Dissertation

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Sadjad Sadeghi

born in: Jahrom, Iran

Oral examination: 14.01.2022

The human mirror neuron system – Effective connectivity and computational models

Referees:

Prof. Dr. Joachim Hass

Prof. Dr. sc. techn. Mark E. Ladd

To my love Lida

Abstract

In this thesis, I attempt to gain a deeper insight into the details of the human mirror neuron system by finding the effective connectivity of its central regions and simulating them with computational modeling. To achieve this aim, I have used the measured functional magnetic resonance imaging (fMRI) data for healthy participants while performing key tasks of social cognition (imitation of emotional faces, theory of mind, and empathy). Using a self-developed firing-rate-based extension of the statistical analysis procedure dynamic causal modeling (DCM), I was able to determine the effective network structure of the human mirror neuron system from the fMRI data and compare it between the different tasks of social cognition. In particular, far more complex processing occurs in imitation than in the other two tasks, which seems plausible given that imitation involves matching observed and self-performed emotional expression. Furthermore, we were able to show that the extended DCM procedure allows for significantly better model evidence, both for our novel data and for previously established datasets from other research groups. Thus, in addition to the substantive insight, this project has provided an important methodological advance for all users of the widely used DCM procedure.

Furthermore, the main regions of the mirror neuron system are modeled in detail by a modification of an existing, completely data-driven spiking network model of the prefrontal cortex. Here, I use the estimated parameters of the modified DCM to match the time series of the simulated and observed data. This two-stage approach allows both to account for the neural mass signals measured by fMRI and assess the fine-scale temporal dynamics of the local dynamics, and thus derive predictions about the physiological details that cannot be obtained from non-invasive recordings alone.

Keywords: mirror neuron system, fMRI, dynamic causal modeling, spiking network model, social cognition.

Kurzzusammenfassung

In dieser Arbeit versuche ich, einen tieferen Einblick in die Details des menschlichen Spiegelneuronensystems zu gewinnen, indem ich die effektive Konnektivität seiner zentralen Regionen finde und sie mit Computermodellierung simuliere. Um dieses Ziel zu erreichen, habe ich die gemessenen Daten der funktionellen Magnetresonanztomographie (fMRT) von gesunden Teilnehmern verwendet, die Schlüsselaufgaben der sozialen Kognition (Imitation von emotionalen Gesichtern, Theory of Mind und Empathie) durchführten. Mit Hilfe einer selbstentwickelten Feuerraten-basierten Erweiterung des statistischen Analyseverfahrens dynamic causal modeling (DCM) konnte ich aus den fMRT-Daten die effektive Netzwerkstruktur des menschlichen Spiegelneuronensystems bestimmen und zwischen den verschiedenen Aufgaben der sozialen Kognition vergleichen. Insbesondere findet bei der Imitation eine weitaus komplexere Verarbeitung statt als bei den beiden anderen Aufgaben, was plausibel erscheint, wenn man bedenkt, dass bei der Imitation ein Abgleich von beobachtetem und selbst ausgeführtem emotionalem Ausdruck stattfindet. Darüber hinaus konnten wir zeigen, dass das erweiterte DCM-Verfahren eine signifikant bessere Modellevidenz ermöglicht, sowohl für unsere neuartigen Daten als auch für bereits etablierte Datensätze anderer Forschergruppen. Somit hat dieses Projekt neben den inhaltlichen Erkenntnissen auch einen wichtigen methodischen Fortschritt für alle Anwender des weit verbreiteten DCM-Verfahrens gebracht.

Darüber hinaus werden die Hauptregionen des Spiegelneuronensystems durch eine Modifikation eines bestehenden, vollständig datengetriebenen Spiking Netzwerkmodells des präfrontalen Kortex im Detail modelliert. Hier verwende ich die geschätzten Parameter des modifizierten DCM, um die Zeitreihen der simulierten und beobachteten Daten abzugleichen. Dieser zweistufige Ansatz erlaubt es, sowohl die durch fMRT gemessenen neuronalen Massensignale zu berücksichtigen als auch die feinskalige zeitliche Dynamik der lokalen Dynamik zu bewerten und so Vorhersagen über die physiologischen Details abzuleiten, die aus nicht-invasiven Aufzeichnungen allein nicht gewonnen werden können.

Schlagworte: Spiegelneuronensystem, fMRT, dynamische kausale Modellierung, Spiking-Netzwerk-Modell, soziale Kognition.

Contents

List of Abbreviations & Symbols **13**

1. Introduction **17**

2. Dynamic causal modeling for fMRI **21**
2.1. Functional magnetic resonance imaging 21
2.2. Effective connectivity 22
2.3. Dynamic causal modeling 22
2.3.1. Bayesian estimation 24
2.3.2. Variational Bayes 27
2.4. Bayesian model comparison 30
2.4.1. Bayes factors 30
2.4.2. Random effects BMS for heterogeneous groups 31
2.4.3. Comparing model families or family-level inference 32
2.4.4. Bayesian model averaging 32
2.4.5. Protected exceedance probability 33

3. Dynamic causal modeling for fMRI with Wilson-Cowan-based neuronal equations **35**
3.1. Introduction 35
3.2. Materials & Methods 37
3.2.1. Wilson-Cowan equations for DCM 37
3.2.2. Two-state DCM with Wilson-Cowan equations 40
3.2.3. Bayesian model selection 41
3.2.4. Data sets 42
3.3. Results 45
3.3.1. Validation of established data set 45
3.3.2. Validation of novel empirical data 45
3.3.3. Validation of synthetic data 49
3.4. Discussion 51
3.4.1. Comparison to other DCM extensions 53
3.4.2. The role of the nonlinearity 54
3.4.3. Implications for the mirror neuron system 55
3.4.4. Constraints of local neural networks 56

4. Effective connectivity of the human mirror neuron system during social cognition **57**
4.1. Introduction 57
4.2. Materials & Methods 60
4.2.1. Data sets 60

4.2.2. DCM 65
4.3. Results 68
4.4. Discussion 71
4.4.1. Limitations and outlook 73

5. Detailed spiking network model of the human mirror neuron system 75
5.1. Spiking network model of the prefrontal cortex 75
5.2. Implementation of the DCM model on the spiking network model . 76

6. General discussion of possible extensions & Outlook 87
6.1. Dopaminergic modulation of the spiking network model 87
6.2. PEB analysis on subjects according to their genotypes 90
6.3. Concluding remarks 93

Appendices 97

A. Maximum likelihood estimation 99

B. Fisher scoring algorithm 101

Bibliography 103

List of Figures 117

List of Tables 119

CV & Publications 121

Acknowledgements 123

List of Abbreviations & Symbols

Abbreviations

AdEx Adaptive Exponential Integrate-and-Fire Model

BA44 Brodmann Area 44

BF Bayes Factor

BMA Bayesian Model Averaging

BMC Bayesian Model Comparison

BMR Bayesian Model Reduction

BMS Bayesian Model Selection

BOLD Blood Oxygen Level Dependent

DCM Dynamic Causal Modeling

EEG Electroencephalogram

EM Expectation Maximization

FFX Fixed Effects

fMRI Functional Magnetic Resonance Imaging

GBF Group Bayes Factor

HRF Hemodynamic Response Function

IFG Inferior Frontal Gyrus

IN Interneuron

IPL Inferior Parietal Lobe

KL Kullback-Leibler

MEG Magnetoencephalography

MLE Maximum Likelihood Estimate

MNS Mirror Neuron System

PC Pyramidal Cell

PEB Parametric Empirical Bayes

PFC Prefrontal Cortex

RFX Random Effects

ROI Region Of Interest

SimpAdEx Simplified AdEx

SNR Signal to Noise Ratio

SPM Statistical Parametric Mapping

STS Superior Temporal Sulcus

TMS Transcranial Magnetic Stimulation

ToM Theory of Mind

TPJ Temporoparietal Junction

VB Variational Bayes

VBA Variational Bayesian Analysis

W-C Wilson-Cowan

Symbols

A directed or intrinsic connectivity among the regions

α Grubb's vessel stiffness exponent

B^j changes in the connectivity influenced by the inputs u_j (experimental manipulations)

C elements of external input u

f blood flow

F free-energy

γ rate constant for flow-dependent elimination

κ rate constant of the vasodilatory signal decay

q deoxyhemoglobin content of the blood

ρ capillary resting net oxygen extraction

s vasodilatory signal

τ transit time

θ^c set of parameters at the neuronal level

v blood volume

w synaptic weight

z neuronal activity

1. Introduction

In the African Bantu language, the word Ubuntu means ‘I am, because you are’. This word is part of a phrase that indicates a person is a person through other people. Neuroscientists have a similar opinion: regular social interactions form humans, and their brains and minds [1]. Social interaction is an exchange between two or more individuals and is a critical building block of society. It involves communication in all its forms, as simple as a word or a threatening gesture or as complex as organizing a social movement or forming a family.

