OXFORD

Social Cognitive and Affective Neuroscience, 2022, **17**, 732–743

DOI: https://doi.org/10.1093/scan/nsab138
Advance Access Publication Date: 27 January 2022
Original Manuscript

Effective connectivity of the human mirror neuron system during social cognition

Sadjad Sadeghi,[1,2] Stephanie N. L. Schmidt,[3] Daniela Mier,[3,*] and Joachim Hass[1,4,5,*]

[1]Department of Theoretical Neuroscience, Central Institute of Mental Health, Mannheim 68159, Germany
[2]Department of Physics and Astronomy, Heidelberg University, Heidelberg 69120, Germany
[3]Department of Psychology, University of Konstanz, Konstanz 78464, Germany
[4]Bernstein Center for Computational Neuroscience (BCCN) Heidelberg, Heidelberg/Mannheim 68159, Germany
[5]Faculty of Applied Psychology, SRH University of Applied Sciences Heidelberg, Heidelberg 69123, Germany
Correspondence should be addressed to Joachim Hass, Faculty of Applied Psychology, SRH University of Applied Sciences, Maria-Probst-Strasse 3A, Heidelberg 69123, Germany. E-mail: joachim.hass@srh.de.
[*]These authors contributed equally.

Abstract

The human mirror neuron system (MNS) can be considered the neural basis of social cognition. Identifying the global network structure of this system can provide significant progress in the field. In this study, we use dynamic causal modeling (DCM) to determine the effective connectivity between central regions of the MNS for the first time during different social cognition tasks. Sixty-seven healthy participants completed fMRI scanning while performing social cognition tasks, including imitation, empathy and theory of mind. Superior temporal sulcus (STS), inferior parietal lobule (IPL) and Brodmann area 44 (BA44) formed the regions of interest for DCM. Varying connectivity patterns, 540 models were built and fitted for each participant. By applying group-level analysis, Bayesian model selection and Bayesian model averaging, the optimal family and model for all experimental tasks were found. For all social-cognitive processes, effective connectivity from STS to IPL and from STS to BA44 was found. For imitation, additional mutual connections occurred between STS and BA44, as well as BA44 and IPL. The results suggest inverse models in which the motor regions BA44 and IPL receive sensory information from the STS. In contrast, for imitation, a sensory loop with an exchange of motor-to-sensory and sensory-to-motor information seems to exist.

Key words: mirror neuron system; social cognition; fMRI; dynamic causal modeling; Bayesian model selection

Introduction

Mirror neurons (MNs) are considered essential building blocks to present the neuronal basis of social cognition (Frith and Frith, 2007; Gallese, 2007). It is assumed that humans obtain an immediate understanding of others' emotions, desires and intentions by representing the others' motor states in their own motor system (Gallese, 2007; Oberman *et al.*, 2007; Kilner and Lemon, 2013). A large body of functional imaging (fMRI) and electroencephalography (EEG) studies have provided indirect evidence for the involvement of MNs in social-cognitive processes (see Bekkali *et al.*, 2020 for a meta-analysis), including imitation (Iacoboni, 2005; Molenberghs *et al.*, 2009), action learning (Cook *et al.*, 2014), emotion recognition (Keuken *et al.*, 2011), theory of mind (ToM) (Gallese and Goldman, 1998; Mier *et al.*, 2010b) and empathy (Iacoboni, 2009; Moore *et al.*, 2012). However, research is needed to explain how regions of the MN system (MNS) interact to understand others' emotions and intentions. Studies using dynamic causal modeling (DCM) or related methods for inference about effective connectivity can help to get a deeper understanding of sensory–motor processing in the MNS (see Schurz *et al.*, 2020a for a review) and inform computational models that model the physiological processes in the MNS.

MNs were discovered by Di Pellegrino *et al.* (1992) in the macaque monkey. The authors found a subset of motor neurons that fire when the animal executes an action and when the animal observes a comparable action (Di Pellegrino *et al.*, 1992; Rizzolatti and Craighero, 2004). Due to ethical reasons, direct examination of MN activity by single-cell recordings is excluded in healthy human participants. Thus, indirect, non-invasive measurements such as fMRI and EEG are applied to investigate the functioning of the MNS. These studies in humans identified several brain regions with mirror properties (Molenberghs *et al.*, 2012). Primate data support the existence of MNs in two of these regions, and they build the basis for models explaining MN function: Broca's area (BA44) located in inferior frontal gyrus (IFG) with adjacent ventral premotor cortex that corresponds to area F5 of the primate brain (Rizzolatti *et al.*, 1996) and inferior parietal lobule (IPL) (Rizzolatti and Craighero, 2004; Rizzolatti, 2005). In addition, the posterior superior temporal sulcus (STS) has been suggested as the region that conveys the visual input to the MNS (Van Overwalle, 2009).

Received: 9 April 2021; **Revised:** 15 November 2021; **Accepted:** 27 January 2022

These regions of the human MNS have been mainly identified in studies with visual images of actions and execution of motor actions (Iacoboni and Dapretto, 2006; Molenberghs *et al.*, 2012). However, a special interest in MNs exists due to their proposed role in social cognition (Gallese, 2007; Mier *et al.*, 2010b). Recently, we demonstrated common activation for imitation, empathy and affective ToM across and within individuals in STS, IPL and BA44 during social cognition tasks (Schmidt *et al.*, 2021). Since dysfunction of the MNS has been assumed to result in core symptoms of mental disorders, a growing number of studies also focusses on the role of the MNS for mental disorders, such as in autism (Rizzolatti and Fabbri-Destro, 2010; Hamilton, 2013), psychopathy (Mier *et al.*, 2014), schizophrenia (Mier *et al.*, 2010b) and borderline personality disorder (Mier *et al.*, 2013). Thus, a deeper understanding of the human MNS could also enhance our knowledge of mental disorders.

To go beyond the activity of the MNS, computational models have been constructed to conceptualize the function of MNs (for a review, see Giese and Rizzolatti, 2015). These models are based on data regarding the anatomical and effective connectivity between the brain regions of the MNS. Anatomically, there are prominent bidirectional connections between STS and IPL on the one hand and IPL and IFG, including BA44, on the other hand, as well as projections of visual areas onto STS (Nelissen *et al.*, 2011). In humans, direct connections from STS to IFG have also been found (Catani *et al.*, 2005; Rilling *et al.*, 2008). Based on this connectivity profile and the mirror properties of IPL and IFG, common assumptions of the models of the MNS are that (i) visual information enters the MNS via the STS, (ii) motor information is transferred from IFG directly or via the parietal cortex to STS, and in several cases, (iii) sensory information is projected from STS, directly or via parietal cortex, to the IFG, closing the sensory–motor loop. Depending on the modeling framework, the core regions of the MNS, STS, IPL and IFG are interpreted as recurrent neural networks (e.g. Yamashita and Tani, 2008), neural fields (e.g. Erlhagen *et al.*, 2006), layers of deep neural networks (e.g. Fleischer *et al.*, 2013) or, more abstractly, as elements of action controller architectures (implementing a forward model from IPL to STS (motor-to-sensory) and an inverse model from STS to IPL (sensory-to-motor), Wolpert *et al.*, 2003) or Bayesian predictors (where IFG and IPL act as empirical priors for STS, Kilner *et al.*, 2007a; Friston *et al.*, 2011). Most models assume a hierarchical organization of the MNS, at least implicitly, with the STS on the bottom, representing visual information, the IPL representing kinematic details of the movement, and the IFG, standing on top of the hierarchy, representing more abstract motor goals (Erlhagen *et al.*, 2006; Kilner *et al.*, 2007a; Yamashita and Tani, 2008; Friston *et al.*, 2011; Fleischer *et al.*, 2013, but see Grafton and Hamilton, 2007 for a different hierarchy with the IPL on top).

