysfunctional); (*iii*) we calculated the partial correlation between dysfunction load in the region of interest and the accuracy in the condition of interest, after having adjusted for the effect of the nuisance covariates; (*iv*) we took as measure of effect size the proportion of variance explained in the correlation (R^2). The reliability of the estimated effect size was tested through 5000 bootstrap resamplings from the original dataset. For each iteration of the resampling procedure, a random selection of 40 patients (with replacement) was used to re-estimate effect size and the associated *p*-value. This will allow to assess the variability of the effects observed, as well as an estimate of the power to obtain significant effects on resampled data (see Lorca-Puls et al., 2018, for more details).

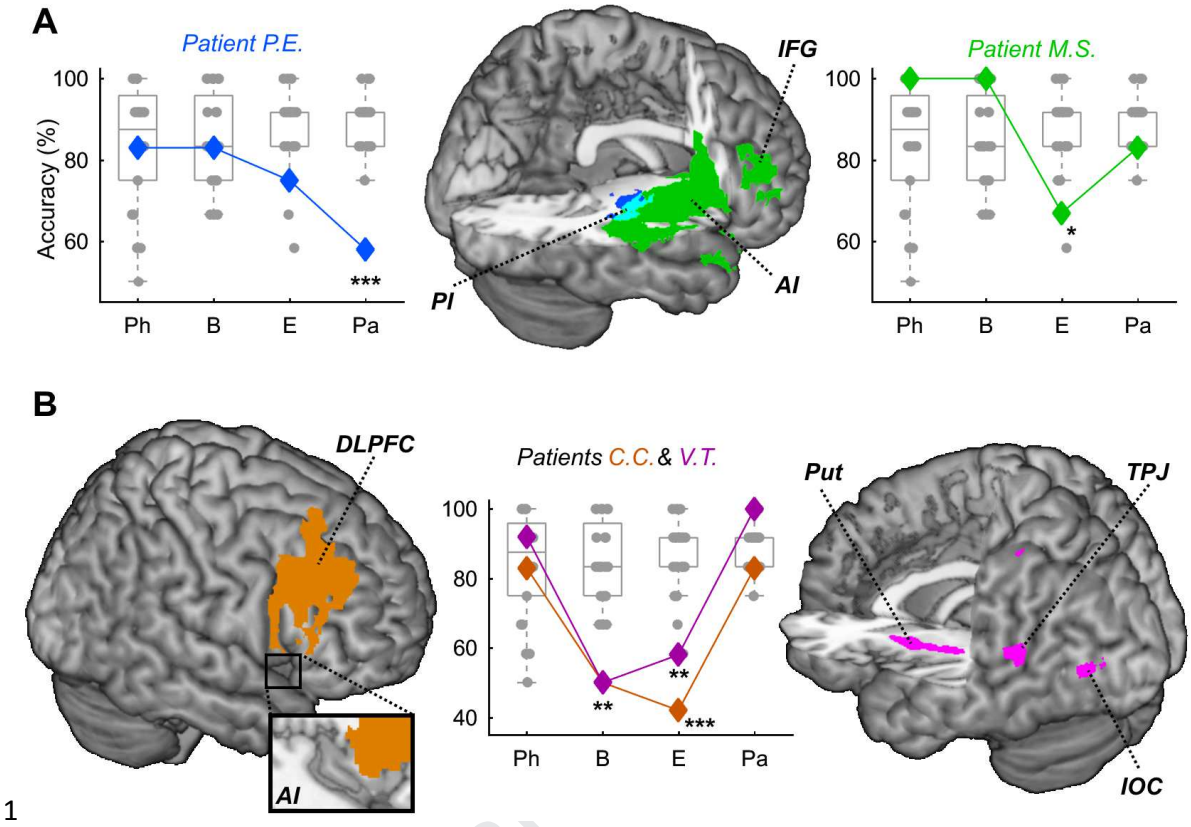
3. Results

3.1 Preliminary Group Analysis

As first step, we run a Repeated Measures ANOVA to assess whether participants' accuracy in the task changed as function of the Condition (*Ph*, *B*, *E*, and *Pa* – modeled s a within-subject factor, and the Group (Left-damaged Patients, Right-damaged Patients and Controls – modeled as between-subjects factor). This analysis revealed a main effect of Group ($F_{(2,61)}$ = 4.42, p = 0.016, $\eta_p^2 = 0.13$) a main effect of Condition ($F_{(3,183)}$ = 5.84, p < 0.001, $\eta_p^2 = 0.09$), but no Group*Condition interaction ($F_{(6,183)}$ = 0.52, p = 0.792, $\eta_p^2 = 0.02$). Bonferroni-corrected post-hoc *t*-tests were used to explore these main effects. For Groups, we found that left-damaged patients had ~10 percentage accuracy points less than controls ($t_{(37)}$ = 2.80, p = 0.008 [critical α = 0.05/3 = 0.016], Cohen's d = 0.83), whereas no difference was observed between right-damaged patients and the other two groups ($|t_{(37)}| \leq 1.93$, $p \geq 0.061$, $d \leq 0.53$). As for the conditions, the inference of pain was associated with ~6 percentage accuracy points more than all other conditions ($t_{(63)} \geq 3.37$, $p \leq 0.001$ [critical α = 0.05/6 = 0.008], $d \geq 0.42$). Instead, conditions *Ph*, *B* and *E* did not differ from one another ($|t_{(63)}| \leq 0.95$, $p \geq 0.345$, Cohen's $d \leq 0.12$). We then repeated the analysis by assessing the role played by age and education level and, within the patients' group, the time post-stroke and global cognitive efficiency score (from MoCA). The inclusion of either of these variables as covariate had no influence on the effects described above. Furthermore, none of the covariates influenced the task, either as a main effect, or as interaction with other variables. The only exception was the MoCA score, which influenced selectively the accuracy of the *B* task (but not *Ph*, *E*, or *Pa*), with individuals with larger scores displaying the higher accuracy. See Supplementary Information for full details.

3.2 Single Case Analysis

We then employed a single-subject approach and compared each individual's performance in each condition to the healthy control population. Across the overall group, 14 patients showed *lower-than-control* performance in at least one condition (see Table 2). In particular, one case (patient *P.E.*) showed a difficulty in the assessment of other people's pain ($t_{(23)}$ = 5.15, p (one-tailed) < 0.001, Z_{CC} = 5.26 [95% confidence intervals: 3.69, 6.81]), but not in the other three conditions ($t_{(23)} \leq 1.08$, p (one-tailed) ≥ 0.145). Critically, interaction analyses confirmed that, in this patient, scores were reliably lower for *Pa* than in *B*, *E,* and *Ph* scenarios ($t_{(23)} \geq 3.29$, $p <$ 0.003, $Z_{DCC} \geq 3.47$ [2.08, 5.07]), supporting the specificity of the trouble. In a similar vein, a second case (patient *M.S.*) showed a difficulty only in the assessment of others' emotions ($t_{(23)}$ = 1.87, p (one-tailed) = 0.037, Z_{CC} = 1.87 [1.22, 2.58]; all other conditions, $t_{(23)} \leq 1.06$, p (one-tailed) ≥ 0.147). Interaction analysis also confirmed the more severe impairment at *E* than at *B* and *Ph* ($t_{(23)} \geq 2.29$, $p < 0.032$, $Z_{DCC} \geq 2.39$ [1.63, 3.23]), but not *Pa* ($t_{(23)}$ = 0.64, p = 0.525), ruling out potential confounds related to language, attention, or memory that would also impact understanding non-affective states. Figure 1A displays these patients' scores as well as their lesions. In both cases, damage included the right insular cortex, with the impairment in assessing pain circumscribed to its most posterior portion (and neighboring white matter tissue). Instead, the impairment in assessing emotions extended to the most anterior section of the insula, as well as part of the inferior frontal gyrus. Three other patients who showed *lower-than-control* proficiency in one condition only (see Table 1), although none was associated with significant interaction effects that could insure the specificity of the trouble.