How do we know about each other’s feelings and intentions? Apart from verbal communications, people can also use nonverbal communications, like gestures, facial expressions, and postures, to share their feelings and intentions. The understanding of nonverbal communications is supported by the mirror neuron system (MNS). The human MNS is considered the neural basis of our interpersonal understanding [2]. These neurons were first found in the monkey motor cortex and activated both when the monkey moves (execution) and when it observes (observation) a similar movement (Fig. 1.1A and B; [3]). This is the defining property of mirror neurons that separates them from other ‘motor’ or ‘sensory’ neurons, which are activated with either execution or observation, but not both. For humans, the assumption was derived that we recognize the emotions and intentions of other persons by simulating their motor state in our motor system [4]. The theory of embodied simulation assumes that this is an automatic process that goes without cognitive exertion. In certain mental disorders such as autism, however, disturbances in the MNS would be present, resulting in deficits in social cognition as well as in emotion recognition and the theory of mind (ToM) [5].

In healthy subjects, direct detection of the mirror neuron activity by single-cell recordings is excluded. Therefore, scientists use other non-invasive measurement methods in humans, in which mass signals from neuronal populations (in the case of the electroencephalogram, EEG) or the changes in oxygen consumption by neuronal activity (as in functional magnetic resonance tomography, fMRI), are detected. Apart from experimental approaches, computational modeling and simulations are also essential tools for studies on cognitive science. Some computational models of the MNS also exist [7, 8], both for scientific understanding and technical applications, such as in robotics. These models are strongly task (mostly a grasping movement) oriented (top-down approach), and in some cases, contain detailed modules for object recognition and execution of the movement. However, approaches of strictly data-based (bottom-up) modeling are not yet available.

Following the theoretical considerations on the role of the MNS, activation in Brodmann area 44 (BA44) of the inferior frontal gyrus and in the inferior parietal lobe (IPL) as well as in the superior temporal sulcus (STS) is found in humans utilizing fMRI both when performing and observing neutral actions and emotional facial expressions [9]. These areas are seen as being homologous to the areas of the

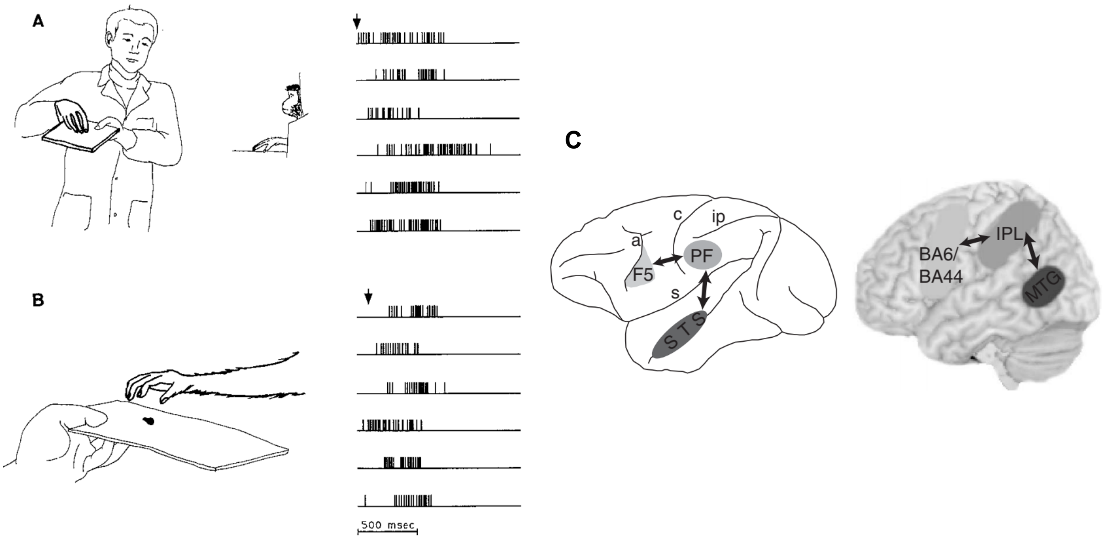


Figure 1.1.: The mirror neuron system. A) The single-cell derivation in the monkey during the observation of a grasping movement, B) Individual cell derivation in the monkey during the execution of a grasping movement, C) Areas of the MNS in the monkey (left) and humans (right). A) and B) from [3], C) from [6].

MNS in the monkey (Fig. 1.1C [6]). Studies in clinical populations find reduced activation in the inferior frontal gyrus (IFG) in autism [10] and other mental disorders such as borderline personality disorder [11] and psychopathy [12]. These results suggest that in these areas, a simulation of the motor expression of a person takes place, and this simulation is disturbed by mental disorders. However, they do not provide direct evidence[1] that the mirror neurons form the neural basis of such a simulation mechanism. The knowledge about the physiology as well as the network structure of the MNS in humans is still scarce, despite the extensive indirect findings.

This project aims to broaden the knowledge about the human MNS by finding the effective connectivity between the activated regions of the MNS and detailed mathematical modeling. Three experiments on healthy subjects have been carried out by my colleagues, in which imitation, empathy, and theory of mind (ToM) using emotional facial expressions have been analyzed and the indirect detection of the mirror neuron activity was achieved by fMRI measurements [14, 15].

Theoretical modeling of the involved cell assemblies is another way to learn more about the human MNS without directly measuring individual neurons' activity. In this thesis, I have used the mathematical methods of theoretical physics and computational neuroscience to simulate the data underlying neuronal cell networks. On the one hand, this allows identification of the global network structure and its possible modulation by neurotransmitters or different tasks. On the other hand, the dynamics of these networks can be comprehended by computer simulations of biologically detailed local network models. Of particular importance for our project is the possibility of adapting and interpreting the parameters of these networks based on neuropsychological data. Thus, the combination of experimental

---

[1] There is only one study directly measuring neurons with mirror property [13].

neuroscientific psychology and computational neuroscience makes it possible for the first time to infer the physiology and dynamics of the cell networks, which are subject to a simulation mechanism of social information.

The fMRI measurements and first-level data analysis were done by my colleagues and the common activation for imitation, empathy, and affective ToM across and within participants in STS, IPL, and BA44 is demonstrated [14, 15]. The main focus of this thesis is the theoretical part of the project. A two-step model approach is developed, which allows conclusions to be drawn from the experimental data on the anatomy and physiology of the MNS. In the first step, I use a modified version of the dynamic causal modeling (DCM) [16] for parameter estimation based on the fMRI data and then use the Bayesian model selection (BMS) [17, 18] for finding the global network of mirror neurons. The second component is a detailed network model of local cell clusters, which simulates the exact temporal dynamics of individual nerve cells [19].

In the first chapter of my thesis, I present the theory of the sophisticated methods, including dynamic causal modeling (DCM) and Bayesian model selection (BMS), which are used in this project. For the next chapter, I present the first part of the project, consisting of an investigation of the global network of human mirror neurons for which I first integrate the Wilson-Cowan (W-C) model [20, 21] into the classical DCM formulation. Thus, it is possible to adapt the parameters of the model to the fMRI data using this sophisticated statistical analysis method. These parameters provide information about the effective connectivity between the active brain regions. Through the W-C model, the global physiological properties of each region are now included in the estimation. Using the DCM method, we can determine the global network structure for the imitation block-designed task for the first part and three event-related tasks (Imitation, Empathy, and Theory of Mind) in the second part. Furthermore, I show that the modified DCM analysis allows a significantly better fit for our empirical data and previously established data set. Therefore, our results can provide an important methodological advance for all users of the widespread DCM method.

The new nonlinear modification of the DCM framework based on the W-C DCM is tested on three different datasets, and its superiority in fitting these data compared with the standard bilinear model is shown. First, I use an established dataset that has been widely used as a test case for DCM [22]. Second, the dynamics of the human MNS using our novel data set are investigated. Finally, I generate synthetic data with different signal-to-noise ratios (SNRs) based on the novel data to examine how W-C DCM performs when the ground truth is known.

Next, I investigate the effective connectivity of the human MNS for the three different social-cognitive processes using the modified version of DCM for parameter estimation. I use a stepwise family level inference to find the best fitting effective connectivity model among STS, IPL, and BA44 with the hypothesis that the external input enters in STS [23]. I try all possible models (540 models) and significantly decrease the model space to do the final BMS on a smaller number of models. I then use the Bayesian model averaging (BMA) [24] to overcome the ambiguity raise from BMS and make our inference based on the parameters that contribute significantly to the model space.

In the final phase of this thesis, I investigate more about the human MNS by simulating it using a highly detailed network model of the prefrontal cortex. All neurons and synapses parameters of this model are determined by anatomical and in vitro electrophysiological data. This spiking network model has previously been shown to statistically reproduce a wide range of measures from in vivo prefrontal data in rodents [19]. In this study, the model is adjusted to the fMRI data using the global connectivity, which was inferred from W-C DCM. The input-output functions of the neurons in the firing-rate model are matched with the predicted data from DCM analysis by comparing the resulting outputs, thus realizing the transfer from the macro- to the micro-level. The neurons of brain regions in the simulated data are connected according to the global network, which was derived from DCM analysis and BMA. The simulated MNS can be used to predict task performance of social cognition and to be validated on a different dataset.