For a given task, the effective connectivity, i.e. the concrete flow of information, including the temporal profile of information flow during that task is of the highest importance. A method that estimates the time course of effective connections between active brain regions is DCM. DCM is a widely used method to find the effective connectivity between activated brain regions by estimating the hidden state parameters of the observed experimental data (Friston *et al.*, 2003). The few existing DCM studies on the MNS have largely confirmed the connectivity profile outlined above (Lebreton *et al.*, 2012; Sasaki *et al.*, 2012; Thioux and Keysers, 2015; Urgen and Saygin, 2020), including the direct pathway from STS to IFG (Sasaki *et al.*, 2012). However, it should be noted that all of the above studies have been conducted for hand movements. Facial expressions, which are the primary source of social information, are a special case in motor processing because (unless using a mirror) we do not get visual feedback from our facial movements. Indeed, facial expressions result in different activation of the MNS compared to social hand movements (Montgomery and Haxby, 2008), and connectivity studies in monkeys have shown that facial communicative expressions are being processed in regions of the MNS that lack a robust parietal component and are more connected to limbic and ventral prefrontal areas (Ferrari *et al.*, 2017). To date, only parts of the MNS have been studied with DCM while watching emotional facial expressions (Sato *et al.*, 2015, 2016) or social scenes (Arioli *et al.*, 2018). A complete DCM network analysis of the MNS during processes of social cognition is missing. Hence, to date, it is an open question whether modeling approaches of the MNS based on hand movements can be transferred to social tasks.

In a recent publication, we showed effective connectivity from STS to IPL and IFG, as well as from IFG to IPL in a facial imitation task (the data resulted from a subgroup of the current sample, Sadeghi *et al.*, 2020). The current fMRI study goes one step further by examining three social-cognitive tasks within the same participants, allowing us to address whether there is a common or 'standard' route of processing across different aspects of social cognition or whether the interaction between these regions is dependent on the exact social-cognitive process. The three social cognitive tasks were empathy, imitation and ToM. All tasks were based on pictures of facial expressions. Out of the three tasks, only imitation included actual movements. Activation patterns of these tasks are published in Schmidt *et al.* (2021). By applying region of interest (ROI) analyses, activation in BA44, IPL, STS, amygdala and fusiform gyrus was found bilaterally for all contrasts of interest for the present analyses, i.e. imitation > control, affective empathy > control and ToM > control. Behavioral data analysis showed the highest empathy ratings for cognitive empathy, followed by affective empathy and lowest ratings for personal distress. In addition, response times were longer for ToM than for emotion recognition, neutral face processing and control. Performance, as indicated by the percentage of correct answers, showed the same pattern with ToM being the most difficult condition. To examine effective connectivity between the brain regions of the MNS, an optimized version of DCM (Sadeghi *et al.*, 2020) was applied. We designed 540 models divided into four families and used Bayesian model comparison and Bayesian model averaging (BMA) for group analysis to find the optimal family and model fitting to our experimental data. We assumed finding direct effective connectivity from STS to IPL and/or BA44 (inverse model), independent of the specific social-cognitive process. Further, we were interested in the mutual connections (forward model) and connections between IPL and BA44. The main question was whether common effective connectivity characterizes the different social-cognitive processes or whether distinct communication is found.

Materials and methods

Data sets

Participants

We invited 80 participants to two appointments. Of these, we excluded one due to excessive head movement (more than two scans with rotation > 3° and translation > 3°), one due to anatomical aberrations (incidental finding in the ventricle area which needed further medical evaluation), three due to technical/inclusion criteria issues ($n = 1$: the fMRI measurement stopped for unknown reasons; $n = 1$: BDI score of 27 despite careful telephone screening; $n = 1$: biased answers in questionnaires and bizarre behavior during the experimental sessions) and eight because they did not show significant activation at $P < 0.5$

in at least one of the three ROIs in all three tasks. These eight participants were excluded to allow a direct comparison between the tasks. Beyond these measures, we did not control for outliers. Our final sample for the DCM analysis contains 67 subjects (39 women, 28 men, mean age = 23.39, SD = 3.60) with university entrance qualifications who reported no history of mental or neurological disorder. The first appointment included a simultaneous EEG-fMRI assessment, the second transcranial magnetic stimulation prior to fMRI. All data analyzed in this manuscript were taken from the fMRI data of the first appointment. It should be noted that 42 of these 67 participants were part of the analyses conducted in Sadeghi *et al.* (2020) for establishing the novel DCM method (only based on the imitation task of which activation was modeled with a boxcar, instead of a stick function, as for the present analyses).

Tasks and experimental procedures

We used three experimental paradigms covering different processes of social cognition: an imitation task, an empathy task and a ToM task. For all three tasks, we used pictures from the Karolinska Directed Emotional Faces stimulus set of five females and five males and control stimuli without social information. Tasks were implemented with Presentation Software (Version 18.1; www.neurobs.com) and presented via video goggles. Responses were given with a diamond-shaped button device (Current Designs, Inc., Philadelphia, PA, USA). Task order with (1) imitation, (2) empathy and (3) ToM was fixed for all participants.

Study procedure

The study was approved by the ethics committee of the Medical Faculty Mannheim, University of Heidelberg. Participants received oral and written information about the study procedure and aims, signed written informed consent and practiced all tasks on a laptop. The three social-cognitive paradigms that were presented during fMRI are shortly described in the following. For more details and the results of the activation analyses, please refer to Schmidt *et al.* (2021).

Imitation. The imitation task contained four conditions (Figure 1A). In the observation and imitation conditions, angry and fearful faces were presented, and participants had to observe or imitate the presented faces, respectively. In the Execution condition, participants read an emotional word (anger or fear) and had to perform the according to facial expression. In the control condition, participants had to read out loud a presented letter (German letters 'Ä' or 'A' to resemble the facial expressions of anger and fear, respectively). Conditions were presented blockwise, with each experimental block containing four stimuli of 5 s (Figure 1B). Control blocks contained two stimuli of 3 s and were presented interleaved with the experimental blocks. A fixation cross with jittered duration was presented between trials (for 1–3 s) and between blocks (for 4–6 s). All blocks were repeated, so there were 20 trials for each experimental condition and 30 trials for the control condition. The task duration was 13 min.

Empathy. The empathy task contained four conditions, again presented blockwise. In the experimental conditions, namely, affective empathy, cognitive empathy and distress, angry and fearful faces were presented. In the Control condition, circles of different sizes were shown, and at the beginning of each block, one out of four instruction cues was displayed. The instruction cues were 'How bad do I feel?' (distress), 'How bad does the presented person feel?' (cognitive empathy), 'How much do I