1

Figure 1. Single Case Analysis. Individual performance of (**A**) two patients (P.E. and M.S.) showing selective impairment in one condition only, and (**B**) two patients (C.C. and V.T.) showing conjoint impairment in two conditions. In all four cases, individual data are confronted with boxplots descriptive of group performances in control neurotypical individuals, with significant decreases highlighted as follows: *** $p < 0.001$; ** $p < 0.01$, * $p < 0.05$, in single case Crawford t-test. Each boxplot is characterized by a horizontal red line referring to the median value of the distribution, a gray rectangle referring to the inter-quartile range, and whiskers referring to overall data range (within 1.5 of the inter-quartile range). Individual data are also plotted as filled grey circles. Conditions are highlighted as *Ph* (Photos), *B* (Beliefs), *E* (Emotions) and *Pa* (Pain). Finally, lesion masks of each patient are overlaid over human brain surface rendering, under the same color coding of accuracy plots. *IFG*: Inferior Frontal Gyrus; *AI*: Anterior Insula; *PI*: Posterior Insula; *DLPFC*: dorsolateral prefrontal cortex; *TPJ*: Temporo-Parietal Junction; *IOC*: Inferior Occipital Cortex; *Put*: Putamen.

In keeping with our hypothesis, we also searched for cases showing conjoint impairment in both *B* and *E*, but not in other conditions. Two patients (*C.C.* & *V.T.*) showed this pattern (*B*/*E* performance: $t_{(23)} \geq 2.66$, p (one-tailed) ≤ 0.007, $Z_{CC} \geq 2.71$ [1.83, 3.58]; *Ph*/*Pa* performance: $t_{(23)} \leq 1.06$, p (one-tailed) ≥ 0.147 – see Table 1). Critically, interaction analyses confirmed that, for

V.T., an impairment at *B* and *E* was significantly stronger than at either *Ph* and *Pa* ($t_{(23)} \geq 2.65$, $p \leq 0.014$, $Z_{DCC} \geq 2.79$ [1.86, 3.82]). For *C.C.*, significant interaction effects were found when comparing *E* with *Ph* and *Pa* ($t_{(23)} \geq 2.56$, $p \leq 0.018$, $Z_{DCC} \geq 2.69$ [1.54, 4.04]), and *B* with *Ph* ($t_{(23)} = 2.24$, $p = 0.035$, $Z_{DCC} \geq 2.35$ [1.48, 3.33]), but not *B* with *Pa* ($t_{(23)} = 1.57$, $p = 0.130$). Figure 1B displays these two patients' scores as well as their lesions, which involve the left TPJ and Putamen (for *V.T.*), as well as the right anterior insula and right DLPFC (for *C.C.*).

Table 2. Single Case Analysis.

	Gender	Age	MoCA	Hemisphere	Task			
					Ph	***B***	***E***	***Pa***
P.A.	F	78	26	R	0.83	**0.58***	0.83	0.83
M.S.	M	45	26	R	1	1	**0.67***	0.83
B.D.	M	72	23	L	0.75	0.67	**0.67***	0.83
P.E.	M	65	28	R	0.83	0.83	0.75	**0.58*****
G.F	F	72	26	R	0.75	0.75	0.83	**0.75****
P.J.	M	70	26	L	**0.33****	0.75	**0.58****	0.83
C.C	M	53	22	R	0.83	**0.50****	**0.42*****	0.83
V.T.	F	49	26	L	0.92	**0.50****	**0.58****	1
F.J.	M	53	22	R	0.67	**0.50****	0.83	**0.50*****
T.G.	M	72	29	L	0.67	0.67	**0.50****	**0.58*****
B.M.	M	74	--	R	0.58	0.92	**0.50****	**0.75***
E.G.	F	55	26	R	**0.42****	**0.50****	**0.67***	0.83
K.A.	F	62	24	L	**0.42****	**0.33*****	**0.42*****	**0.58*****
F.C.	F	77	26	L	**0.50***	**0.33*****	**0.42*****	**0.33*****

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Individual performance of 14 unilateral brain damaged patients, showing selective impairment in at least one condition, as opposed to a control population of neurotypical individuals. Bold values refer to significant effects in single case Crawford t-test. For visualization purposes, patients impaired in only one conditions are grouped on top, followed by patients impaired in multiple conditions. Individual patients are also described by Gender (F = females), Age (years), lesioned hemisphere (R = right), and their score on the Montreal Cognitive Assessment battery.

3.3 Network-based Lesion-Symptom Mapping

To consider the whole spectrum of performance and potentially weaker patterns of deficit across all brain-damaged individuals, we used lesion-symptom mapping to identify neural structures associated with social inferential impairments on our ToM task from the whole

population of 40 patients. In particular, in our study we employed a network-based lesion-symptom mapping (NLSM; Boes et al., 2015; Darby et al., 2017; Laganiere et al., 2016; Wawrzyniak et al., 2018), in which the lesion masks were "extended" to include those brain structures that are most frequently functionally connected with the damaged site based on normative connectome data (see methods). These network masks were modeled against the patients' accuracy in the three conditions of interest (*B*, *E*, *Pa*), with accuracy in the control *Ph* task, the overall lesion size, patients' gender, age, education level, MoCA score, and time post-stroke specified as nuisance predictors (see methods).

Table 3. NLSM Results.

	SIDE	***Coordinates***			$t_{(30)}$	***% Pat***	***Cluster size***	R^2
		X	***Y***	***Z***				
Effects of Beliefs *(covaring per Ph)*								
DLPFC	R	32	50	2	3.45	10	362	0.28
DMPFC	R	12	18	30	3.45	10	322	0.31
Effects of Beliefs *(covaring per E)*								
DLPFC	R	32	50	2	3.86	10	359	0.33
DMPFC	L	16	12	46	4.10	12.5	448	0.36
Effects of Emotions *(covaring per Ph)*								
Anterior Insula	L	-38	2	-2	4.83	35	271	0.41

Regions significantly associated with deficits in inferring Emotions (*E*), after having controlled for *Ph*, *Pa* or *B*. All clusters survived permutation-based correction for multiple comparisons at the cluster level (with an underlying height threshold corresponding to $p < 0.001$, uncorrected). Each region is described in terms of the local maxima's MNI coordinates, *t*-statistics, and percentage amount of patients implicated. Furthermore, we also provide information about the cluster size (number of contiguous voxels), and amount of variance explained in the overall cluster on patients' behavior (R^2, see Lorca-Puls et al., 2018).