2. Dynamic causal modeling for fMRI

This chapter presents the conceptual and mathematical foundations behind dynamic causal modeling (DCM) for fMRI data. In the first step, I give a brief description of fMRI data and effective connectivity and then, in more detail, describe the DCM framework used for parameter estimation and model selection.

For the estimation section, I have described the expectation-maximization (EM) algorithm and variational Bayes (VB), in which both methods are used to find the posterior distribution of the parameters. They are the most important approaches used in the DCM framework and are iterative procedures that successively converge on optimum parameter values. EM algorithm computes point estimates (mean and variance) of the posterior distribution of parameters, and VB computes estimates of the complete distribution of the posterior density [25].

For the model selection section, I present the basic concepts and formulations of the Bayesian model comparison (BMC), including fixed effects and random effects for heterogeneous groups. Furthermore, a short introduction to the family-level inference and Bayesian model averaging (BMA) is given at the end of this chapter.

2.1. Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) is a non-invasive technique that measures brain activity by identifying changes in levels of oxygenated blood in different regions of the brain using a method called blood-oxygen-level-dependent (BOLD) contrast. During an fMRI experiment, a series of brain images are acquired across time while subjects perform a set of tasks. The signal changes in reaction to these tasks are described by hemodynamic response function (HRF), which represents changes in the fMRI signal triggered by neuronal activity [26]. Hemodynamics is the dynamics of blood flow, and the coupling between neuronal activity and blood flow is also referred to as neurovascular coupling [27].

As mentioned above, during the fMRI experiments, we measure the same brain volume multiple times across time; each of these brain volumes consists of roughly 100,000 ‘voxels’, which are cubic volumes that span the three-dimensional space of the brain. Each voxel corresponds to a spatial location and has an intensity. During an experiment, several hundred images are acquired ($\sim$ one every 2s). To analyze the fMRI data, we can extract information from a single voxel and study what is going on and how the intensity is changing across a specific voxel. By doing this, we can extract the time series of these intensities. This shows that fMRI data analysis is fundamentally a time series analysis as every data from every voxel is a time series [28].

There are three main goals of fMRI data analysis. The first one is localization, the process of determining which regions of the brain are active during a specific task. The second thing we can look at is connectivity which determines how different brain regions are connected. The third one is prediction, in which we can use a person's brain activity to predict perceptions, behavior, or health status [28].

Two popular varieties of connectivity are functional and effective connectivity. Functional connectivity describes correlations among different brain regions across time, and effective connectivity rests on the causal influences that neural units exert over one another [29]. In particular, I focus on effective connectivity among activated regions in this project.

2.2. Effective connectivity

Effective connectivity is defined as the directed influence of one brain region on the physiological activity recorded in other brain regions [29]. It claims to make statements about causal effects among tasks and regions. In effective connectivity methods we typically make anatomically assumptions and restricts inference to networks comprising of a number of pre-selected regions of interest[1] (ROIs).

A goal of effective connectivity analysis is to make statements about causal effects among tasks and regions and usually computed using several methods, including DCM, Granger causality (GC), and structural equation modeling (SEM) [29,30], among which DCM is the focus of this thesis.

2.3. Dynamic causal modeling

Dynamic causal modeling is a non-linear system identification procedure using the Bayesian approach to infer the effective connectivity (and other parameters) of deterministic multiple-input multiple-output dynamic systems from brain imaging data, e.g., fMRI, Magnetoencephalography (MEG), or Electroencephalography (EEG) [16,31–33]. The general concept describes the hidden neuronal states using differential equations and combines them with a forward model to produce measured data. The hidden state and the forward model differ for fMRI and EEG/MEG data.

This study focuses on fMRI data, which has a relatively poor temporal resolution (in seconds) but high spatial resolution (typical voxel size is $3mm \times 3mm \times 3mm$). In the first step, I introduce the neuronal state equations in the DCM framework, which show the effective connectivity of the brain regions, then briefly explain the translation of neuronal activity into hemodynamic responses with the differential equations that constitute the hemodynamic model for each region [16].

This dynamical system describes the hidden state:

$$\dot{z}_t = F(z_t, u_t, \theta), \tag{2.1}$$

---

[1] Regions of interest are samples within a data set that show a particular response and are identified for a specific purpose.

where z_t is the neural activity, and F is some nonlinear differentiable function - called the neural response - describing how the rate of neural activity in all brain regions depends on a set of parameters θ and external input u. The parameters define the dynamics of the system and we need their posterior density for inference.

Using Taylor expansion for Eq. 2.1 around $z = 0$ and $u = 0$, defining the baseline neural response $F(0, 0) = 0$, and keeping only the first-order, we would have a bilinear form of the Eq. 2.1:

$$\begin{aligned}
F(z, u, \theta) = \dot{z} &\approx \left(A + \sum_{j=1}^{m} u_j B^j\right) z + Cu \\
A &= \frac{\partial F}{\partial z} = \frac{\partial \dot{z}}{\partial z}|_{u=0} \\
B^j &= \frac{\partial^2 F}{\partial z \partial u_j} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z} \\
C &= \frac{\partial F}{\partial u}|_{z=0},
\end{aligned} \tag{2.2}$$

where the matrix A describes the directed or intrinsic connectivity among the regions in the absence of input. The matrix B^j embodies the changes in the connectivity influenced by the inputs u_j (experimental manipulations) and finally the matrix C represents the elements of external input u on each area. With these matrices, we can specify the models, how different regions are connected to each other. DCM estimates these parameters $\theta^c = \{A, B^j, C\}$ at the neuronal level.

At the next level, in each region, the neuronal activities mapped to the measured data with the hemodynamic model which is modeled by four state equations with 5 area-specific parameters [16, 34]. The neuronal activity z_i is an input in the vasodilatory signal s_i which is subject to autoregulatory feedback f_i by blood flow. This blood inflow f_i influences simultaneously the blood volume v_i and deoxyhemoglobin content q_i:

$$\begin{aligned}
\dot{s}_i &= z_i - \kappa_i s_i - \gamma_i(f_i - 1) \\
\dot{f}_i &= s_i \\
\tau_i \dot{v}_i &= f_i - v_i^{1/\alpha} \\
\tau_i \dot{q}_i &= f_i(1 - (1 - \rho_i)^{1/f_i})/\rho_i - v_i^{1/\alpha} q_i / v_i,
\end{aligned} \tag{2.3}$$

in which κ is the rate constant of the vasodilatory signal decay, γ is the rate constant for flow-dependent elimination, τ transit time, α Grubb's vessel stiffness exponent and ρ is the capillary resting net oxygen extraction. These are the biophysical parameters $\theta^h = \{\kappa, \gamma, \tau, \alpha, \rho\}$ which are estimated with DCM. By integrating the state equations of v and q we can form the predicted BOLD signal equation:

$$\begin{aligned} y_i &= g(q_i, v_i) \\ &= V_o(k_1(1-q_i) + k_2(1-q_i/v_i) + k_3(1-v_i)) \\ k_1 &= 7\rho_i \\ k_2 &= 2 \\ k_3 &= 2\rho_i - 0.2, \end{aligned} \tag{2.4}$$

where $V_o = 0.02$ is resting blood volume fraction (see [16, 34] for more details and a complete description). In the following section, I present the mathematical description to estimate the parameters of the hidden state in the DCM for fMRI data.

2.3.1. Bayesian estimation

This section describes the expectation-maximization (EM) algorithm to estimate the hidden state introduced above in the DCM framework [16, 35, 36]. EM algorithm seeks to find the maximum likelihood estimate (MLE) of the parameters (see appendix A) by iteratively applying two steps (E-step and M-step). As a general definition, in E-step, the distribution of the hidden parameters is estimated given the known values for the observed dataset. In M-step, the parameters are re-estimated to be those with maximum likelihood, assuming the distribution found in the E-step is correct. These two steps are repeated until convergence [37]. In the DCM framework, for the E-step, the conditional expectations and covariances of the parameters are estimated, and in the M-step, the maximum likelihood of the hyperparameters (parameters of the error covariance) is estimated to be used in the next E-step [16, 35].