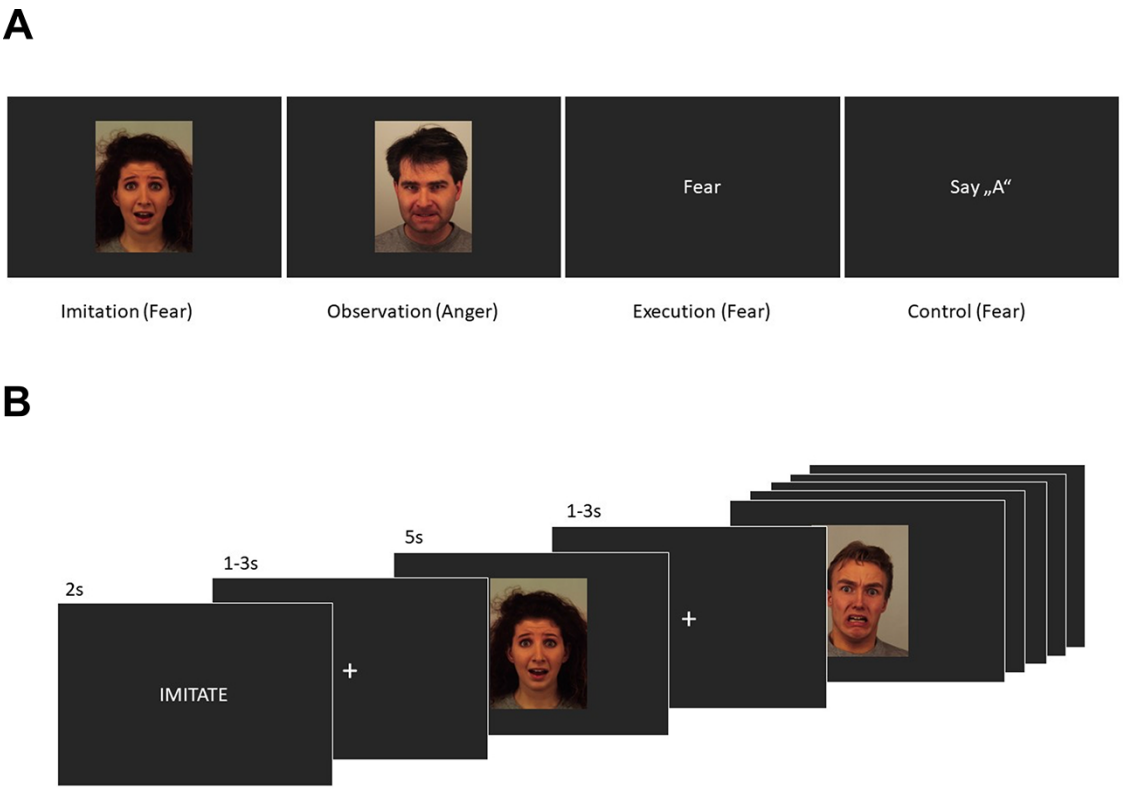

Fig. 1. Imitation Task. (A) Overview over the four conditions imitation, observation, execution and control, with exemplary stimuli. In all conditions except control, half of the stimuli showed angry, the other half fearful facial expressions or word cues. The control condition served as a motor control without emotional information. (B) Task flow with presentation times. At the beginning of each block, a cue word served as an instruction. In the experimental blocks, four stimuli with a duration of 5 s were presented, in the control blocks, two stimuli with 3 s duration.

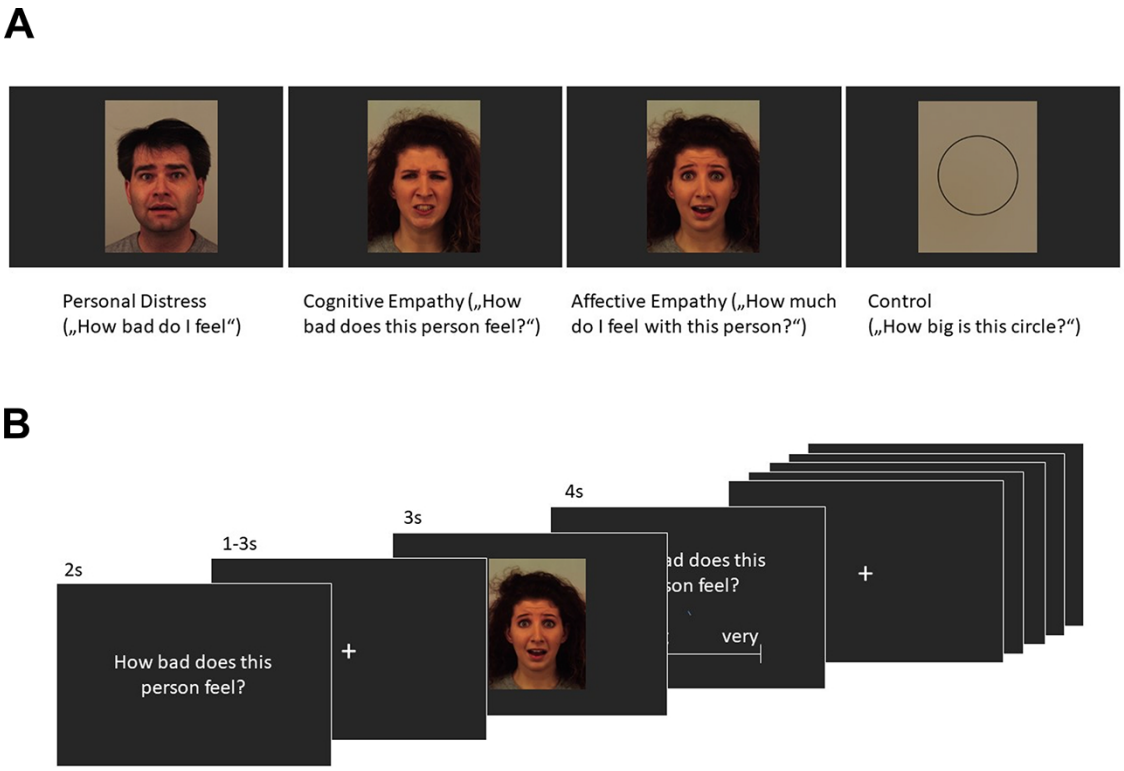

Fig. 2. Empathy task. (A) Overview over the four conditions personal distress, cognitive empathy, affective empathy and control with exemplary stimuli. In all empathy conditions, half of the stimuli showed angry, the other half fearful facial expressions or word cues. The control condition showing a circle of different sizes served as a visual control that also required rating on a visual analog scale. (B) Task flow with presentation times. At the beginning of each block, the cue question was presented. In the experimental blocks, four stimuli with a duration of 3 s were presented, in the control blocks, two stimuli with 3 s duration.

empathize with the presented person?' (affective empathy), or 'How big is the circle?' (control condition) (Figure 2A). The participants' task was to think about the cued question while watching the stimuli. After each stimulus to answer the question on a continuous visual analog scale from 'not at all' to 'very much' (control condition: 'small' to 'large'; Figure 2B).

Analogous to the imitation task, we chose a design with experimental blocks of four stimuli of 3 seconds alternating with a control block of two stimuli of 3 s. A fixation cross with jittered duration was presented between trials (1–3 s) and between blocks (4–6 s). Again, there were 20 trials for each experimental block and 30 total control trials. The task duration was 17 min.

Theory of mind (ToM). The ToM task had three experimental conditions such as affective ToM, emotion recognition and neutral face processing and a control condition (Figure 3A). Conditions were presented in pseudo-randomized order in an event-related design (Figure 3B). For each condition, 20 trials were shown. One trial consisted of a statement (e.g. 'This person is about to run away' for affective ToM, 'This person is angry' for emotion recognition, 'This person is female' for neutral face processing, and 'This is a circle' for Control), followed by an angry or fearful emotional facial expression (for affective ToM and emotion recognition), a neutral facial expression (for neutral face processing) or a circle or triangle (for control). Participants had to select 'yes' or 'no' as the appropriate answer. Each statement and face was presented for 2 s, and the inter-stimulus interval lasted between 1 and 3 s. The task duration was 8 min.

Data acquisition

fMRI data were acquired with a 3T Siemens Magnetom Trio with a 12-channel head coil at the Central Institute of Mental Health in Mannheim, Germany. At first, an MPRage was measured (TR = 1570 ms, TE = 2.75 ms, flip angle = 15°, field of view = 256 mm, matrix = 256×256, voxel size 1×1×1 mm^3). For the recording of task activation, echo-planar imaging was set to 32 descending 3×3×3 mm^3 slices with 1 mm gap, TR of 2000 ms, TE of 30 ms, flip angle of 80°, field of view 192 mm and matrix of 64 × 64. The volume was aligned to AC-PC and tilted by –20°. The imitation task was measured with 397 volumes, the empathy task with 518 volumes and the ToM task with 248 volumes. Movement correction was performed for scans exceeding 3 mm translation or 3° rotation, by replacing the scan with the mean of the below-threshold scans before and after.

Time series extraction

Data were preprocessed and analyzed with Statistical Parametric Mapping 8 (SPM8, http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). Pre-processing consisted of slice time correction, realignment to the mean image, normalization with segmentation and co-registration to the individual MPRage, and resampling with 3 mm^3 voxel size, as well as smoothing with 8 mm Gaussian kernel. First-level-analyses were achieved by general linear models with the onsets of the conditions (for the imitation task: imitation, observation, execution and control; for the empathy task: affective empathy, cognitive empathy, distress and control; for the theory of mind task: affective theory of mind, emotion recognition, neutral face processing and control) and the six movement parameters from the realignment procedure as covariates.