Table 3 displays the regions implicated in the analysis. More specifically, we found a network centered in the right dorsolateral prefrontal cortex (DLPFC) and on the dorsomedial prefrontal cortex (DMPFC, see Figure 2A), that was predictive of an impairment at inferring

others' beliefs. Figure 2B displays the performance of those patients with dysfunction at the level of DMPFC local maxima, who showed drastically lower scores in the *B* task relative to other patients and controls. This was however not the case of the performance of the same patients in the other three conditions *Ph*, *E* and *Pa* (data from the right DLPFC local maxima are almost identical to those displayed in Figure 2B for DMPFC). We then also examined to which extent such effects might reflect an actual structural lesion in prefrontal cortex, and found that no patient was damaged over the peak coordinates of neither the right DLPFC nor DMPFC. In fact, the patients implicated in these effects are characterized by large and heterogeneous lesion sites, including insula, precentral, and superior parietal in the right hemisphere. To inspect the degree of specificity of this effect with respect to the other conditions of interest, we repeated the analysis, by replacing the nuisance covariate *Ph*, with either *B* or *Pa*. When accounting for *E*, the same regions in DMPFC and DLPFC were found (see Table 3), whereas no effect were observed when accounting for *Pa*. Finally, we found no role played by TPJ. We then restricted the analysis on those brain regions mapped in our previous neuroimaging study where neurotypical individuals underwent the same narratives used here (Corradi-Dell'Acqua et al., 2014). More specifically, we took the activation maps when responding to questions probing about people's beliefs, contrasted with the control photo condition (Table 2 in Corradi-Dell'Acqua et al., 2014). This maps were binarized and used as an explicit inclusive mask for analysis. Under this constrained hypothesis, we still found no suprathreshold effect.

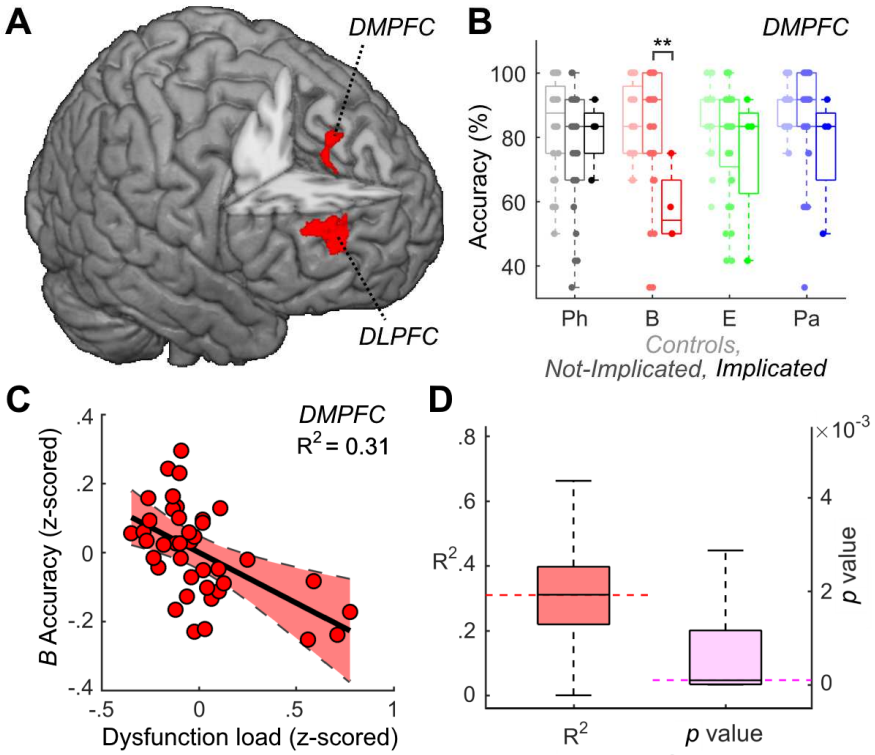


Figure 2. NLSM: beliefs effects. (**A**) Surface rendering displaying regions with dysfunctional connectivity with lesions disrupting the appraisal of others' beliefs. *DLPFC*: dorsolateral prefrontal cortex; *DMPFC*: dorsomedial prefrontal cortex. (**B**) For DMPFC, we also plotted the performance of patients whose lesion site was functionally-connected with the local maxima (dark tones), and compared it with patients who displayed different connectivity-patterns (medium tones) and with neurotypical controls (light tones). Accuracy is plotted in terms of boxplots. Individual data are also displayed as filled circles. Conditions are highlighted as *Ph* (Photos), *B* (Beliefs), *E* (Emotions), and *Pa* (Pain), color-coded in black, red, green, and blue respectively. "**" refers to independent samples *t*-test associated with $p < 0.01$ (**C**) Negative relationship between patients' ability at inferring beliefs (vertical axis) and the dysfunction in DMPFC (horizontal axis), after having accounted for all nuisance variables. R^2 is used as effect size (Lorca-Puls et al., 2018). (**D**) Boxplots describing the variability of the effect size (R^2) and the associated *p* value, as estimated through 5000 bootstrap-resamples of the original dataset.

Subsequently, we found a network centered in the left Anterior Insula (see Figure 3A), that was predictive of an impairment at inferring others' emotions. Figure 3B displays the performance of those patients with dysfunction at the level of region's local maxima, who showed drastically lower scores in the *E* task relative to other patients and controls. This was however not the case of the performance of the same patients in either *Ph*, *B* and *Pa*. As for

previous NLSM effects, the highlighted region was seldom lesioned directly, as only one patient had damage overlapping with the cluster peak, and additional five had a damage contralateral to the site. Indeed, this effect is also the result a dysfunction remotely caused by losses of connections to the damaged parietal, temporal, prefrontal and subcortical subcortical structures. We then repeated the analysis, by replacing the nuisance covariate *Ph*, with either *B* or *Pa*, but found no significant effect (at least at the employed threshold). Finally, we then restricted the analysis on those brain regions mapped in our previous neuroimaging study where neurotypical individuals responding to questions probing about people's emotions (Table 2 in Corradi-Dell'Acqua et al., 2014). No effect was found within this mask.

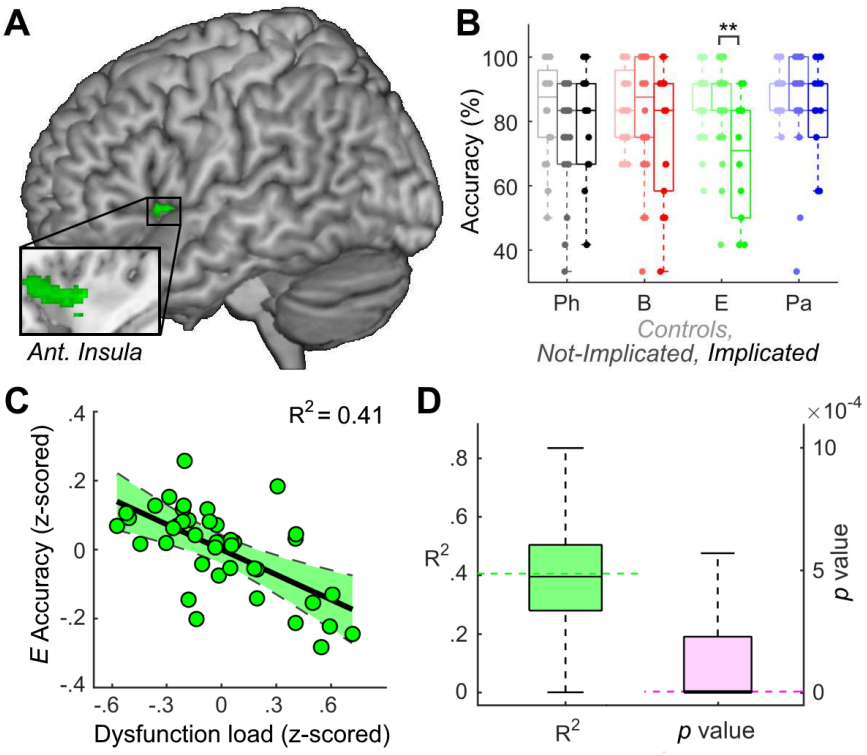


Figure 3. NLSM: emotion effects. (**A**) Surface rendering displaying a region at the level of the Anterior Insula with dysfunctional connectivity with lesions disrupting the appraisal of others' beliefs. (**B**) We also plotted the performance of patients whose lesion site was functionally-connected with the local maxima, and compared it with patients who displayed different

connectivity-patterns and with neurotypical controls. (**C**) Negative relationship between patients' ability at inferring emotions and the dysfunction in int Anterior insula, after having accounted for all nuisance variables. R^2 is used as effect size. (**D**) Boxplots describing the variability of the effect size (R^2) and the associated *p* value, as estimated through bootstrap-resamples of the original dataset.