The forward model is specified with combining the neuronal and hemodynamic states introduced above $x = \{z, s, f, v, q\}$:

$$\begin{aligned} \dot{x} &= f\,(x, u, \theta) \\ y &= \lambda(x) \\ &= h(u, \theta). \end{aligned} \tag{2.5}$$

The predicted response $h(u, \theta)$ is the integration of the state equations with any set of parameters $\theta = \{\theta^c, \theta^h\}$ and inputs u (h can be any system, independent of DCM). By adding error ϵ to the forward model, the observation model is achieved of the form:

$$y = h(u, \theta) + \epsilon. \tag{2.6}$$

In this formulation, it is assumed that the parameters and error are Gaussian ($\epsilon \sim N\{0, C_\epsilon\}$, C_ϵ is the covariance of the observation error). This assumption is motivated by the central limit theorem arising from the averaging implicit in most imaging applications [38]. Bayesian inference is based on the conditional probability of the parameters given the data $p(\theta|y)$ (posterior probability) and is proportional to the likelihood of obtaining the data given θ, times the prior probability of θ (Bayes' theorem):

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{2.7}$$

These probabilities are Gaussian and specified in terms of their expectation η and covariance C values and the problem reduces to finding the conditional expectation $\eta_{\theta|y}$ and covariance $C_{\theta|y}$ of the posterior probability.

To approximate the likelihood, we expand $h(u, \theta)$ in Eq. 2.6 about the conditional mean of the parameters, $\eta_{\theta|y}$, given the data y:

$$\begin{aligned} h\left(u, \theta\right) &\approx h\left(u, \eta_{\theta|y}^{(i)}\right) + J \times \left(\theta - \eta_{\theta|y}^{(i)}\right) \\ J &= \frac{\partial h(u, \eta_{\theta|y}^{(i)})}{\partial \theta}, \end{aligned} \tag{2.8}$$

and use Eq. 2.8 in Eq. 2.6:

$$y - h\left(u, \eta_{\theta|y}^{(i)}\right) \approx J\Delta\theta + \epsilon, \tag{2.9}$$

where $\Delta\theta = \theta - \eta_{\theta|y}^{(i)}$. Let $r = y - h(u, \eta_{\theta|y}^{(i)})$ such that $\epsilon \approx r - J\Delta\theta$. In this way, the likelihood and prior probabilities can be proved under Gaussian assumptions:

$$\begin{aligned} p(y \mid \theta) &\propto \exp\{-\frac{1}{2}(r - J\Delta\theta)^T C_\epsilon^{-1}(r - J\Delta\theta)\} \\ p(\theta) &\propto \exp\{-\frac{1}{2}(\theta - \eta_\theta)^T C_\theta^{-1}(\theta - \eta_\theta)\}. \end{aligned} \tag{2.10}$$

Using the Eq. 2.7 and by substituting the Eq. 2.10 in it, the log posterior defined as below:

$$\begin{aligned} I &= \ln p(\theta|y) = \ln p(y|\theta) + \ln p(\theta) \\ &= -\frac{1}{2}\{(r - J\Delta\theta)^T C_\epsilon^{-1}(r - J\Delta\theta) + (\theta - \eta_\theta)^T C_\theta^{-1}(\theta - \eta_\theta)\}. \end{aligned} \tag{2.11}$$

Following the Fisher scoring algorithm (see appendix B) by taking the gradients with respect to the parameters we can provide the basis for recursive estimation of the conditional mean and covariance:

$$\eta_{\theta|y}^{(i+1)} = \eta_{\theta|y}^{(i)} - \left\langle \frac{\partial^2 I}{\partial \theta^2} \right\rangle^{-1} \frac{\partial I}{\partial \theta}(\eta_{\theta|y}^{(i)}), \tag{2.12}$$

with the initial value $\eta_{\theta|y}^{(1)} = \eta_\theta$, where

$$\begin{aligned} \frac{\partial I}{\partial \theta}(\eta_{\theta|y}^{(i)}) &= J^T C_\epsilon^{-1} r + C_\theta^{-1}(\eta_\theta - \eta_{\theta|y}^{(i)}) \\ -\left\langle \frac{\partial^2 I}{\partial \theta^2} \right\rangle &= J^T C_\epsilon^{-1} J + C_\theta^{-1} = C_{\theta|y}^{-1}. \end{aligned} \tag{2.13}$$

This was the E-step of the EM algorithm in which the conditional expectation and covariance of the parameters were estimated under Gaussian assumptions.

In the case of unknown error covariance, it can be estimated through some hyperparameters λ_j, where $C_\epsilon = \sum \lambda_j Q_j$, $Q_j = \frac{\partial C_\epsilon}{\partial \lambda_j}$ represents a basis set for the covariance matrices corresponds to the partial derivatives of the covariances with respect to the hyperparameters. Here, the basis sets (also called error covariance components) are known[2], and the hyperparameters should be estimated (see [35] and chapters 8 and 10 in [39] for more details). In the M-step, these hyperparameters of the error covariance are estimated by integrating the parameters out of the log-likelihood function using their conditional distribution from the E-step. The estimated hyperparameters are then used in the next E-step [35, 36]. As the log-likelihood depends on the unknown parameters, we can integrate them out by using the approximate conditional distribution $q(\theta)$ which is a distribution over model parameters [40]. The goal is to minimize the observed data $p(y \mid \lambda)$ conditional on some hyperparameters in the presence of unobserved parameters θ, which is equivalent to maximize the log-likelihood (note that there are no priors on the hyperparameters):

$$\begin{aligned} \ln p(y \mid \lambda; u) &= \ln \int q(\theta) \frac{p(\theta, y \mid \lambda; u)}{q(\theta)} d\theta \geq \\ F(q, \lambda) &= \int q(\theta) \ln \frac{p(\theta, y \mid \lambda; u)}{q(\theta)} d\theta \\ &= \int q(\theta) \ln p(\theta, y \mid \lambda; u) d\theta - \int q(\theta) \ln q(\theta) d\theta. \end{aligned} \tag{2.14}$$

This equation is based on Jensen's inequality which follows from the concavity of the log function. F corresponds to the negative free energy[3] in statistical physics and comprises two terms related to the energy (first term) and entropy (second term) [40]. By substituting Eqs. 2.12 and 2.13 from the E-step into F and taking the gradients with respect to the hyperparameters and using the fisher scoring scheme again, the following terms can be derived:

$$\lambda^{i+1} = \lambda^i - \left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle^{-1} \frac{\partial F}{\partial \lambda}(\lambda), \tag{2.15}$$

with the initial value $\lambda^1 = \lambda^0$, where

$$\begin{aligned} \frac{\partial F}{\partial \lambda_i}(\lambda) &= \frac{1}{2} tr\{PQ_i\} - \frac{1}{2} r^T P^T Q_i P r \\ -\left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle_{ij} &= \frac{1}{2} tr\{PQ_i P Q_j\}, \end{aligned} \tag{2.16}$$

and $P = C_\epsilon^{-1} - C_\epsilon^{-1} J C_{\theta|y}^{-1} J^T C_\epsilon^{-1}$. In this way, hyperparameters are estimated. In brief, the E-step computes the conditional mean and covariance of the unobserved

---

[2] In fMRI data, the basis sets have two components Q_1 and Q_2. Q_1 is the identity matrix ($Q_1 = I_N$) and:

$$Q_{2_{ij}} = \begin{cases} e^{-|i-j|} & i \neq j \\ 0 & i = j. \end{cases}$$

[3] This is also related to the free energy principle, which states that systems change to decrease their free energy and free energy minimization is mandatory in biological systems. It rests upon the fact that self-organizing biological agents resist a tendency to disorder and, therefore, minimize their sensory states' entropy. In thermodynamics, free energy is a measure of the amount of work that can be extracted from a system, and at a constant temperature, it is minimized at equilibrium [41, 42].

parameters to enable the M-step to optimize the hyperparameters in a maximum likelihood sense. These new hyperparameters re-enter into the estimation of the conditional distribution and so on until convergence.

The EM algorithm is robust and has found multiple applications in data analysis, ranging from ReML (restricted maximum likelihood) of serial correlations in fMRI to hyperparameter estimation in hierarchical observation models using empirical Bayes [16, 35, 36].

It is also important to note, the parameter ascent on the log posterior I, Eq. 2.11 in the E-step is closely related to an ascent on the negative free energy F used for the hyperparameters in the M-step, with exact equivalence when $q(\theta)$ is deterministic (fixed value). Hence for both EM steps, the same function (F) can be used. This can be seen if we write the Eq. 2.14 as below:

$$\begin{aligned} F &= \int q(\theta) \ln \frac{p(y \mid \theta, \lambda; u) p(\theta)}{q(\theta)} d\theta \\ &= \int q(\theta) \ln p(y \mid \theta, \lambda; u) d\theta - \int q(\theta) \ln \frac{q(\theta)}{p(\theta)} d\theta \\ &= \left\langle \ln p(y \mid \theta, \lambda; u) \right\rangle_q - KL(q(\theta), p(\theta)). \end{aligned} \tag{2.17}$$

F comprises the expected log-likelihood under $q(\theta)$ (first term) and describes the accuracy of the model in fitting the data (i.e., the goodness of fit). The second term is the Kullback-Leibler (KL) divergence between the conditional and prior densities. It is a measure of the difference between these two distributions and is always a non-negative value. The second term is sensitive to the number of parameters and the form of the densities and can be regarded as a measure of model complexity. These two terms illustrate F has properties, which are required for model selection. I write more about these properties in the Bayesian model comparison section. In the following section, I introduce the variational Bayes (VB) method, which is also used in the DCM framework for estimating the hidden state and the model evidence.