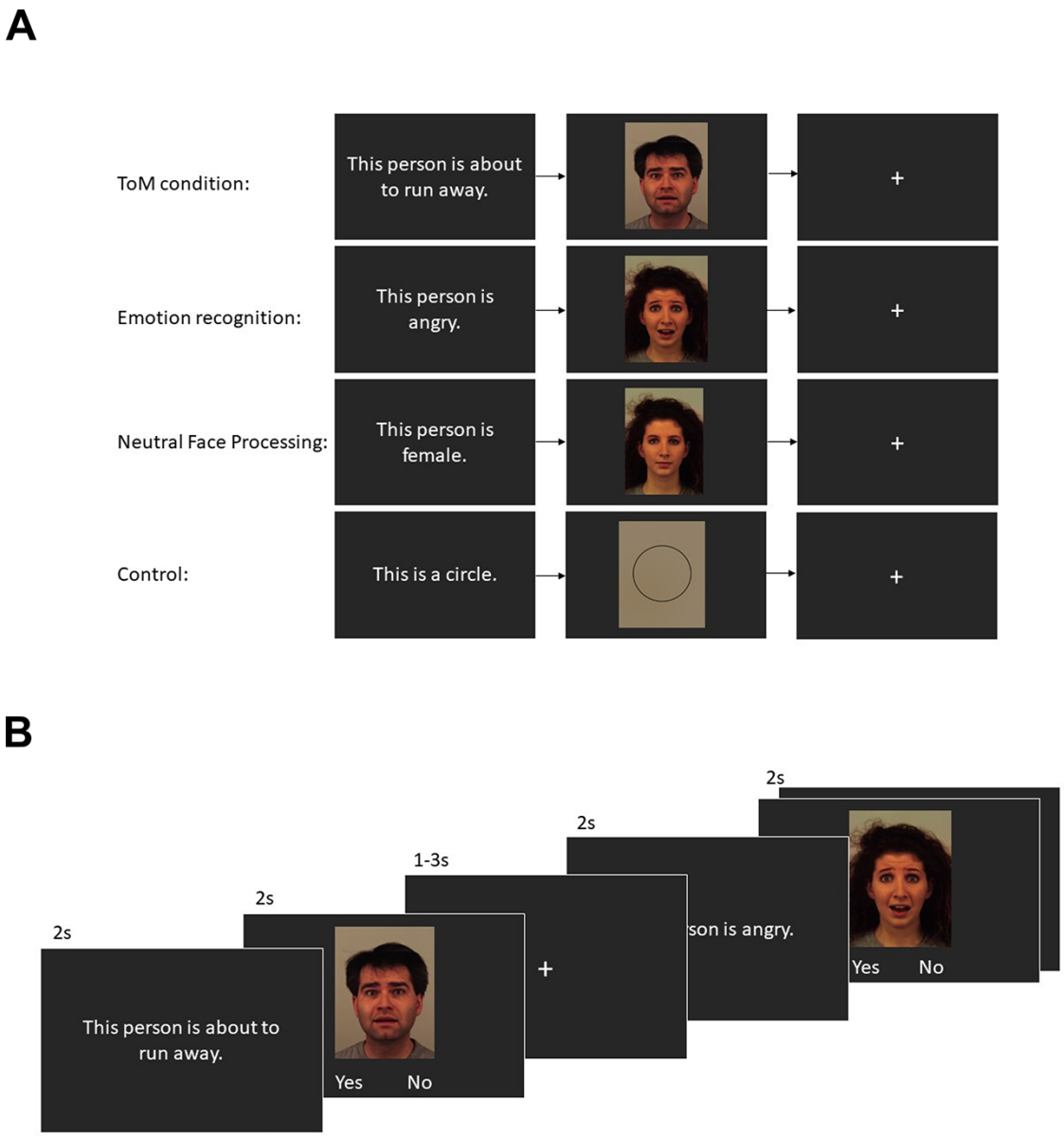

Fig. 3. ToM task. (A) Overview over the four conditions ToM, emotion recognition, neutral face processing and control, with exemplary stimuli. Faces in the ToM and emotion recognition condition showed angry and fearful expressions, in the neutral face processing condition neutral expressions, and geometric figures served as stimuli in the control condition. (B) Task flow with presentation times. The stimuli are presented in pseudo-randomized order in an event-related design.

The tasks were analyzed as an event-related design, by convoluting the HRF with a stick function. First eigenvariates of the time series of imitation > control, affective empathy > control and ToM > control were extracted with $P < 0.5$ without a cluster size threshold while adjusting for the activation during imitation, affective empathy and ToM, respectively. All trials of the conditions were included, independent of whether participants responded correctly or incorrectly. The threshold of $P < 0.5$ was selected liberally to ensure the inclusion of a majority of participants. The first eigenvariates were extracted from the individual peak voxels with a sphere of 8 mm from the main regions associated with the human MNS: BA 44, IPL and STS, all on the right hemisphere, to avoid confounding effects of language processing. The masks for BA 44 (Brodmann atlas) and IPL (AAL atlas) were taken from the WFU_pickatlas. The BA44 mask was smoothed with a dilation factor of 1 to allow a continuous mask. The posterior STS mask was based on activation in a study on social cognition with a similar design for the ToM task, used in this study and has been used as a ROI in previous publications (Mier *et al.*, 2010a; 2010b).

DCM

We used DCM to estimate effective connectivity between the ROIs in the three tasks. DCM uses a Bayesian framework to estimate the posterior values of intrinsic connections between brain regions and exogenous or driving inputs on different nodes (task stimuli). In the bilinear state equation, all connections can be modulated by contextual inputs (i.e. task-related changes in effective connectivity), and all brain regions have self-connections (Friston *et al.*, 2003).

We performed the DCM10 (r6313) in SPM8, which we have modified before, by replacing the standard linear equations with Wilson–Cowan-based models (Sadeghi et al., 2020). This model shows the changes in neural activity according to

$$\dot{z}_t = -z_t + S(x)$$

$$S(x) = \frac{1}{1 + exp(-\alpha x)} - \frac{1}{2}$$

$$x = (A + \sum_{j=1}^{m} u_j B^j) z_t + Cu$$

where z_t denotes the time derivative of neuronal activity and function $S(x)$ shows a sigmoid function in which the parameter α determines its slope. Matrix A describes the endogenous connectivity between the neural nodes and B^j shows which connection is modulated by the direct contextual input u_j and C embodies the direct influences of external input u on brain regions. We can specify these parameters $\theta_c = \{A, B^j, C\}$ and build different models to compare them to find the best model fitted to the observed data. Note that all parameters in θ_c represent effective connectivity that may vary, e.g. across different tasks, while using the same set of anatomical connections. In particular, a non-significant entry in one of the matrices does not imply a missing synaptic connection between two regions, but merely that this connection is not used in this particular task. Here, we estimate independent sets of parameters θ_c for each task (imitation, empathy, and ToM) and compare them afterward. The contextual inputs u_j are restricted to the external input u ($u_j = u$) for simplicity. Thus, the B matrix mostly regulates the activity dynamics at the beginning and the end of the external stimulus u.

In our previous study (Sadeghi et al., 2020), we showed that the modified DCM allows a significantly better fit to the empirical data than the standard bilinear model. We tested it on three different datasets and showed its superiority in fitting these datasets. This kind of neuronal equation can infer the sigmoid transfer function as an averaged *f–I* curve of activation in brain regions that have a sigmoidal format and has the potential of adopting generative models for fMRI time-series to be informed by physiological principles. In this way, the parameters obtained by DCM show different and more robust results and can be directly interpreted physiologically.

Model specification

In DCM, one can construct different models according to these factors: (1) which regions receive external inputs (Matrix C), (2) how the activated regions are connected (Matrix A) and (3) which of these connections are modulated by the contextual inputs (Matrix B). However, to avoid extensive numbers of models, a hypothesis-driven approach is warranted to decrease the model space. In this way, we constructed models according to the hypothesis that visual input always integrates into the STS region, and from STS, this input would propagate to the IPL and BA44. This assumption is based on previous research on the MNS (Iacoboni et al., 2001; Barraclough et al., 2005; Kilner et al., 2007b). The effective connectivity between the two regions can be both unidirectional or bidirectional. We have three nodes, and these nodes can maximally have six connections in case of mutual connectivity. In addition, we have considered the modulatory input on the interregional connections. Thereby, each combination of the intrinsic connection between different regions can have 2^n modulatory inputs, n (in our case, n can be 2, 3, 4, 5 and 6) being the number of endogenous connections between the regions of interest. In this way, we built 540 models partitioned into four families as explained in the following.