Finally, we sought for regions predictive of an impairment at appraising other people's pain. This analysis led to no suprathreshold effects, neither when controlling for *Ph*, nor when employing *B* or *E* as nuisance covariates. Furthermore, no suprathreshold effects were observed when restricting to regions previously implicated in the appraisal of pain through the same narratives (Corradi-Dell'Acqua et al., 2014).

All NLSM findings were associated with large effect sizes, ranging between R^2 = 0.29 (DLPFC in Figure 2A) to 0.40 (Anterior Insula in Figure 3; see also Table 3). Supplementary Tables 1 and 2 report the outcome of the bootstrap-stimulation run to assess the reliability of the estimated effects. Overall, the R^2 and the associated *p* values were almost identical to the median from all bootstrap resamplings (see Figures 2D and 3D for a graphical representation). Even when focusing only the resampled R^2 associated with significant effects, the median values were quite in line with those of the original sample. This is due to the fact that the majority of the resamplings were associated with a significant effect (~ 92% at $\alpha = 0.05$; ~ 70% at $\alpha = 0.001$). Overall, the effect sizes observed in the present study appear to be reliable and replicable.

4. Discussion

Our results reveal a striking dissociation in the information represented within the human brain networks mediating emotional processing and theory of mind (ToM). Combining neuropsychological approach with lesion mapping and normative brain connectome data, we provide novel evidence that impairments in the appraisal of other people's emotions and beliefs underlie dissociable lesional correlates, with the former linked with dysfunction at a network centered on the anterior insula, and the latter specifically associated with a network involving the lateral (DLPFC) and medial (DMPFC) portions of the dorsal prefrontal cortex. No evidence was found for networks common between these two social inferential abilities.

4.1 Neural systems for understanding others' beliefs

The neural correlates of cognitive ToM abilities have been systematically investigated in previous neuroimaging studies, which implicated a network comprising the temporo-parietal junction, middle temporal gyrus, precuneus, as well as lateral and medial prefrontal cortices (Bzdok et al., 2012; Krall et al., 2015; Molenberghs et al., 2016; Schurz et al., 2017; Van Overwalle, 2009; van Veluw & Chance, 2014). Furthermore, the understanding others' cognitive states can be affected following damage/interference to many among these regions (Balaban et al., 2016; Biervoye et al., 2016; Campanella et al., 2014; Domínguez D et al., 2019; Happé et al., 1999; Kalbe et al., 2010; Krause et al., 2012; Le Bouc et al., 2012; Leigh et al., 2013; Leopold et al., 2012; Mai et al., 2016; Muller et al., 2010; Rowe et al., 2001; Samson et al., 2004, 2005; Shamay-Tsoory et al., 2005, 2006, 2010; Shamay-Tsoory & Aharon-Peretz, 2007; Stuss et al., 2001; Yeh & Tsai, 2014; Young et al., 2010). However, this network might not be homogeneous

in its function, and different parts most likely contribute to appraisal of others' cognitive states through distinct and dissociable subprocesses.

Indeed, in the typical brain, the temporo-parietal cortex is also active when people appraise others' emotions (Corradi-Dell'Acqua et al., 2014; Hynes et al., 2006; Schlaffke et al., 2015; Sebastian et al., 2012; Völlm et al., 2006), with very similar neural activity patterns than those observed for the inference of cognitive states (Corradi-Dell'Acqua et al., 2014). Furthermore, inhibition of this region's activity through cathodal electrical stimulation leads to joint impairment in both cognitive ToM and in the assessment of others' affect (Mai et al., 2016). In this perspective, it has been argued that the temporo-parietal cortex may represent a key structure for a mentalistic strategy during emotion inference, according to which representations of people's ongoing beliefs/thoughts/goals play a crucial role for also for the inference of affective states. This is consistent with appraisal theories of emotions (Scherer, 2009), who propose that affective experience is strongly determined by the contextual evaluation of events, including beliefs about their implications for one's goals and how they can be coped with (e.g., sadness is often the consequence to the believe that there is no way to resolve a bad situation). Our data offer little evidence on the role of temporo-parietal cortex in our paradigm, presumably due to the fact that our lesion and network maps do not extend too frequently in this region (see Supplementary Figure 1). The only insight comes from, single-case analysis, who showed how deficit in appraising both beliefs and emotion can result from a variety of impairments, sometimes implicating directly the temporo-parietal cortex (patient *VT*, Figure 1B), but other times being associated with different regions (patient *CC*). However, joint deficits in single cases should be considered with caution, as it is unclear whether they result

from damage to a unique neural structure involved in two processes, or whether the lesion affects multiple sites each subserving one specific function.

Instead, in our study, NLSM analysis revealed that dysfunctions at the level of lateral (DLPFC) and medial (DMPFC) portions of the prefrontal cortex are associated with difficulties in inferring beliefs (Figure 2), an effect that did not generalize to affect attribution. This evidence converges with, but also extends, previous neuroimaging investigations showing how judgments of cognitive and affective states lead to dissociated responses in neighboring portions of DMPFC (Corradi-Dell'Acqua et al., 2014), but also how damage/interference to dorsal prefrontal structure lead to selective impairments in the assessment of cognitive states (Bejanin et al., 2017; Kalbe et al., 2010; Shamay-Tsoory & Aharon-Peretz, 2007). It is unclear which subprocess underlying ToM abilites characterizes the dorsal prefrontal cortex, although previous studies suggest it might relate to a mechanism for the inhibition of one's own point of view. Indeed, one element common to many paradigms testing ToM abilities (including the current one) is the awareness that the representation of the world of one character is different from that of the participant himself/herself: e.g., stories where someone believes that an object is in the wrong location are usually framed in such way that participants are aware where the object truly is. Hence, by asking to explicitly assess others' beliefs about an event, ToM paradigms are typically forcing individuals to inhibit their own perspective about the same event. This is not necessarily the case in implicit paradigms, where a representation of others' beliefs might influence participants' performance without being the object of the task. Indeed, whereas impairments of the temporo-parietal cortex could lead to difficulties in mental states attribution under both explicit (Krall et al., 2015; Samson et al., 2004) and implicit settings

(Biervoye et al., 2016; Young et al., 2010), the right prefrontal cortex seems associated with more selective impairments when the state to be predicted explicitly contrasts with participants' own (Samson et al., 2005).

This model suggesting that prefrontal contributions to ToM tasks might be limited to explicit (but not implicit) mental state attribution (Samson et al., 2005), could also be applied to the inference of affective *vs*. cognitive states in the present study, ultimately explaining our results. Indeed, if inferences about others' emotions are partly grounded on representations of their beliefs/thoughts/goals, then condition *E* should be considered as an implicit ToM paradigm, where cognitive states contribute to the judgment without being themselves the object of the task (participants are asked to choose between two target emotions). Hence, although matched for difficulty, conditions *B* and *E* differ in the degree to which participants are overtly asked to respond about the point of view of the character with respect to their own. This might explain why paradigms comparing the inference of cognitive and affective states often report dissociated responses in the prefrontal cortex (see Figure 2; see also Bejanin et al., 2017; Kalbe et al., 2010; Shamay-Tsoory & Aharon-Peretz, 2007), as only cognitive ToM tasks require overt inhibition of one's own perspective in favor of a representation of others' view.