2.3.2. Variational Bayes

In variational Bayes (VB) approach [39, 43], the true posterior distribution $p(\theta|y)$ over a set of unobserved parameters θ given observed data y is approximated by a density $q(\theta)$, which is described by a few parameters (e.g. its mean and covariance). This is often done by assuming a particular form (e. g. Gaussian distributions) for q and then optimizing its sufficient statistics.

We can measure this optimality by KL divergence which is a quantity that expresses the dissimilarity between $q(\theta)$ and true posterior. If the two distributions are identical, KL divergence is zero and if they are different from each other, it gives a positive value. In this way, inference is performed by selecting the distribution $q(\theta)$ that minimizes KL divergence.

We can decompose the log evidence into negative free energy (Fig. 2.1) which depends on approximate posterior q and observed data, and a KL divergence between approximate posterior and true posterior. Given a probabilistic model of some data, we can write the log of evidence according to the Bayes' theorem 2.24 and marginalization 2.25:

$$
\begin{aligned}
\ln p(y) &= \int q(\theta) \ln p(y) d\theta \\
&= \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{p(\theta|y)} d\theta \\
&= \int q(\theta) \ln \frac{p(y,\theta)q(\theta)}{q(\theta)p(\theta|y)} d\theta \\
&= \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} d\theta + \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta \\
&= F(q,y) + KL[q(\theta)||p(\theta|y)],
\end{aligned}
\tag{2.18}
$$

in which F is the negative free energy (Eq. 2.17) and KL divergence is a non-negative value (due to Gibbs inequality). This is not to be confused with the KL in Eq. 2.17 which was between the approximate posterior and the prior. Here, we can not compute F as we don't know the true posterior. All we can do is to maximize $F(q,y)$, which is equivalent to minimizing $KL[q||p]$. By choosing better parameters for approximate posterior $q(\theta)$, we will get closer to the true log evidence, and accordingly, the KL divergence is minimized. In this way, at the same time, we will have the best approximate to our posterior and an approximation to the log evidence.

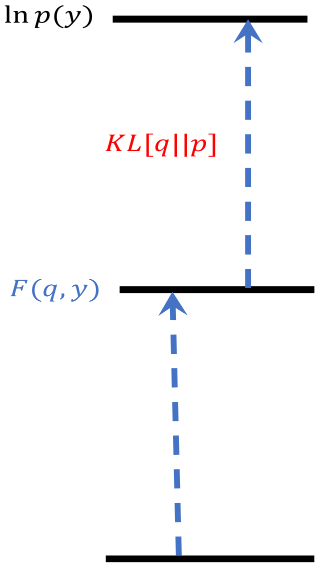


Figure 2.1.: The negative free energy, F, provides a lower bound on the log model evidence. When $F(q,y)$ is maximized, $q(\theta)$ is our best estimate of the posterior $p(\theta|y)$.

One way to make the maximization of F easier is using the mean-field approximation [44, 45]. The basic idea behind this assumption is to approximate a very high dimensional probability distribution with the product of a number of simple densities, and in this way, the dynamics of one node are determined by the mean or average activity in another [46, 47]. In this approach, we factorize the approximation density over groups of parameters:

$$q(\theta) = \prod_i q_i(\theta_i), \tag{2.19}$$

where θ_i is the *ith* group of parameters. It can also be written as below:

$$q(\theta) = q_i(\theta_i) q_j(\theta_j),$$

where θ_j denotes all parameters not in the *ith* group. Here we substitute this approximation in F:

$$\begin{aligned}
F &= \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{q(\theta)}\, d\theta = \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)}\, d\theta \\
&= \iint q_i(\theta_i) q_j(\theta_j) \ln \left[\frac{p(y,\theta)}{q_i(\theta_i) q_j(\theta_j)} \right] d\theta_i \, d\theta_j \\
&= \int q_i(\theta_i) \left[\int q_j(\theta_j) \ln p(y,\theta) d\theta_j \right] d\theta_i - \int q_i(\theta_i) \ln q_i(\theta_i)\, d\theta_i + C \\
&= \int q_i(\theta_i) I_i(\theta_i)\, d\theta_i - \int q_i(\theta_i) \ln q_i(\theta_i)\, d\theta_i + C,
\end{aligned} \tag{2.20}$$

in which, constant C contains terms not dependent on $q_i(\theta_i)$. By defining

$$I_i(\theta_i) = \int q_j(\theta_j) \ln p(y,\theta) d\theta_j,$$

and writing $I_i(\theta_i) = \ln \exp I_i(\theta_i)$ we have:

$$\begin{aligned}
F &= \int q_i(\theta_i) \ln \left[\frac{\exp(I_i(\theta_i))}{q_i(\theta_i)} \right] d\theta_i + C \\
&= KL[q_i(\theta_i) || \exp(I_i(\theta_i))] + C.
\end{aligned} \tag{2.21}$$

The above term is minimized (maximizing F, the negative free energy, is the same as minimizing $-F$, the free energy) when:

$$q_i(\theta_i) = \frac{\exp(I_i(\theta_i))}{Z_i}, \tag{2.22}$$

where Z_i is a normalization constant. With the same process, we have the same equation for $q_j(\theta_j)$, and in this way, we approximated the approximation density as below:

$$\begin{aligned}
q_i(\theta_i) &\propto \exp(I_i(\theta_i)) = \exp\left[\langle \ln p(y|\theta) + \ln p(\theta) \rangle_{q_j(\theta_j)} \right] \\
q_j(\theta_j) &\propto \exp(I_j(\theta_j)) = \exp\left[\langle \ln p(y|\theta) + \ln p(\theta) \rangle_{q_i(\theta_i)} \right].
\end{aligned} \tag{2.23}$$

Iterating these updates between $q_i(\theta_i)$ and $q_j(\theta_j)$ with any initial values until convergence provides a simple deterministic optimization of the free energy with respect to the approximate posterior density.

2.4. Bayesian model comparison

Having specified a forward model, one can simulate data under different models (e.g., with varying connectivity patterns, as models differ in connectivity) and ask which simulation best characterizes the observed data. Practically, this is done in two stages: Bayesian model inversion and comparison, respectively, which I will introduce briefly.

According to Bayes' theorem, the posterior distribution equals the likelihood times prior divided by model evidence:

$$p(\theta|y,m) = \frac{p(y|\theta,m)p(\theta,m)}{p(y,m)}. \tag{2.24}$$

The model evidence is the probability of obtaining observed data y given a particular model m. To evaluate the evidence for a given model m (defined by the pattern of connectivity), one needs to integrate out the unknown parameters, which means model inversion is usually needed before model comparison:

$$p(y,m) = \int p(y|\theta,m)p(\theta,m)d\theta. \tag{2.25}$$

This averaging or marginalisation is why model evidence is sometimes called the marginal likelihood of a model.

Using the VB approximations this integration can be computed (see above). According to the equations 2.17 and 2.18, we can approximate the log model evidence with the negative free energy, when the KL divergence between the true and approximate conditional density is suppressed, therefore we have:

$$\ln p(y,m) \approx F. \tag{2.26}$$

Model selection is based on this approximation, where the best model is characterized by the greatest log-evidence (i.e., the smallest free-energy) [48]. Accordingly, the model inversion (i.e., estimation) is the process of finding the parameters that offer the best trade-off between accuracy (the fit of the predicted time-series to the data) and the complexity of the model (how far the parameters had to move from their prior values to explain the data). The model evidence quantifies this trade-off between accuracy and complexity.

2.4.1. Bayes factors

In the next stage, hypotheses are tested by comparing the evidence for different models, either at the single-subject or the group level. To compare two models m_i and m_j, we can compare their log evidences [49] by defining the Bayes factor (BF):

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)}. \tag{2.27}$$

When $BF_{ij} > 1$, the data favour model i over model j, and when $BF_{ij} < 1$, the data favour model j. Table 2.1 shows the interpretation of BF, which is an established convention to prefer one model over another if the BF is > 3 and the posterior probability of model i, $p(m_i|y)$, is larger than 75% (positive evidence) [50, 51].

Table 2.1.: Interpretation of Bayes factors.

BF_{ij}	$p(m_i\|y)$	Evidence in favor of model i
1 to 3	50-75%	weak
3 to 20	75-95%	positive
20 to 150	95-99%	strong
≥ 150	$\geq 99\%$	Very strong

In practice, we want to study multiple individuals and make an inference at the group level. In this regard, a group Bayes factor (GBF) can be computed by multiplying the individual Bayes factors for $1...K$ subjects:

$$GBF_{ij} = \prod_k BF_{ij}^{(k)}.$$

This form of Bayesian model comparison (BMC) is called fixed effects (FFX) BMC at the group level as it assumes that all the subjects' data are generated by the same model [17, 52]. However, this method has some problems. It is blind with regard to group heterogeneity and sensitive to strong outliers [17].