As shown in Figure 4, a feed-forward connection from STS to BA44 and IPL is always available for family 1. However, they can also have mutual connections. Within this family, IPL and BA44 can have a connection or not; the connection between them can be unidirectional or bidirectional. Solid lines in Figure 4 show the connections that always are present (input from STS to BA44 and IPL), and dashed lines show the connections which can be present or not (e.g. the backward connection from IPL to STS). For family 2, the activation propagates from STS to IPL and then from IPL to BA44, and for family 3, from STS to BA44 and then to the IPL region. In family 4, the common feature is that STS either gives input into IPL, or BA44 and this information is forwarded back to STS via the regions that are not getting input from STS. Furthermore, each of these families has modulations on the interregional connections. We have considered one experimental condition (imitation, empathy and ToM) as modulatory input for each task separately. So, in this approach, we have first defined the baseline A matrix and partitioned the models into four families and then inserted the modulatory input on each connection. For example, sub-family 1 within family 1 includes 36 models with different A and B matrices variations.

Model selection

To find the most probable model from the model space above, which fitted best to the observed data, we used group analysis Bayesian model selection (BMS) (Stephan et al., 2009) among all single models with inference over families of models (Penny et al., 2010). To account for the heterogeneity of the model structure across subjects, we used random effect (RFX) BMS, which uses the hierarchical Bayesian modeling to estimate the parameters of a Dirichlet distribution considering the probabilities of all models. With this technique, subjects can have different best models, and the effects of outliers are very limited in the BMS results. The results of RFX BMS are reported in terms of exceedance and expected probabilities, which are the probability that a particular model is more likely than any other model tested, and the probability of obtaining the model for a random subject from the population respectively. The best model is the one with the largest exceedance or expected probability.

Since in RFX BMS, the exceedance and expected probability sum up to one, a large model space reduces the probabilities for each model, which hampers finding a single winning model. Thus, models are implemented in groups as a model or as a family of models, in which all models share some features (Stephan et al., 2010) (e. g. a fixed particular set of connections). This technique can compensate for the issue of large numbers of models and narrow the search for the optimal model. Note that the number of families should also be small. It is also possible to have different numbers of models per family, as the prior for each model is weighted by the number of models in its family (i.e. the prior is that all families (rather than all models) are equally likely) (Penny et al., 2010).

Still, finding sufficient evidence for one model or family of models being optimal is not always possible. BMA helps resolve

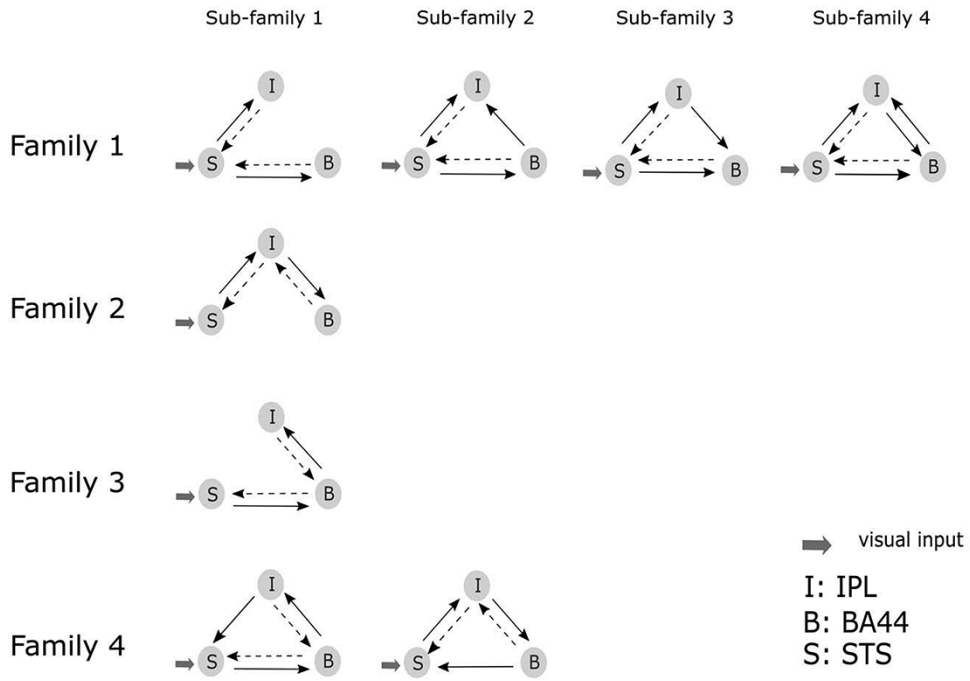

Fig. 4. Model space. Schemata of parameters that made up the models included in four families. Solid lines show the connections that always are present, and dashed lines the connections that can be present or not. The modulatory input can be exerted on these interregional connections. Family 1 consists of four sub-families, and family 4 includes two sub-families. For each family, we assumed visual input external input always integrates into the STS region.

this inference uncertainty by averaging over all models within the family or even the whole model space. It is the average of the connectivity parameters over models, weighted by the models' posterior probabilities. Thus, the most probable models will contribute the most to the model averaging (Penny *et al.*, 2010). Applied to our data, we divided models into families based on the A matrix, i.e. their intrinsic connectivity structure. This division was conducted in a stepwise manner. First, we identified how the input from the STS region would activate other regions. Then we entered models from the winning family (family 1 in Figure 4), as achieved by RFX BMS, into the second set of BMS analyses to find how the BA44 and IPL regions connect. Finally, we used BMA to make the inference on parameters.

Results

DCM results

Using the Bayesian model comparison, we first used the family level inference to find which set of models in Figure 4 (divided into four families) is selected over other families. Results indicated that family 1, in which input from the STS region propagates to BA44 and IPL with feed-forward connections, is the most probable structure for all three experiments. The exceedance probability for all experiments is larger than 0.9 (Figure 5) for family 1.

Next, we performed the Bayesian model comparison for the winning family to determine which one of the four sub-families in family 1 has the highest probability for each of the experimental tasks (Table 1). Models in family 1 differ in connectivity between BA44 and IPL regions, with no connection, directional and bidirectional connections. For imitation, sub-family 4 with an exceedance probability of 0.95 (Table 1), in which IPL and BA44 have mutual connections, is substantially more probable than any of the alternatives. Results demonstrated for the empathy and ToM tasks, sub-family 3 to be most likely, where IPL has a feed-forward connection to BA44.

However, as shown in Table 1, for empathy and ToM, the exceedance probabilities of the winning sub-families do not provide definitive evidence (0.49 max). Therefore, as a third step, we use BMA to account for model uncertainty by averaging over all models in family 1. In this regard, we performed BMA to obtain the estimates of effective connectivity and their modulation to incorporate the group-level inference on the parameters. The BMA results are shown in Tables 2 and 3 and Figure 6 for all three experimental tasks. These results for empathy and ToM are on all models within family 1 and for imitation only within the winning sub-family 4. In Figure 6, we only illustrate the parameters with a probability greater than 95% (i.e. deviate significantly from zero) for all three tasks. According to these results, visual stimuli integrated into the STS are fed forward to IPL and BA44 with unmodulated connections for empathy and ToM and modulated STS→BA44 connection for imitation. Furthermore, for the imitation task, there is additional feedback from BA44 to STS and bidirectional connections between BA44 and IPL.