4.2 Neural systems for understanding others' emotions and pain

Finally, both our single-case and network analyses revealed lesion patterns leading to impairments in the inference of others' emotions and pain. In particular, two single cases with damage to the insula, extending to the inferior frontal gyrus, showed difficulties in inferring emotions or pain, an effect that could not be explained in terms of more general mentalistic abilities given that the same individuals showed spared performance in the assessment of

beliefs (Figures 1A). In particular, one case with selective damage to the posterior portion of the insula showed a specific impairment at inferring pain, but not other conditions (Figure 1A, case *P.E.*). The other case, with damage extending to the anterior insular and inferior frontal gyrus, showed a selective impairment at inferring emotions, but not beliefs (case *M.S.*). This analysis was complemented by a group-wise NLSM who showed that a network centered on the left anterior insula underlay impairments in the evaluation of emotions (Figure 3). Unfortunately, the same result was not observed (at least under correction for multiple comparisons) when accounting for abilities to judge other people's beliefs, although it should be underscored that none of the patients damaged in this network showed significant impairments in the assessment of cognitive states (Figure 3B).

The idea of a ToM-independent process for affect attribution in the anterior insula fits well seminal models pointing to parallel and dissociable pathways for the understanding of people's emotions. On top of a "cognitive" pathway, grounded on the same temporo-prefrontal processes underlying ToM, some authors proposed an "affective" pathway mediating a mechanism of affective resonance in which others' behavioral/physiological reactions (smiles, tears, shivers, etc.) are simulated on oneself (Shamay-Tsoory, 2011; Stietz et al., 2019). The neural underpinnings of such mechanism have been extensively investigated in neuroimaging studies, implicating a network comprising the anterior insula, supramarginal/postcentral gyrii, and cingulate cortex (Bzdok et al., 2012; Del Casale et al., 2017; Ding et al., 2019; Fan et al., 2011; Jauniaux et al., 2019; Lamm et al., 2011; Timmers et al., 2018, as meta-analyses). Formal comparisons between tasks engaging these pathways have revealed clear-cut segregated neural responses (both at the meta-analytic level and in single studies, Bzdok et al., 2012;

Kanske et al., 2016), with anterior insula implicated in affective resonance exerting inhibitory effects on ToM-related activity in temporo-parietal cortex (Kanske et al., 2016). Despite this wealth of neuroimaging evidence, very little data exist concerning whether the same dissociations could be observed following brain damage. To our knowledge, only one study described a dissociation between anterior insula and medial prefrontal cortex which led to selective impairments in questionnaire scores testing affective empathy and perspective taking, respectively (Shamay-Tsoory et al., 2009). Our data extend previous research by confirming a functional segregation between networks cantered in insular and prefrontal cortex, with the former associated with deficits in emotional processing and the latter selectively implicated in cognitive ToM.

It is less clear to which extent the insular cortex processes different kinds of emotions and affective states in selective fashion, given that all deficits in emotion processing observed in this study (either at the single-case or network level) did not dissociate from deficits in appraising pain. As none of the patients damaged in this network showed significant impairments in *Pa* condition (Figure 3B), it is plausible to assume that the effects observed in Figure 3 do not generalize to pain. However, in a previous research, Gu and colleagues (2012) described three cases with selective damage to the anterior insula who showed difficulties at appraising pain. The neuroscience community debated extensively as to whether the insular cortex processes one's and others' affect through state-specific or state-independent neural representations. For instance, neuroimaging studies showed that the insula processes a wide range of affective events in others, such as pain (Corradi-Dell'Acqua et al., 2011, 2016; Lamm et al., 2011), disgust (Corradi-Dell'Acqua et al., 2016; Jabbi et al., 2007; Wicker et al., 2003), and

even happiness (Hennenlotter et al., 2005; see also Ding et al., 2019; Timmers et al., 2018 as meta-analyses). In some cases, the most anterior portion of insula seems to encode supra-ordinal dimensions of affect for self and others, common between pain and aversive pictures (Corradi-Dell'Acqua et al., 2011), or between pain, disgust, and unfairness (Corradi-Dell'Acqua et al., 2016), or between empathetic reactions to a wide range of states (Timmers et al., 2018). Instead, the middle-posterior portion of the insula seems mainly involved in the appraisal of pain, as shown in paradigms employing text-based stories (Bruneau et al., 2012, 2013; Corradi-Dell'Acqua et al., 2014) or pictures of injured hands (Corradi-Dell'Acqua et al., 2011). This posterior-to-anterior gradient in insula function described by previous neuroimaging research fits well our results, according to which one patient with selective damage to the posterior section displayed specific difficulties in appraising pain (but not emotions and beliefs), whereas dysfunctions related to the more anterior portions underlie impairment in understating other emotional states. However, as no selective deficit for pain was corroborated at the group-level, caution should be advised for interpreting this anatomo-functional association.

4.3 Limitations of the study and conclusions

In keeping with a long tradition, the present study employed a verbal ToM task. This represents only one of the possible ToM paradigms, which impacts on neural structures partially dissociated from those of other non-verbal tasks (see Mar, 2011; Molenberghs et al., 2016, as a meta-analyses). Verbal ToM paradigms weight heavily on individual's linguistic, attentional and mnemonic abilities, thus opening the possibility that low performance scores might underlie a deficit to these processes. Keep in mind, however, that impairments at the level of language, memory or attention are expected to impact all conditions alike, that is beliefs, emotions, pain,

but also the control photo condition (e.g., left-damaged patients are more impaired in the task, regardless of the condition, presumably due to their low linguistic proficiency). However, the effects highlighted in the present study are all described in terms of dissociations, with impairments in one condition of interest associated with spared performance in a control. For this reason, we feel our results safe from any confound idiosyncratic to the verbal nature of the paradigm employed.

Although all lesions were mapped during patients' acute/sub-acute phase of the stroke, the behavior was collected at a much later time, breaching into chronic phase. This might potentially complicate the interpretation of the brain-behavior association, as chronic patients might have had the time to recuperate their deficits due to brain organization (de Haan & Karnath, 2018). We feel unlikely that this concern might apply to our study, as ToM deficits often persist across many months/years following brain damage (Balaban et al., 2016; Happé et al., 1999; Yeh & Tsai, 2014), and the performance of our patients did not change significantly as function of the time following the stroke (see Supplementary Information for more details). However, we nevertheless minimized this potential confound by including the days post-stroke as a nuisance covariate in the lesion analysis. In this way, all NLSM effects should be interpreted as occurring *regardless* of any linear effect of brain reorganization and patients' recovery.

A recent simulation of lesion-mapping analyses found that small samples (e.g., N = 30) lead to low replicability, with only inflated effect sizes reaching significance (Lorca-Puls et al., 2018). We feel unlikely that this might be the case also of our study. A bootstrap-resampling approach similar to that of Lorca-Puls et al. (2018) suggests that the effects observed are fairly stable in magnitude, with the majority of the simulations leading to a significant *p* values (see

Supplementary Information for more details). However, we cannot exclude that effects of smaller magnitude might have just gone undetected, and that a bigger patient cohort could have allowed a better sensitivity.