2.4.2. Random effects BMS for heterogeneous groups

Here it is assumed that the model is a random variable, and each subject's data is generated with different models. An approach to tackle this is to build a hierarchical generative model of the log-evidences. Given the data across subjects, one can estimate the distribution of model probabilities in the population in several ways. One can ask how likely it is to draw a subject randomly from the population, and his/her data are generated by one specific model. In this way, each subject can have their best model (in contrast to FFX, which assumes that the same model generates all the subjects' data).

Here again, the VB approach is used to estimate the log-evidence for each model and subject (for more detail, see [17]), and report the results of random effects (RFX) BMS in different ways which all of them give the same ranking:

1. **Dirichlet parameter** estimates α, which shows the occurrences of models in the population. It can be thought of as the effective number of subjects, in which model k generated the observed data and describes the probabilities for all models considered.

2. **Expected posterior probability** of obtaining the k-th model for any randomly selected subject:

$$\langle r_k \rangle_q = \alpha_k / (\alpha_1 + ... + \alpha_K).$$

3. **Exceedance probability** that a particular model k is more likely than any other model (of the K models tested), given the group data:

$$\exists k \in \{1...K\}, \forall j \in \{1...K | j \neq k\} : \varphi_k = p(r_k > r_j | y; \alpha). \tag{2.28}$$

4. **Protected exceedance probability** that each model is the most likely model across all subjects taking into account the null possibility that differences in model evidence are due to chance. I will present this in more detail after introducing the Bayesian Model Averaging.

The exceedance probability φ_k, conditional expectations of model probabilities $\langle r_k \rangle$, and Protected exceedance probability sum up to one over all models tested.

As the number of ROIs increases, the number of the possible models, i.e., the model space, rises sharply, resulting in a high risk of overfitting. Definition of a hypothesis and restricting the model space to a certain number of models (choosing priors) is a solution to the overfitting. Furthermore, family-level BMS and Bayesian model averaging are some other solutions to this problem. In the following, I explain these two approaches.

2.4.3. Comparing model families or family-level inference

If we have to cover a large model space, we can effectively reduce it with family-level comparison [24]. In this way, we consider the models which share a common property (e.g., presence or absence of a connection) in a family and compare between the families of models.

Partitioning model space into K subsets or families: $M = \{f_1, ..., f_K\}$ and then pooling information over all models in these subsets allows one to compute the probability of a model family, given the data, and it corresponds to summing posterior model probabilities within each of the families. This can effectively remove the uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partitions into families). Furthermore, to avoid bias by differently sized families, the priors of a family K is defined according to how many models N_k it contains

$$p(m_k) = \frac{1}{N_k K}.$$

By this means, the inference is not dominated by large families, and the number of models within a family determines the tightness of prior.

The above equation is for FFX BMS. In the same way for RFX BMS, fair priors over families are achieved by setting the

$$\alpha_{prior}(m) = \frac{1}{N_k}.$$

2.4.4. Bayesian model averaging

Bayesian model averaging abandons inference in models and provides inferences about parameter space by creating average posterior and integrating out the models [24]. It also abandons the dependence of parameter inference on a single model and considers the model uncertainty. It is defined as the probability of the parameters given the data equals the sum of conditional posterior on a particular model multiplied by the posterior probability of that model:

$$p(\theta|y) = \sum_m p(\theta|y, m)p(m|y). \tag{2.29}$$

This equation is the BMA for the single subject and for a group of subjects (N is the number of subjects) represented as below:

$$p(\theta_n|y_{1\ldots N}) = \sum_m p(\theta_n|y_n, m)p(m|y_{1\ldots N}).$$

Notice that $p(m|y_{1\ldots N})$ can be obtained by either FFX (all subjects use the same model) or RFX (each subject uses their own model) BMS.

BMA uses the entire model space considered (or an optimal family of models) and averages parameter estimates, weighted by posterior model probabilities. In this way, the models with a higher likelihood contribute more to the final average probability. Furthermore, BMA represents a beneficial alternative when none of the considered models (or model subspaces) outperform all others and when the optimal model differs within the comparing families.

2.4.5. Protected exceedance probability

Exceedance probabilities (EPs) express our confidence that the posterior probabilities of models (r_k) are different under the hypothesis H_1: $r_k \neq 1/K$ (K, number of models). Hence it does not account for the possibility of 'null hypothesis' H_0: $r_k = 1/K$, in which the posterior probabilities of models are all equal to each other [18]. Under the null, any observed differences in the frequencies are due to chance. In this regards, the Bayesian omnibus risk (BOR) is defined to exclude chance as a likely explanation for an observed difference in model frequencies.

Bayesian omnibus risk of wrongly accepting H_1 over H_0, given the observed model m, is defined as below:

$$P_0 = \frac{1}{1 + \frac{p(m|H_1)}{p(m|H_0}}.$$

P_0 evaluates the probability that the observed sample may have occurred by chance. Protected EP uses BOR to compute a BMA of the exceedance probability and accounts for H_0. It is a BMA over H_0 and H_1. Using Eqs. 2.28 and 2.29, we have:

$$\begin{aligned}\tilde{\varphi}_k &= p(r_k \geq r_{k'\neq k}|y) \\ &= p(r_k \geq r_{k'\neq k}|y, H_1)p(H_1|y) + p(r_k \geq r_{k'\neq k}|y, H_0)p(H_0|y) \\ &= \varphi_k(1 - P_0) + \frac{1}{K}P_0,\end{aligned} \tag{2.30}$$

where φ_k is the exceedance probability. Protected EPs also sum to one and, in short, quantify the probability that any model is more frequent than the others, above and beyond chance.

3. Dynamic causal modeling for fMRI with Wilson-Cowan-based neuronal equations

Dynamic causal modeling (DCM)[1] is an analysis technique that has been successfully used to infer about directed connectivity between brain regions based on imaging data such as functional magnetic resonance imaging (fMRI). Most variants of DCM for fMRI rely on a simple bilinear differential equation for neural activation, making it difficult to interpret the results in terms of local neural dynamics. In this work, we introduce a modification to DCM for fMRI by replacing the bilinear equation with a nonlinear Wilson-Cowan-based equation and use Bayesian model comparison (BMC) to show that this modification improves the model evidences. Improved model evidence of the nonlinear model is shown for our empirical data (imitation of facial expressions) and validated by synthetic data as well as an empirical test dataset (attention to visual motion) used in previous foundational papers. For our empirical data, we conduct the analysis for a group of 42 healthy participants who performed an imitation task, activating regions putatively containing the human mirror neuron system (MNS). In this regard, we build 540 models as one family for comparing the standard bilinear with the modified Wilson-Cowan models on the family-level. Using this modification, we can interpret the sigmoid transfer function as an averaged f-I curve of many neurons in a single region with a sigmoidal format. In this way, we can make a direct inference from the macroscopic model to detailed microscopic models. The new DCM variant shows superior model evidence on all tested data sets.

3.1. Introduction

Since its invention, fMRI has been developed into a powerful and versatile measurement technique. Apart from localizing a wide range of brain functions, it can now also be used to make statistical inferences about the neural network underlying these functions. This kind of inference has been made possible by sophisticated analysis techniques such as DCM. DCM is a well-established method to investigate the causal structure (effective connectivity) of a system of brain regions. It uses a Bayesian framework to deduce hidden neuronal states from time series of observed data measured by fMRI or other neuroimaging tools such as electroencephalography (EEG) or magnetoencephalography (MEG). DCM provides posterior estimates of intrinsic synaptic coupling strengths among neuronal populations, the

---

[1] This chapter is published in Front. Neurosci. 14, 593867 (2020) [53].

inputs that modulate those couplings, and extrinsic inputs driving the neuronal states [16, 31].

The interpretability of DCM is limited by the expressiveness or complexity of the underlying neural model. This complexity is constrained by the nature of the data at hand. As we will see below, the best generalizing models have the most significant model evidence. Log model evidence is accuracy minus complexity. This means that there is an optimal model complexity for any given kind of data. In what follows, we ask whether typical fMRI data could support more expressive or complex models that incorporate sigmoid activation functions, which are characteristic of neuronal dynamics. The most current versions of DCM for fMRI rely on a relatively simple, completely linear model of neuronal activity, which is justified as a Taylor expansion of more complex dynamics near a fixed point [16, 45, 54–56]. This is mainly due to the low temporal resolution of the fMRI data, making it necessary to estimate parameters from a very limited number of data points and restricts the number of parameters that can reasonably be inferred [57]. DCM has also been used in EEG and MEG with considerably more complex neuronal state equations than in standard bilinear DCM for fMRI [31–33], as the finer time resolution allows to constrain a wider range of neuronal processes at different time scales. Very recently, more complex models have also been applied to fMRI data, including simulated superficial and deep pyramidal cells, spiny-stellate excitatory and inhibitory interneurons, all contributing to the ongoing dynamics [58–60] (for a more detailed comparison of the existing DCM variants, see section 'Comparison to other DCM extensions' in the Discussion).