Discussion

Here, for the first time, we present results of effective connectivity within the human MNS for three different social-cognitive processes. We used a stepwise family level inference to find the best fitting effective connectivity model among STS, IPL and BA44, with the prior assumption that the external visual input enters the STS. We tried different models and significantly decreased the model space for a final BMS on a smaller number of models. Subsequent BMA revealed that effective connectivity for imitation, empathy and ToM is always characterized by a feed-forward information processing from STS to IPL and BA44, suggesting an inverse (sensory-to-motor) internal model. In addition, we show

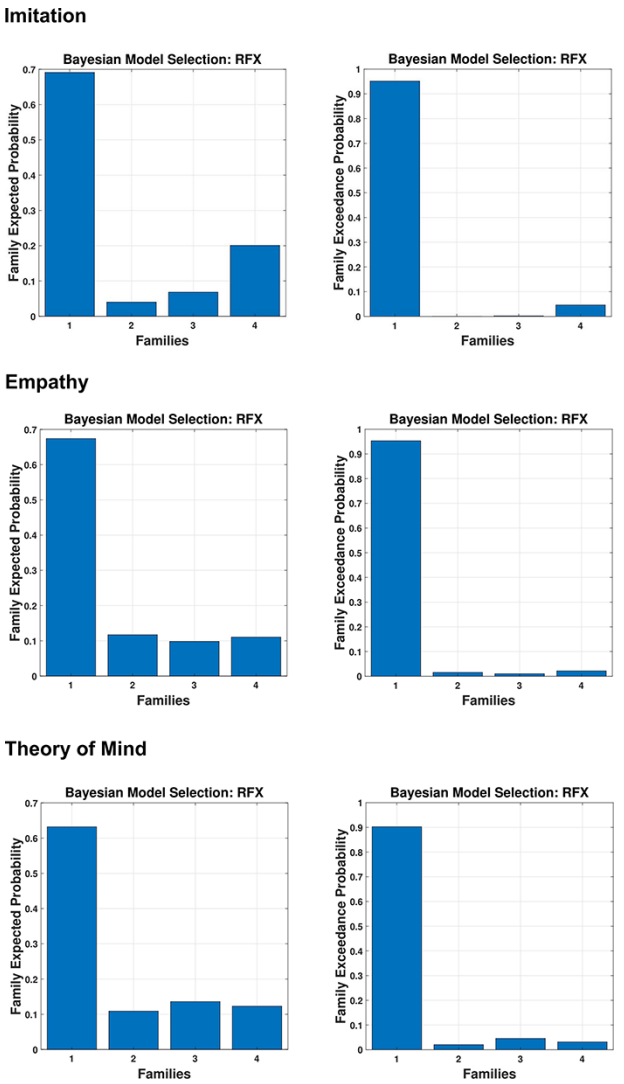

Fig. 5. Family level inference performed on models within the families in Figure 1. All experiments demonstrate that family 1 has the highest expected and exceedance probability, in which the models within this family have forward connections from STS to IPL and BA44.

Table 1. RFX BMS results on models within family 1 for all experiments

Imitation	Sub-family 1	Sub-family 2	Sub-family 3	Sub-family 4
Expected probability	0.04	0.14	0.11	0.71
Exceedance probability	0	0.04	0.01	0.95
Empathy				
Expected probability	0.20	0.22	0.38	0.20
Exceedance probability	0.16	0.18	0.49	0.17
Theory of Mind				
Expected probability	0.25	0.25	0.28	0.22
Exceedance probability	0.26	0.24	0.31	0.19

Table 2. BMA results for endogenous connectivity (Matrix A) (in Hz). Next to each parameter is the posterior probability, which is different from the test statistic (zero). We consider PP > 0.95 as the threshold at which parameters are significant

Imitation from to	STS	IPL	BA44
STS	−0.4625, 1.00	0.0348, 0.71	0.3322, 0.99
IPL	1.1059, 1.00	−0.4738, 1.00	0.9853, 1.00
BA44	1.4284, 1.00	0.7655, 1.00	−0.4768, 1.00
Empathy			
STS	−0.4843, 1.00	0.1978, 0.88	0.0551, 0.66
IPL	0.4127, 0.99	−0.4948, 1.00	0.0005, 0.50
BA44	0.4318, 0.99	0.1042, 0.77	−0.4918, 1.00
Theory of Mind			
STS	−0.4885, 1.00	0.1330, 0.80	0.0810, 0.72
IPL	0.4048, 0.99	−0.4964, 1.00	0.0528, 0.66
BA44	0.5219,0.99	0.0701, 0.70	−0.4958, 1.00

Table 3. BMA results for modulatory connectivity (Matrix B) (in Hz). Next to each parameter is the posterior probability, which is different from the test statistic (zero). We consider PP > 0.95 as the threshold at which parameters are significant

Imitation from to	STS	IPL	BA44
STS	–	0.0046, 0.57	0.0747, 0.80
IPL	0.1102, 0.88	–	0.0371, 0.65
BA44	0.1886, 0.97	0.0806, 0.78	–
Empathy			
STS	–	0.0064, 0.53	0.0019, 0.52
IPL	0.0745, 0.80	–	0.0008, 0.51
BA44	0.0347, 0.65	0.0003, 0.50	–
Theory of Mind			
STS	–	0.0030, 0.52	0.0025, 0.52
IPL	0.0466, 0.70	–	0.0007, 0.51
BA44	0.0489, 0.72	0.0003, 0.50	–

that information flow between these regions of the MNS is more complex for imitation than for Empathy and ToM, including both forward and inverse information flow.

Information flow from STS to IPL and BA44 is in general agreement with the assumption that STS is passing visual information to the MNS. In computational models, it is assumed that this path reflects visual information to be converted into a motor representation (Kilner, 2011; Giese and Rizzolatti, 2015), termed an inverse model in the context of controller architectures (Wolpert *et al.*, 2003). However, contrary to most MNS models (Kilner, 2011), a forward model from motor-to-sensory areas is missing, as well as a clear hierarchy of areas within the MNS (Erlhagen *et al.*, 2006; Grafton and Hamilton, 2007; Kilner *et al.*, 2007a; Yamashita and Tani, 2008; Friston *et al.*, 2011; Fleischer *et al.*, 2013). We can only speculate on the reasons. One possibility is that social-cognitive processes associated with pictures of facial expressions without actual movements necessitate a direct information flow to both MNS regions because the processing of social information neither relies on action goal recognition (and with this, further feed-forward information flow of the action goal representation from IFG to the parietal cortex for detailed kinematics) nor, on the opposite, information about exact motor states that is transferred to IFG for matching with possible motor aims. The latter is in contrast to the assumption that intention is inferred by recognition of the current emotional state plus the simulation of possible further actions (Mier *et al.*, 2010b). Overall, our results support the

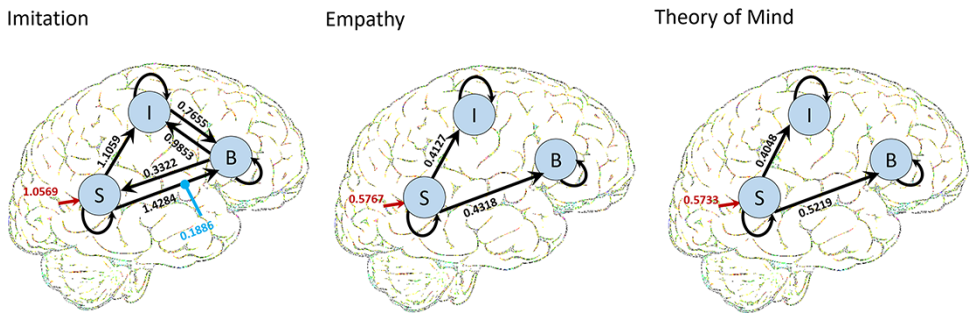

Fig. 6. BMA results for all three tasks for the winning family 1. Here, we illustrate only the parameters which are significantly > 0. The values for external inputs (Matrix C), which are not reported in Tables 2 and 3, are shown here (all parameters are in Hz).

notion of a different processing route for emotional facial expression as connectivity data in monkeys suggest (Ferrari *et al.*, 2017), revealing a direct flow of information from STS to IFG, bypassing the IPL. Interestingly, combining results from different effective connectivity studies in humans, links between STS and IFG can also be established via the amygdala (Ćurčić-Blake *et al.*, 2012; Bruneau *et al.*, 2015; Seok and Cheong, 2019) and the prefrontal cortex (Schuwerk *et al.*, 2014; Arioli *et al.*, 2018; Esménio *et al.*, 2019), both in agreement with the results from monkey studies (Ferrari *et al.*, 2017). Another explanation for the effective connectivity from STS to IPL and BA44 is higher attention demands for emotional facial expressions than hand movements. For example, Schuwerk *et al.* (2017) showed the role of the TPJ for ToM and attention. Their activation cluster labeled with anterior TPJ reaches into the IPL, while their posterior TPJ cluster overlaps with the posterior STS region (Schuwerk *et al.*, 2017). Furthermore, both regions share effective connectivity with the anterior cingulate cortex (ACC) in this study, which has a prominent role in attention (Davis *et al.*, 2000).