We run NLSM as an innovative way to account for lesion heterogeneity (Boes et al., 2015; Darby et al., 2017; Laganiere et al., 2016; Wawrzyniak et al., 2018). Yet, caution should be advised in assuming that results from an independent resting-state dataset is a good estimate of patients' impairment at the network level. First, networks maps were estimated using a seed-based approach, which works effectively on small homogeneous lesions. However, when the damage is extended and involves multiple brain structures, some information loss is expected, as connectivity-patterns from different areas might cancel each other out when averaged within the same seed. In this view, modeling the lesion size as covariate of no interest is foremost important for the NLSM. Second, networks maps describe only information from the grey matter, and do not take into account damage in white matter tracts which can also affect the interaction between brain areas in an unforeseen way. Finally, connectivity data collected under rest do not take into account task-specific interactions between regions. Although resting state patterns highly resemble task-positive co-activation maps (Smith et al., 2009), a better estimate of stroke-induced connectivity impairments needs to be explored in future studies by measuring neural activity and functional connectivity during the execution of the task.

Keeping these limitations aside, the approach used here allowed us to efficiently map impairments in social inferential abilities in stroke patients, and unveil functional dissociations within networks underlying cognitive and emotional ToM processes which are strongly in line

with those of previous neuroimaging investigations. More specifically, modeling patients' deficits at the network level (NLSM) proofed an efficient way to reconcile an heterogeneous set of brain lesions with similar behavioural impairments, reflecting the fact that complex social behavior emerges from the interaction of a heterogeneous network. In this view, this study underscores the importance of overcoming standard region-based approaches when investigating patients' deficits in understanding others.

5. Acknowledgments

CCD is supported by the Swiss National Science Foundation (SNSF), grant numbers PP00O1_157424 and PP00O1_157424, RR is supported by the SNSF grant n. PMPDP3_171376, whereas PV is supported by SNSF grant n. 32003B_138413. We like to thank Marina Almató for her assistance in collecting data from the last controls. De-identified data and code for the experiment/analysis are available under the Open Science Framework: https://osf.io/v94pf/

6. References

Adenzato, M., Cavallo, M., & Enrici, I. (2010). Theory of mind ability in the behavioural variant of frontotemporal dementia: An analysis of the neural, cognitive, and social levels. *Neuropsychologia*, *48*(1), 2–12. https://doi.org/10.1016/j.neuropsychologia.2009.08.001

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews. Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Balaban, N., Friedmann, N., & Ziv, M. (2016). Theory of mind impairment after right-hemisphere damage. *Aphasiology*, *30*(12), 1399–1423. https://doi.org/10.1080/02687038.2015.1137275

Bartolomeo, P., Thiebaut de Schotten, M., & Doricchi, F. (2007). Left unilateral neglect as a disconnection syndrome. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(11), 2479–2490. https://doi.org/10.1093/cercor/bhl181

Bejanin, A., Chételat, G., Laisney, M., Pélerin, A., Landeau, B., Merck, C., Belliard, S., Sayette, V. de L., Eustache, F., & Desgranges, B. (2017). Distinct neural substrates of affective and cognitive theory of mind impairment in semantic dementia. *Social Neuroscience*, *12*(3), 287–302. https://doi.org/10.1080/17470919.2016.1168314

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23. https://doi.org/10.1146/annurev-neuro-062111-150536

Biervoye, A., Dricot, L., Ivanoiu, A., & Samson, D. (2016). Impaired spontaneous belief inference following acquired damage to the left posterior temporoparietal junction. *Social Cognitive and Affective Neuroscience*, *11*(10), 1513–1520. https://doi.org/10.1093/scan/nsw076

Binder, J. R., Pillay, S. B., Humphries, C. J., Gross, W. L., Graves, W. W., & Book, D. S. (2016). Surface errors without semantic impairment in acquired dyslexia: A voxel-based lesion–symptom mapping study. *Brain*, *139*(5), 1517–1526. https://doi.org/10.1093/brain/aww029

Bodden, M. E., Kübler, D., Knake, S., Menzler, K., Heverhagen, J. T., Sommer, J., Kalbe, E., Krach, S., & Dodel, R. (2013). Comparing the neural correlates of affective and cognitive theory of mind using fMRI: Involvement of the basal ganglia in affective theory of mind.

Advances in Cognitive Psychology, *9*(1), 32–43. https://doi.org/10.2478/v10053-008-0129-6

Boes, A. D., Prasad, S., Liu, H., Liu, Q., Pascual-Leone, A., Caviness, V. S., & Fox, M. D. (2015). Network localization of neurological symptoms from focal brain lesions. *Brain*, *138*(10), 3061–3075. https://doi.org/10.1093/brain/awv228

Bora, E., Velakoulis, D., & Walterfang, M. (2016). Social cognition in Huntington's disease: A meta-analysis. *Behavioural Brain Research*, *297*, 131–140. https://doi.org/10.1016/j.bbr.2015.10.001

Bora, E., Walterfang, M., & Velakoulis, D. (2015). Theory of mind in behavioural-variant frontotemporal dementia and Alzheimer's disease: A meta-analysis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *86*(7), 714–719. https://doi.org/10.1136/jnnp-2014-309445

Bottiroli, S., Cavallini, E., Ceccato, I., Vecchi, T., & Lecce, S. (2016). Theory of Mind in aging: Comparing cognitive and affective components in the faux pas test. *Archives of Gerontology and Geriatrics*, *62*, 152–162. https://doi.org/10.1016/j.archger.2015.09.009

Bruneau, E., Dufour, N., & Saxe, R. (2013). How We Know It Hurts: Item Analysis of Written Narratives Reveals Distinct Neural Responses to Others' Physical Pain and Emotional Suffering. *PLOS ONE*, *8*(4), e63085. https://doi.org/10.1371/journal.pone.0063085

Bruneau, E., Pluta, A., & Saxe, R. (2012). Distinct roles of the "Shared Pain" and "Theory of Mind" networks in processing others' emotional suffering. *Neuropsychologia*, *50*(2), 219–231. https://doi.org/10.1016/j.neuropsychologia.2011.11.008

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A., Langner, R., & Eickhoff, S. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, *217*(4), 783–796. https://doi.org/10.1007/s00429-012-0380-y

Campanella, F., Shallice, T., Ius, T., Fabbro, F., & Skrap, M. (2014). Impact of brain tumour location on emotion and personality: A voxel-based lesion–symptom mapping study on mentalization processes. *Brain*, *137*(9), 2532–2545. https://doi.org/10.1093/brain/awu183

Charlton, R. A., Barrick, T. R., Markus, H. S., & Morris, R. G. (2009). Theory of mind associations with other cognitive functions and brain imaging in normal aging. *Psychology and Aging*, *24*(2), 338–348. https://doi.org/10.1037/a0015225

Clas, P., Groeschel, S., & Wilke, M. (2012). A semi-automatic algorithm for determining the demyelination load in metachromatic leukodystrophy. *Academic Radiology*, *19*(1), 26–34. https://doi.org/10.1016/j.acra.2011.09.008

Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2011). Felt and Seen Pain Evoke the Same Local Patterns of Cortical Activity in Insular and Cingulate Cortex. *The Journal of Neuroscience*, *31*(49), 17996–18006. https://doi.org/10.1523/JNEUROSCI.2686-11.2011

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*(8), 1175–1184. https://doi.org/10.1093/scan/nst097

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., & Singer, T. (2016). Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nature Communications*, *7*, 10904. https://doi.org/10.1038/ncomms10904

Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using monte carlo simulations and revised tests for dissociations. *Neuropsychology*, *19*(3), 318–331. https://doi.org/10.1037/0894-4105.19.3.318

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*(3), 245–260. https://doi.org/10.1080/02643294.2010.513967