While the increased complexity of such extended DCMs opens the possibility to make more detailed inference about the networks underlying brain functions, it also makes those models harder to fit the data, as increasing the number of fitted parameters increases both computational cost and the risk of obtaining suboptimal fits. Furthermore, it has been shown in other contexts that complex models with a large number of parameters can be seriously underconstrained, i.e., several qualitatively different sets of parameters fit the data equally well, making it hard to interpret the results [61]. Thus, fitting of complex models to data with limited resolution can result in solutions that produce good fits, but unphysiological parameter regimes. Even worse, fitting may result in physiologically plausible solutions, which point towards neural mechanisms that are nevertheless entirely different from those being used by the brain.

We propose a solution to the dilemma between detailed inference and underconstrained modeling using a DCM, which is relatively simple but involves a more realistic, nonlinear neuron model. More precisely, Wilson-Cowan-type equations [20], which describe the evolution of excitatory and inhibitory activity in a population of neurons, are implemented instead of standard bilinear equations for both single and two-state DCM. In this way, the parameters obtained by DCM can be directly interpreted physiologically (see section 3.2). In the future, these DCM results can be used to constrain a spiking network model [19] to derive predictions about physiological details that cannot be obtained from non-invasive recordings.

We test the new nonlinear modification of the DCM framework based on the Wilson-Cowan model (W-C DCM) on three different data sets and show its superiority in model evidences compared with the standard bilinear model. First, we use an established data set that has been widely used as a test case for DCM [22]. This section compares W-C DCM with the bilinear DCM for the two best models achieved from previous studies [51, 62]. Second, we investigate the dynamics of the human MNS using our own novel data set. Here, W-C DCM is shown to unravel connections which are overlooked by bilinear DCM. Finally, we generate synthetic data with different signal-to-noise ratios (SNRs) based on the novel data to investigate how W-C DCM performs when the ground truth is known. We show that W-C DCM provides explanations with more significant evidence compared to bilinear DCM for low SNRs, which are typical for fMRI data.

3.2. Materials & Methods

3.2.1. Wilson-Cowan equations for DCM

In this section, we briefly review single-state DCM for fMRI data [16] as well as the Wilson-Cowan model [20, 21] before introducing the modifications of the neuronal state equation.

DCM describes a system characterized by m inputs and l outputs with one output in each brain region. The experimental manipulations are modelled as changes in the inputs. In each of these regions, the output is measured, which corresponds to the observed BOLD signal. Normally these time series are considered as average or typical values of given brain regions. Each region is described by five state variables, four of which correspond to the hemodynamic model, i.e., the dilatation of the blood vessels, the normalized blood flow, the normalized venous volume, and the deoxyhemoglobin content of the blood [27, 34]. These variables are independent of the state of other brain regions. The fifth state variable is the neuronal or synaptic activity in each brain region, modulated by the neuronal states in other regions.

The effective connectivity of the regions is described at the neuronal level. This neuronal activity is modeled by a multivariate differential equation that has a bilinear form in the original format [16] to describe the dynamics:

$$\dot{z}_t = (A + \sum_{j=1}^{m} u_j B^j) z_t + Cu \tag{3.1}$$

$\dot{z}_t$ denotes the time derivative of neuronal activity (z) and u_j is the j-th of m inputs at time t. Matrix A, also called the connectivity matrix, describes the interconnections between the brain regions and the influence that a neural system exerts on another. The matrices B^j describe the change in connectivity through the j-th modulatory input u_j. Finally, the matrix C embodies the direct influences of the external inputs u on the neuronal activity. This equation can be achieved from a Taylor expansion of any nonlinear function, $F(z, u, \theta)$, around the system's resting state ($z = 0, u = 0$). Such a nonlinear function can be thought to describe both the synaptic transmission between regions and the neural computations within each region. Thus, when estimating the connectivity matrices,

$\theta = \{A, B^j, C\}$, the estimated numbers also reflect neural computations and synaptic transmission together. Thus, one goal of the presented framework based on the W-C model is to disentangle local computation (using a nonlinear transmission function S) and transmission between regions (using the linear connectivity matrices $\theta = \{A, B^j, C\}$, which thus are to be interpreted differently compared to the bilinear model).

The Wilson-Cowan model describes the evolution of firing rates of a large population of densely coupled neurons. Assuming both excitatory (E) and inhibitory neurons (I) in this population to be homogeneous, their firing rates $R_E(t)$ and $R_I(t)$ are governed by two differential equations:

$$\begin{aligned}
\tau_E \dot{R}_E(t) &= -R_E(t) + S_E(x_E) \\
\tau_I \dot{R}_I(t) &= -R_I(t) + S_I(x_I) \\[1em]
S_E(x_E) &= \frac{1}{1 + \exp(-\alpha_E * (x_E - \theta_E))} \\
S_I(x_I) &= \frac{1}{1 + \exp(-\alpha_I * (x_I - \theta_I))} \\[1em]
x_E &= w_{EE} * R_E(t) - w_{EI} * R_I(t) + I_E \\
x_I &= w_{IE} * R_E(t) - w_{II} * R_I(t) + I_I
\end{aligned} \tag{3.2}$$

τ_E and τ_I are the membrane time constants of the two subpopulations, and $S(x)$ denotes the sigmoidal nonlinearity as an activation or transfer function with slope α and threshold θ, which are also specific for E and I. w_{XY} is the synaptic weight of the connection from subpopulation X to Y and I_X represents the external input to each subpopulation, where X and Y can be E or I. The first differential equation describes an exponential relaxation of the firing rate $R_E(t)$ with time constant tau_E to its steady-state value $S_E(x_E)$, which is determined by a weighted sum x_E of both firing rates $R_E(t)$ and $R_I(t)$ as well as the external input I_E, filtered by the sigmoid nonlinearity S_E. The same is true for $R_I(t)$ with its respective variables and parameters. The weights in the sum can be directly interpreted as synaptic efficiencies between the subpopulations, while the sigmoid mimics the nonlinear input-output relations of the neurons in the subpopulation (Fig. 3.1). For large values α, this relation is very steep, so S is zero for inputs x below the threshold θ and one for input above. The relation becomes more gradual for lower slopes, but still saturates into zero and one for very low and very high inputs, respectively.

Wilson and Cowan used phase-plane analysis to show that the system described by equation 3.2 allows for a variety of dynamic phenomena that are relevant to the function of the brain [20], including multiple stable fixed points (a simple mechanism e.g., for working memory) and oscillations.

We propose to replace the standard bilinear equation (equation 3.1) with a Wilson-Cowan-type equation:

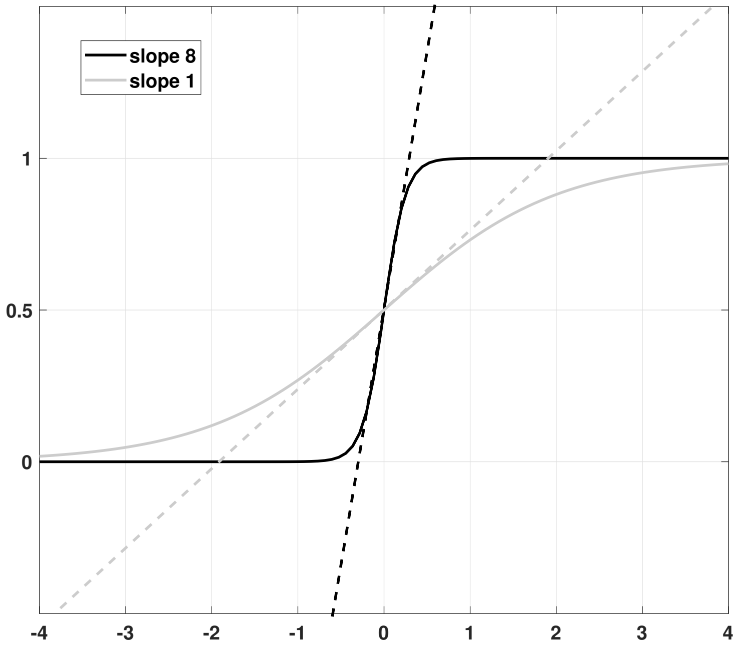


Figure 3.1.: A schematic illustration of two sigmoid functions with different slopes and the corresponding linear functions.

$$\begin{aligned} \dot{z}_t &= -z_t + S(x) \\ S(x) &= \frac{1}{1+\exp(-\alpha * x)} - \frac{1}{2} \\ x &= (A + \sum_{j=1}^{m} u_j B^j) z_t + Cu \end{aligned} \tag{3.3}$$

In this way, the different components of the model relate to the underlying biological elements of the brain: The matrices A, B, and C are the synaptic weights (w) in equation 3.2 (parameters merge excitatory and inhibitory synaptic weights), and the sigmoid nonlinearity directly relates to the f-I curve of single neurons [19]. This contrasts with the bilinear model, where the matrices intermingle synaptic weights and Taylor approximations of the nonlinearities. Note that this implementation of the sigmoid function allows for negative firing rates to ensure the neuronal system has a stable fixed point when all states are equal to zero and changes in state variable can be interpreted as deviations from the fixed point (cf. [31]). As a negative firing rate is not physiologically plausible, the model can be interpreted to capture activity relative to a baseline resting-state rather than total activity.