Taken together, these results suggest that IPL and BA44 may independently encode different aspects of emotional facial expressions during social cognition: the interplay between STS and IPL might be due to attentional processes during social cognition, while the STS-BA44 connectivity could reflect information flow of motor information to the MNS, potentially enriched or gated by emotional or cognitive aspects of the task. Consistent with such a possible division of labor between the different components of the MNS, we have recently shown that emotional valence can be discriminated in the human MNS, but BA44 does so in a more differentiated way compared to IPL (Schmidt *et al.*, 2020). Future studies are needed to disentangle these possible processes, e.g. by using an attentional condition and a social-cognitive condition, as in Schuwerk *et al.* (2017). In addition, further studies with independent samples are needed that examine effective connectivity in the MNS with different stimuli, including face and hand movements as well as pictures or videos. These studies would help to elucidate whether the connectivity pattern we found is mostly due to the unmoving pictures or to face instead of hand stimuli that were used in most studies on the MNS. Furthermore, including additional regions into the DCM analysis, most notably the amygdala, the ACC, as well as further regions of the prefrontal cortex, may elucidate whether the information in the STS is passed to IPL and BA44 directly or via any of these brain regions, thus further constraining models of MNS function.

While all social-cognitive processes showed effective connectivity from STS to IPL and BA44, the imitation task differed from the ToM and the empathy task, showing more connections between regions and modulation by the condition. In the imitation task, there is additionally a mutual connection between IPL and BA44. We assume these additional mutual connections to be explained by the demands of the imitation task. The imitation task was the only social-cognitive process that needed facial movements and matching these movements with the observed facial expression from the participants. The effective connectivity patterns during imitation suggest a sensory–motor loop with forward and inverse information flow, allowing the matching of motor and visual sensory states. The additional effective connectivity between BA44 and IPL suggests active information exchange between the motor goal (e.g. a fearful facial expression) and kinematics (e.g. corrugator muscle contraction). Also, the STS region embedded in such a closed-loop might serve as a comparator of own and observed movements, as suggested in agency models (Isoda, 2016). For empathy and ToM such interconnections between IPL and BA44 are not task-relevant because these processes seem to afford to process the motor expression, but no fine-tuning and matching of the own facial expression with the observed emotional state, as it is necessary for imitation. Thus, for imitation, our DCM results agree with internal inverse and forward models of sensory input and motor commands (Wolpert *et al.*, 2003; Kilner, 2011).

Limitations and outlook

The exceedance probability for sub-family 3 for empathy and ToM was not at a level that allows clear support. Thus, we used BMA. In comparison to the imitation of facial expressions, empathy and ToM are more complex social-cognitive processes that might result in more variance across participants, lowering the probability of finding a winning model. Future studies should investigate how personality traits or self-reported empathy influences the effective connectivity between these regions. Further, interplay with additional regions, such as the amygdala, might be even more important for empathy and ToM than for imitation. We can also extend the DCM models with the additional regions of the limbic system or the medial frontal cortex that plays an essential role in more cognitively effortful social-cognitive tasks (Schurz *et al.*, 2020a, 2020b). As we used the time series of activation from the right hemisphere to avoid confounding effects with language processing, it is open to further analyses and studies on whether the effective connectivity patterns in the left hemisphere or even across hemispheres are comparable. Also, replicating these connectivity patterns based on EEG data would be of high interest. Albeit connectivity and activity patterns suggest an active imitation of the participants (Schmidt *et al.*, 2021), since we did not apply a camera, or measure the activity of facial muscles, we have no proof that participants indeed imitated the facial expressions.

Notwithstanding these limitations, our results build the foundation for further advanced models of the MNS. First of all, albeit our findings warrant replication, they can inform further social cognition models, including those of direct matching and embodied simulation (Gallese, 2007; Donnarumma *et al.*, 2017), to involve information flow from STS to IPL and BA44. Further, the results of the modified DCM analysis (Sadeghi *et al.*, 2020) allow the estimation of physiological models of the MNS. One way to approach the human MNS without directly measuring the activity of individual neurons consists of the theoretical modeling of the involved cell assemblies (Hass *et al.*, 2016, 2019). The mathematical description of the activity of neuronal networks and the simulation of the dynamics makes it possible to calculate the indicators of the non-invasive measurement methods and compare them with the measured values. This approach would pave the way for statements about the physiology of the cell assemblies, which would become possible since the parameters of the model are directly related to biophysical properties such as cellular activation functions or synaptic conductivities.

Keeping in mind that fMRI does not allow the assessment of individual neurons and with these conclusions about MNs, the effective connectivity patterns suggest directed information flow between the regions of the MNS during social cognition, which might be the basis for embodied simulation (Gallese, 2007). This information flow can represent an inverse model transferring sensory information to motor neurons in mirroring regions. In addition, for imitation, a sensory–motor loop exists for matching between external and internal sensory and motor states, allowing us to match our facial movements with the observed emotion of our interaction partners (Lee *et al.*, 2006; Bastiaansen *et al.*, 2009; Prochazkova and Kret, 2017).

Acknowledgements

The authors are grateful to Zhimin Yan, Vera Eymann and Manuel Vietze for their assistance in data acquisition.

Funding

This work was supported by the WIN-Kolleg of the Heidelberg Academy of Sciences and Humanities.

Conflict of interest

The authors declare no conflict of interest.

References

Arioli, M., Perani, D., Cappa, S., *et al.* (2018). Affective and cooperative social interactions modulate effective connectivity within and between the mirror and mentalizing systems. *Human Brain Mapping*, **39**, 1412–27.

Barraclough, N.E., Xiao, D., Baker, C.I., *et al.* (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience*, **17**(3), 377–91.

Bastiaansen, J.A.C.J., Thioux, M., Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**, 2391–404.

Bekkali, S., Youssef, G.J., Donaldson, P.H., *et al.* (2020). Is the putative mirror neuron system associated with empathy? A systematic review and meta-analysis. *Neuropsychology Review*, **31**, 14–57.

Bruneau, E.G. Jacoby, N., Saxe, R. (2015). Empathic control through coordinated interaction of amygdala, theory of mind and extended pain matrix brain regions. *NeuroImage*, **114**, 105–19.

Catani, M., Jones, D.K., Ffytche, D.H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, **57**, 8–16.

Cook, R., Bird, G., Catmur, C., *et al.* (2014). Mirror neurons: from origin to function. *Behavioral and Brain Sciences*, **37**, 177–92.

Ćurčić-Blake, B., Swart, M., Aleman, A. (2012). Bidirectional information flow in frontoamygdalar circuits in humans: a dynamic causal modeling study of emotional associative learning. *Cerebral Cortex*, **22**, 436–45.

Davis, K.D., Hutchison, W.D., Lozano, A.M., *et al.* (2000). Human anterior cingulate cortex neurons modulated by attention-demanding tasks. *Journal of Neurophysiology*, **83**, 3575–7.

di Pellegrino, G., Fadiga, L., Fogassi, L., *et al.* (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, **91**, 176–80.

Donnarumma, F., Costantini, M., Ambrosini, E., *et al.* (2017). Action perception as hypothesis testing. *Cortex*, **89**, 45–60.

Erlhagen, W., Mukovskiy, A., Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, **1083**, 174–88.

Esménio, S., Soares, J.M., Oliveira-Silva, P., *et al.* (2019). Using resting-state DMN effective connectivity to characterize the neurofunctional architecture of empathy. *Scientific Reports*, **9**, 1–9.

Ferrari, P.F., Gerbella, M., Coudé, G., *et al.* (2017). Two different mirror neuron networks: the sensorimotor (hand) and limbic (face) pathways. *Neuroscience*, **358**, 300–15.

Fleischer, F., Caggiano, V., Their, P., *et al.* (2013). Physiologically inspired model for the visual recognition of transitive hand actions. *Journal of Neuroscience*, **33**, 6563–80.

Friston, K., Mattout, J., Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, **104**, 137–60.

Friston, K.J., Harrison, L., Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, **19**, 1273–302.

Frith, C.D., Frith, U. (2007). Social cognition in humans. *Current Biology*, **17**, R724–32.

Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society Biological Sciences*, **362**, 659–69.

Gallese, V., Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, **2**, 493–501.