Crawford, J. R., & Howell, D. C. (1998). Comparing an Individual's Test Score Against Norms Derived from Small Samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. https://doi.org/10.1076/clin.12.4.482.7241

Darby, R. R., Laganiere, S., Pascual-Leone, A., Prasad, S., & Fox, M. D. (2017). Finding the imposter: Brain connectivity of lesions causing delusional misidentifications. *Brain*, *140*(2), 497–507. https://doi.org/10.1093/brain/aww288

de Haan, B., Clas, P., Juenger, H., Wilke, M., & Karnath, H.-O. (2015). Fast semi-automated lesion demarcation in stroke. *NeuroImage. Clinical*, *9*, 69–74. https://doi.org/10.1016/j.nicl.2015.06.013

de Haan, B., & Karnath, H.-O. (2018). A hitchhiker's guide to lesion-behaviour mapping. *Neuropsychologia*, *115*, 5–16. https://doi.org/10.1016/j.neuropsychologia.2017.10.021

Del Casale, A., Kotzalidis, G. D., Rapinesi, C., Janiri, D., Aragona, M., Puzella, A., Spinazzola, E., Maggiora, M., Giuseppin, G., Tamorri, S. M., Vento, A. E., Ferracuti, S., Sani, G., Pompili, M., & Girardi, P. (2017). Neural functional correlates of empathic face processing. *Neuroscience Letters*, *655*, 68–75. https://doi.org/10.1016/j.neulet.2017.06.058

Ding, R., Ren, J., Li, S., Zhu, X., Zhang, K., & Luo, W. (2019). Domain-general and domain-preferential neural correlates underlying empathy towards physical pain, emotional situation and emotional faces: An ALE meta-analysis. *Neuropsychologia*, 107286. https://doi.org/10.1016/j.neuropsychologia.2019.107286

Domínguez D, J. F., Nott, Z., Horne, K., Prangley, T., Adams, A. G., Henry, J. D., & Molenberghs, P. (2019). Structural and functional brain correlates of theory of mind impairment post-stroke. *Cortex*, *121*, 427–442. https://doi.org/10.1016/j.cortex.2019.09.017

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 903–911. https://doi.org/10.1016/j.neubiorev.2010.10.009

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*, 21–45. https://doi.org/10.1146/annurev.psych.50.1.21

Freedman, M., Binns, M. A., Black, S. E., Murphy, C., & Stuss, D. T. (2013). Theory of mind and recognition of facial emotion in dementia: Challenge to current concepts. *Alzheimer Disease and Associated Disorders*, *27*(1), 56–61. https://doi.org/10.1097/WAD.0b013e31824ea5db

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition*, *101*(1), 129–152. https://doi.org/10.1016/j.cognition.2005.05.007

Goldman, A., & de Vignemont, F. (2009). Is social cognition embodied? *Trends in Cognitive Sciences*, *13*(4), 154–159. https://doi.org/10.1016/j.tics.2009.01.007

Gu, X., Gao, Z., Wang, X., Liu, X., Knight, R. T., Hof, P. R., & Fan, J. (2012). Anterior insular cortex is necessary for empathetic pain perception. *Brain: A Journal of Neurology*, *135*(Pt 9), 2726–2735. https://doi.org/10.1093/brain/aws199

Happé, F., Brownell, H., & Winner, E. (1999). Acquired "theory of mind" impairments following stroke. *Cognition*, *70*(3), 211–240.

He, B. J., Snyder, A. Z., Vincent, J. L., Epstein, A., Shulman, G. L., & Corbetta, M. (2007). Breakdown of functional connectivity in frontoparietal networks underlies behavioral deficits in spatial neglect. *Neuron*, *53*(6), 905–918. https://doi.org/10.1016/j.neuron.2007.02.013

Hennenlotter, A., Schroeder, U., Erhard, P., Castrop, F., Haslinger, B., Stoecker, D., Lange, K. W., & Ceballos-Baumann, A. O. (2005). A common neural basis for receptive and expressive communication of pleasant facial affect. *NeuroImage*, *26*(2), 581–591. https://doi.org/16/j.neuroimage.2005.01.057

Henry, J. D., Phillips, L. H., & von Hippel, C. (2014). A meta-analytic review of theory of mind difficulties in behavioural-variant frontotemporal dementia. *Neuropsychologia*, *56*, 53–62. https://doi.org/10.1016/j.neuropsychologia.2013.12.024

Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia*, *44*(3), 374–383. https://doi.org/10.1016/j.neuropsychologia.2005.06.011

Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage*, *34*(4), 1744–1753. https://doi.org/10.1016/j.neuroimage.2006.10.032

Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage*, *126*, 39–48. https://doi.org/10.1016/j.neuroimage.2015.11.025

Jauniaux, J., Khatibi, A., Rainville, P., & Jackson, P. L. (2019). A meta-analysis of neuroimaging studies on pain empathy: Investigating the role of visual information and observers' perspective. *Social Cognitive and Affective Neuroscience*, *14*(8), 789–813. https://doi.org/10.1093/scan/nsz055

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., Brand, M., Shamay-Tsoory, S., Onur, O. A., & Kessler, J. (2010). Dissociating cognitive from affective theory of mind: A TMS study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *46*, 769–780.

Kanske, P., Böckler, A., Trautwein, F.-M., Lesemann, P., H, F., & Singer, T. (2016). Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. *Social Cognitive and Affective Neuroscience*, *11*(9), 1383–1392. https://doi.org/10.1093/scan/nsw052

Kipps, C. M., & Hodges, J. R. (2006). Theory of mind in frontotemporal dementia. *Social Neuroscience*, *1*(3–4), 235–244. https://doi.org/10.1080/17470910600989847

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., Fink, G. R., & 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*, *220*(2), 587–604. https://doi.org/10.1007/s00429-014-0803-z

Krause, L., Enticott, P. G., Zangen, A., & Fitzgerald, P. B. (2012). The role of medial prefrontal cortex in theory of mind: A deep rTMS study. *Behavioural Brain Research*, *228*(1), 87–90. https://doi.org/10.1016/j.bbr.2011.11.037

Kumfor, F., Hazelton, J. L., De Winter, F.-L., de Langavant, L. C., & Van den Stock, J. (2017). Clinical Studies of Social Neuroscience: A Lesion Model Approach. In A. Ibáñez, L. Sedeño, & A. M. García (Eds.), *Neuroscience and Social Science: The Missing Link* (pp. 255–296). Springer International Publishing. https://doi.org/10.1007/978-3-319-68421-5_12

Laganiere, S., Boes, A. D., & Fox, M. D. (2016). Network localization of hemichorea-hemiballismus. *Neurology*, *86*(23), 2187–2195.

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502.