Giese, M.A., Rizzolatti, G. (2015). Neural and computational mechanisms of action processing: interaction between visual and motor representations. *Neuron*, **88**, 167–80.

Grafton, S.T., Hamilton, A.F.D.C. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, **26**, 590–616.

Hamilton, A.F.D.C. (2013). Reflecting on the mirror neuron system in autism: a systematic review of current theories. *Developmental Cognitive Neuroscience*, **3**, 91–105.

Hass, J., Hertäg, L., Durstewitz, D. (2016). A detailed data-driven network model of prefrontal cortex reproduces key features of in vivo activity O. Sporns (ed). *PLoS Computational Biology*, **12**, e1004930.

Hass, J., Ardid, S., Sherfey, J., *et al.* (2019). Constraints on persistent activity in a biologically detailed network model of the prefrontal cortex with heterogeneities. *bioRxiv*, 645663.

Iacoboni, M., Koski, L.M., Brass, M., *et al.* (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the National Academy of Sciences*, **98**, 13995–9.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Current Opinion in Neurobiology*, **15**, 632–7.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, **60**, 653–70.

Iacoboni, M., Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, **7**, 942–51.

Isoda, M. (2016). Understanding intentional actions from observers' viewpoints: a social neuroscience perspective. *Neuroscience Research*, **112**, 1–9.

Keuken, M.C., Hardie, A., Dorn, B.T., *et al.* (2011). The role of the left inferior frontal gyrus in social perception: an rTMS study. *Brain Research*, **1383**, 196–205.

Kilner, J.M., Friston, K.J., Frith, C.D. (2007a). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, **8**, 159–66.

Kilner, J.M., Friston, K.J., Frith, C.D. (2007b). The mirror-neuron system: a Bayesian perspective. *Neuroreport*, **18**, 619–23.

Kilner, J.M. (2011). More than one pathway to action understanding. *Trends in Cognitive Sciences*, **15**, 352–7.

Kilner, J.M., Lemon, R.N. (2013). What we know currently about mirror neurons. *Current Biology*, **23**, R1057–62.

Lebreton, M., Kawa, S., D'Arc, B.F., *et al.* (2012). Your goal is mine: unraveling mimetic desires in the human brain. *Journal of Neuroscience*, **32**, 7146–57.

Lee, T.-W., Josephs, O., Dolan, R.J., *et al.* (2006). Imitating expressions: emotion-specific neural substrates in facial mimicry. *Social Cognitive and Affective Neuroscience*, **1**, 122–35.

Mier, D., Sauer, C., Lis, S., *et al.* (2010a). Neuronal correlates of affective theory of mind in schizophrenia out-patients: evidence for a baseline deficit. *Psychological Medicine*, **40**, 1607–17.

Mier, D., Lis, S., Esslinger, C., *et al.* (2013). Neuronal correlates of social cognition in borderline personality disorder. *Social Cognitive and Affective Neuroscience*, **8**, 531–7.

Mier, D., Haddad, L., Diers, K., *et al.* (2014). Reduced embodied simulation in psychopathy. *World Journal of Biological Psychiatry*, **15**, 479–87.

Mier, D., Lis, S., Neuthe, K., *et al.* (2010b). The involvement of emotion recognition in affective theory of mind. *Psychophysiology*, **47**, 1028–39.

Molenberghs, P., Cunnington, R., Mattingley, J.B. (2009). Is the mirror neuron system involved in imitation? A short review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, **33**, 975–80.

Molenberghs, P., Cunnington, R., Mattingley, J.B. (2012). Brain regions with mirror properties: a meta-analysis of 125 human fMRI studies. *Neuroscience and Biobehavioral Reviews*, **36**, 341–9.

Montgomery, K.J., Haxby, J.V. (2008). Mirror neuron system differentially activated by facial expressions and social hand gestures: a functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, **20**, 1866–77.

Moore, A., Gorodnitsky, I., Pineda, J. (2012). EEG mu component responses to viewing emotional faces. *Behavioural Brain Research*, **226**, 309–16.

Nelissen, K., Borra, E., Gerbella, M., *et al.* (2011). Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience*, **31**, 3743–56.

Oberman, L.M., Pineda, J.A., Ramachandran, V.S. (2007). The human mirror neuron system: a link between action observation and social skills. *Social Cognitive and Affective Neuroscience*, **2**, 62–6.

Penny, W.D., Stephan, K.E., Daunizeau, J., *et al.* (2010). Comparing families of dynamic causal models K. P. Kording (ed). *PLoS Computational Biology*, **6**, e1000709.

Prochazkova, E., Kret, M.E. (2017). Connecting minds and sharing emotions through mimicry: a neurocognitive model of emotional contagion. *Neuroscience and Biobehavioral Reviews*, **80**, 99–114.

Rilling, J.K., Glasser, M.F., Preuss, T.M., *et al.* (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience*, **11**, 426–8.

Rizzolatti, G., Fadiga, L., Gallese, V., *et al.* (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, **3**, 131–41.

Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anatomy and Embryology*, **210**, 419–21.

Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, **27**, 169–92.

Rizzolatti, G., Fabbri-Destro, M. (2010). Mirror neurons: from discovery to autism. *Experimental Brain Research*, **200**, 223–37.

Sadeghi, S., Mier, D., Gerchen, M.F., *et al.* (2020). Dynamic causal modeling for fMRI with wilson-cowan-based neuronal equations. *Frontiers in Neuroscience*, **14**, 1205.

Sasaki, A.T., Kochiyama, T., Sugiura, M., *et al.* (2012). Neural networks for action representation: a functional magnetic-resonance imaging and dynamic causal modeling study. *Frontiers in Human Neuroscience*, **6**, 1–17.

Sato, W., Kochiyama, T., Uono, S. (2015). Spatiotemporal neural network dynamics for the processing of dynamic facial expressions. *Scientific Reports*, **5**, 12432.

Sato, W., Kochiyama, T., Uono, S., *et al.* (2016). Direction of amygdala–neocortex interaction during dynamic facial expression processing. *Cerebral Cortex*, **27**, bhw036.

Schmidt, S.N.L., Sojer, C.A., Hass, J., *et al.* (2020). fMRI adaptation reveals: the human mirror neuron system discriminates emotional valence. *Cortex*, **128**, 270–80.

Schmidt, S.N.L., Hass, J., Kirsch, P., *et al.* (2021). The human mirror neuron system—a common neural basis for social cognition? *Psychophysiology*, **58**, e13781.

Schurz, M., Maliske, L., Kanske, P. (2020a). Cross-network interactions in social cognition: a review of findings on task related brain activation and connectivity. *Cortex*, **130**, 142–57.

Schurz, M., Radua, J., Tholen, M.G., *et al.* (2020b). Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, **147**(3), 293.

Schuwerk, T., Döhnel, K., Sodian, B., *et al.* (2014). Functional activity and effective connectivity of the posterior medial prefrontal cortex during processing of incongruent mental states. *Human Brain Mapping*, **35**, 2950–65.

Schuwerk, T., Schurz, M., Müller, F., *et al.* (2017). The rTPJ's overarching cognitive function in networks for attention and theory of mind. *Social Cognitive and Affective Neuroscience*, **12**, 157–68.

Seok, J.W., Cheong, C. (2019). Dynamic causal modeling of effective connectivity during anger experience in healthy young men: 7T magnetic resonance imaging study. *Advances in Cognitive Psychology*, **15**, 52–62.

Stephan, K.E., Penny, W.D., Daunizeau, J., *et al.* (2009). Bayesian model selection for group studies. *NeuroImage*, **46**, 1004–17.

Stephan, K.E., Penny, W.D., Moran, R.J., *et al.* (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, **49**, 3099–109.

Thioux, M., Keysers, C. (2015). Object visibility alters the relative contribution of ventral visual stream and mirror neuron system

to goal anticipation during action observation. *NeuroImage*, **105**, 380–94.

Urgen, B.A., Saygin, A.P. (2020). Predictive processing account of action perception: evidence from effective connectivity in the action observation network. *Cortex*, **128**, 132–42.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**, 829–58.

Wolpert, D.M., Doya, K., Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **358**, 593–602.

Yamashita, Y., Tani, J. (2008). Correction: emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Computational Biology*, **4**, e1000220.