Le Bouc, R., Lenfant, P., Delbeuck, X., Ravasi, L., Lebert, F., Semah, F., & Pasquier, F. (2012). My belief or yours? Differential theory of mind deficits in frontotemporal dementia and Alzheimer's disease. *Brain*, *135*(10), 3026–3038. https://doi.org/10.1093/brain/aws237

Leigh, R., Oishi, K., Hsu, J., Lindquist, M., Gottesman, R. F., Jarso, S., Crainiceanu, C., Mori, S., & Hillis, A. E. (2013). Acute lesions that impair affective empathy. *Brain*, *136*(8), 2539–2549. https://doi.org/10.1093/brain/awt177

Leopold, A., Krueger, F., Monte, O. dal, Pardini, M., Pulaski, S. J., Solomon, J., & Grafman, J. (2012). Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Social Cognitive and Affective Neuroscience*, *7*(8), 871–880. https://doi.org/10.1093/scan/nsr071

Lorca-Puls, D. L., Gajardo-Vidal, A., White, J., Seghier, M. L., Leff, A. P., Green, D. W., Crinion, J. T., Ludersdorfer, P., Hope, T. M. H., Bowman, H., & Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. *Neuropsychologia*, *115*, 101–111. https://doi.org/10.1016/j.neuropsychologia.2018.03.014

Lough, S., Kipps, C. M., Treise, C., Watson, P., Blair, J. R., & Hodges, J. R. (2006). Social reasoning, emotion and empathy in frontotemporal dementia. *Neuropsychologia*, *44*(6), 950–958. https://doi.org/10.1016/j.neuropsychologia.2005.08.009

Mai, X., Zhang, W., Hu, X., Zhen, Z., Xu, Z., Zhang, J., & Liu, C. (2016). Using tDCS to Explore the Role of the Right Temporo-Parietal Junction in Theory of Mind and Cognitive Empathy. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00380

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*, 103–134. https://doi.org/10.1146/annurev-psych-120709-145406

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O. K., Coslett, H. B., & Schwartz, M. F. (2015). Neural organization of spoken language revealed by lesion–symptom mapping. *Nature Communications*, *6*(1), 1–9. https://doi.org/10.1038/ncomms7762

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, *65*, 276–291. https://doi.org/10.1016/j.neubiorev.2016.03.020

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*, *46*(9), 1088–1099. https://doi.org/10.1016/j.cortex.2009.08.014

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. *J. Neurosci.*, *30*(30), 10127–10134. https://doi.org/10.1523/JNEUROSCI.2161-10.2010

Pillay, S. B., Binder, J. R., Humphries, C., Gross, W. L., & Book, D. S. (2017). Lesion localization of speech comprehension deficits in chronic aphasia. *Neurology*, *88*(10), 970–975. https://doi.org/10.1212/WNL.0000000000003683

Pillay, S. B., Stengel, B. C., Humphries, C., Book, D. S., & Binder, J. R. (2014). Cerebral localization of impaired phonological retrieval during rhyme judgment. *Annals of Neurology*, *76*(5), 738–746. https://doi.org/10.1002/ana.24266

Poletti, M., Enrici, I., & Adenzato, M. (2012). Cognitive and affective Theory of Mind in neurodegenerative diseases: Neuropsychological, neuroanatomical and neurochemical levels. *Neuroscience & Biobehavioral Reviews*, *36*(9), 2147–2164. https://doi.org/10.1016/j.neubiorev.2012.07.004

Qiao-Tasserit, E., Corradi-Dell'Acqua, C., & Vuilleumier, P. (2018). The good, the bad, and the suffering. Transient emotional episodes modulate the neural circuits of pain and empathy. *Neuropsychologia*, *116*, 99–116. https://doi.org/10.1016/j.neuropsychologia.2017.12.027

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). `Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain*, *124*(3), 600–616. https://doi.org/10.1093/brain/124.3.600

Ruff, C. C., Driver, J., & Bestmann, S. (2009). Combining TMS and fMRI: From 'virtual lesions' to functional-network accounts of cognition. *Cortex*, *45*(9), 1043–1049. https://doi.org/10.1016/j.cortex.2008.10.012

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nat Neurosci*, *7*(5), 499–500. https://doi.org/10.1038/nn1223

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain: A Journal of Neurology*, *128*(Pt 5), 1102–1111. https://doi.org/10.1093/brain/awh464

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124. https://doi.org/10.1146/annurev.psych.55.090902.142044

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., Schmidt-Wilcke, T., & Brüne, M. (2015). Shared and nonshared neural networks of cognitive and affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Human Brain Mapping*, *36*(1), 29–39. https://doi.org/10.1002/hbm.22610

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping*, *38*(9), 4788–4805. https://doi.org/10.1002/hbm.23675

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, *7*(1), 53–63. https://doi.org/10.1093/scan/nsr023

Shamay-Tsoory, S. G. (2011). The Neural Bases for Empathy. *The Neuroscientist*, *17*(1), 18–24. https://doi.org/10.1177/1073858410379268

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia*, *45*(13), 3054–3067. https://doi.org/10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617–627. https://doi.org/10.1093/brain/awn279

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, *46*(5), 668–677. https://doi.org/10.1016/j.cortex.2009.04.008

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, *1*(3–4), 149–166. https://doi.org/10.1080/17470910600985589

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *18*(1), 55–67.

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right Supramarginal Gyrus Is Crucial to Overcome Emotional Egocentricity Bias in Social Judgments. *Journal of Neuroscience*, *33*(39), 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Stietz, J., Jauk, E., Krach, S., & Kanske, P. (2019). Dissociating Empathy From Perspective-Taking: Evidence From Intra- and Inter-Individual Differences Research. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00126

Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for `theory of mind'. *Brain*, *124*(2), 279–286. https://doi.org/10.1093/brain/124.2.279

Timmers, I., Park, A. L., Fischer, M. D., Kronman, C. A., Heathcote, L. C., Hernandez, J. M., & Simons, L. E. (2018). Is Empathy for Pain Unique in Its Neural Correlates? A Meta-Analysis of Neuroimaging Studies of Empathy. *Frontiers in Behavioral Neuroscience*, *12*. https://doi.org/10.3389/fnbeh.2018.00289

Torralva, T., Gleichgerrcht, E., Torres Ardila, M. J., Roca, M., & Manes, F. F. (2015). Differential Cognitive and Affective Theory of Mind Abilities at Mild and Moderate Stages of Behavioral Variant Frontotemporal Dementia. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *28*(2), 63–70. https://doi.org/10.1097/WNN.0000000000000053

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: An ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging and Behavior*, *8*(1), 24–38. https://doi.org/10.1007/s11682-013-9266-8

Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J. F. W., & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. *NeuroImage*, *29*(1), 90–98. https://doi.org/10.1016/j.neuroimage.2005.07.022

Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J., & Dolan, R. J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience*, *7*(11), 1271–1278. https://doi.org/10.1038/nn1341

Vuilleumier, P., Schwartz, S., Verdon, V., Maravita, A., Hutton, C., Husain, M., & Driver, J. (2008). Abnormal Attentional Modulation of Retinotopic Cortex in Parietal Patients with Spatial Neglect. *Current Biology*, *18*(19), 1525–1529. https://doi.org/10.1016/j.cub.2008.08.072

Wang, Y., & Olson, I. R. (2018). The Original Social Network: White Matter and Social Cognition. *Trends in Cognitive Sciences*, *22*(6), 504–516. https://doi.org/10.1016/j.tics.2018.03.005

Wang, Z., & Su, Y. (2013). Age-related differences in the performance of theory of mind in older adults: A dissociation of cognitive and affective components. *Psychology and Aging*, *28*(1), 284–291. https://doi.org/10.1037/a0030876

Wawrzyniak, M., Klingbeil, J., Zeller, D., Saur, D., & Classen, J. (2018). The neuronal network involved in self-attribution of an artificial hand: A lesion network-symptom-mapping study. *NeuroImage*, *166*, 317–324. https://doi.org/10.1016/j.neuroimage.2017.11.011

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron*, *40*(3), 655–664.

Yeh, Z.-T., & Tsai, C.-F. (2014). Impairment on theory of mind and empathy in patients with stroke. *Psychiatry and Clinical Neurosciences*, *68*(8), 612–620. https://doi.org/10.1111/pcn.12173

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(15), 6753–6758. https://doi.org/10.1073/pnas.0914826107