In this form, we replace separate excitatory and inhibitory variables with a single neuronal variable that can be positive or negative. However, it is possible to consider separate excitatory and inhibitory variables explicitly, using an extension of the two-state DCM.

3.2.2. Two-state DCM with Wilson-Cowan equations

We also incorporated the W-C model (equation 3.2) into two-state DCM [55] to compare its application between single- and two-state models and use it in the future to constrain local spiking neural networks in a more detailed way. In this version of DCM, each region consists of excitatory and inhibitory subpopulations, and in this way, it is biologically more plausible and less constrained than the single state model. This form of the DCM is more similar to the original W-C model (equation 3.2, [20, 21]). The standard bilinear equations of a two-state model are:

$$\begin{aligned} \dot{x}_E &= w_{EE}x_E - w_{SE}x_E - w_{IE}x_I + Cu \\ \dot{x}_I &= w_{EI}x_E - w_{SI}x_I \end{aligned} \tag{3.4}$$

Similar to the W-C model, the x_E and x_I summarize the dynamics of the excitatory and inhibitory neurons. w_{EE} is the extrinsic connection between excitatory neurons of different regions (The connections between the two regions are provided only by the excitatory neurons), and w_{IE} and w_{EI} are the within region (intrinsic) connections from excitatory (E) to inhibitory (I) populations and vice versa. Finally, w_{SI} and w_{SE} represent the intrinsic inhibitory self-connection on I, and excitatory self-connection on E, respectively. Due to the difference between intrinsic and extrinsic connections, x_E has two different meanings in equation 3.4: In the term including w_{EE}, x_E represents extrinsic input from different brain regions, while in all other terms, intrinsic input from the same region is meant. Furthermore, in this formulation, the between regions connections (w_{EE}) and intrinsic inhibitory to excitatory connections (w_{IE}) are split up into a direct and modulatory part, analogous to the A and B matrix components for single-state DCM.

For excitatory and inhibitory subpopulations, we modified these equations as below:

$$\begin{aligned} \dot{z}_E &= -z_E + S_E(x_E) \\ \dot{z}_I &= -z_I + S_I(x_I) \\[2ex] S_E(x_E) &= \frac{1}{1+\exp(-\alpha_E * x_E)} - \frac{1}{2} \\ S_I(x_I) &= \frac{1}{1+\exp(-\alpha_I * x_I)} - \frac{1}{2} \\[2ex] x_E &= w_{EE}z_E + w_{SE}z_E - w_{IE}z_I + Cu \\ x_I &= w_{EI}z_E - w_{SI}z_I \end{aligned} \tag{3.5}$$

Where α_E and α_I are the slope of sigmoid function in the excitatory and inhibitory subpopulations. In the original model in SPM [55], only w_{EE} and w_{IE} are estimated, but here we estimate all the parameters as well as the sigmoidal slopes for excitatory and inhibitory neurons.

3.2.3. Bayesian model selection

We use Bayesian model selection for comparing the W-C based equations with bilinear DCMs. Bayesian Model Selection is widely used for finding the model that fits the data best among several alternatives. Model evidence (the probability of obtaining observed data given the model) is widely used in this approach, using the free-energy criterion. This criterion is composed of two components: the accuracy term (log-likelihood of data), which computes the data fit, and the complexity term, which depends on the number of parameters and also the deviation of posterior densities from their prior. Two models m_1 and m_2 are compared using the Bayes Factor (BF_{12}), which is the ratio of model evidence of two models reported on a log scale. Its value equals the difference between the free energy of the models ($|F_1 - F_2|$). By convention, if the value of the $\log BF_{12}$ is about three or more, it indicates strong evidence in favor of model 1 over 2 [49, 50].

There are two different approaches at the group level for model inference: Fixed Effects (FFX) and Random Effects (RFX) analysis. In the FFX, Group Bayes Factors (GBF) [52] are widely used for model selection when a common model is being assumed for each subject, i.e., the most likely model structure is the same across subjects [17, 24]. This method is sensitive to outliers and blind concerning group heterogeneity. Hierarchical Random Effects analysis (RFX), on the other hand, models inference on the level of group analysis that allows each subject to have a different best model and computes the probability of all subjects' data given each model. In contrast to FFX, outliers have minimal effect on RFX results, which accounts for group heterogeneity. The results of RFX group analyses are reported in terms of expected, exceedance, and protected exceedance probabilities. The expected probability is the expected posterior probability of obtaining the n-th model for any randomly selected subject, and the exceedance probability is the probability that one model is more likely than any other model between all models tested. As the exceedance probability does not consider the null hypothesis that all model frequencies are due to chance, the protected exceedance probability is also utilized here [18], which considers this null hypothesis. Each of these measures can be used for finding the best model, and higher expected, exceedance, or protected exceedance probability independently means that a model is more probable. However, fixed effects BMS is also used in this study to show that the modified version of DCM has a better result in both Random and Fixed effects analysis. Furthermore, for the established data set [22], as it is only for one subject, FFX BMS usage for testing our modification is mandatory.

In this study, we performed the BMS on the family-level [24] and grouped all the possible models in one family to compare the Bilinear and W-C DCMs. The implementation was originally developed based on DCM10 (r6313) provided with SPM8 (Statistical Parametric Mapping 8, `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`), which was the most recent version when this work was begun. However, we repeated several analyses with DCM12 (r7487) in SPM12 and did not observe any qualitative differences. Furthermore, we computed the protected exceedance probabilities with the VBA toolbox as it is not implemented in the SPM software at the family level [18, 63].

Moreover, to characterize the two models at the microcircuitry level, we perform

Bayesian model averaging. This method averages each connectivity parameter over all models within the family or whole model space, weighted by each models' posterior probabilities. Thus, the most probable models will contribute the most to the model averaging [24].

In order to have a fair comparison between the Bilinear and W-C DCMs, we used the original format of shrinkage priors and the identical hyperpriors as the classical DCM [16, 55] for both single and two-state W-C DCMs. Furthermore, we also used the same inference Variational Bayes under the Laplace assumption (VBL) scheme as the original DCM [38].

3.2.4. Data sets

Established data set

We used well-studied data from an experiment on visual attention-modulated connectivity during visual motion processing (available from the SPM website `http://www.fil.ion.ucl.ac.uk/spm`, the full description of the experimental paradigm can be found in Büchel and Friston, 1997 [22]). In brief, the experimental variables were three exogenous inputs: A 'photic stimulation' variable indicated when dots were shown on a screen, a 'motion' variable indicated that the dots were moving, and the 'attention' variable indicated that the subject should attend to possible velocity changes. These are also the three input variables that we used in the DCM analyses shown in Fig. 3.3A. This data set for a single subject has been used several times to validate DCM for fMRI [51, 55, 56, 58, 64, 65].

Novel empirical data on imitation

Empirical data were acquired within the framework of a larger project on the human MNS. Participants underwent a simultaneous EEG-fMRI measurement. Here data of the fMRI measurement is presented. The reported analyses are conducted on a subset of 42 healthy participants out of the total final sample of 75 participants that were available by the time the analyses were conducted. The study was approved by the local ethics board at the Medical Faculty Mannheim, University of Heidelberg (2015-501N-MA), and participants signed written informed consent before participating in the study.

The imitation paradigm in Fig. 3.2 consists of 3 conditions (Observation, Imitation, Execution) and a motor control condition (Control). During observation, participants simply look at emotional faces, expressing anger or fear. During imitation, participants additionally imitate the facial expression displayed in the pictures. The words 'anger' or 'fear' are presented in the execution condition, and participants have to mimic the corresponding facial expression. The control condition requires participants to say out loud the German letters 'A' (pronounced similar to 'a' in 'car') or 'Ä' (pronounced similar to 'a' in 'anger'), which should resemble the facial expressions of fear and anger, respectively. The experimental trials are presented in blocks of 4 pictures. Experimental blocks are alternated with control blocks, consisting of 2 control stimuli. Stimuli within the blocks are presented in pseudo-randomized order and separated by a jittered inter-stimulus-interval of 1-3 seconds. Experimental stimuli are shown for 5 seconds, and the control stimuli for 3 seconds. Before each block, an instruction cue is presented for 2 seconds, which is preceded by a jittered inter-block-interval of 4-6 seconds. In total, each experimental block is presented 5 times, the control block 15 times, resulting in a total of 20 trials in each experimental condition, and 30 in the control condition.

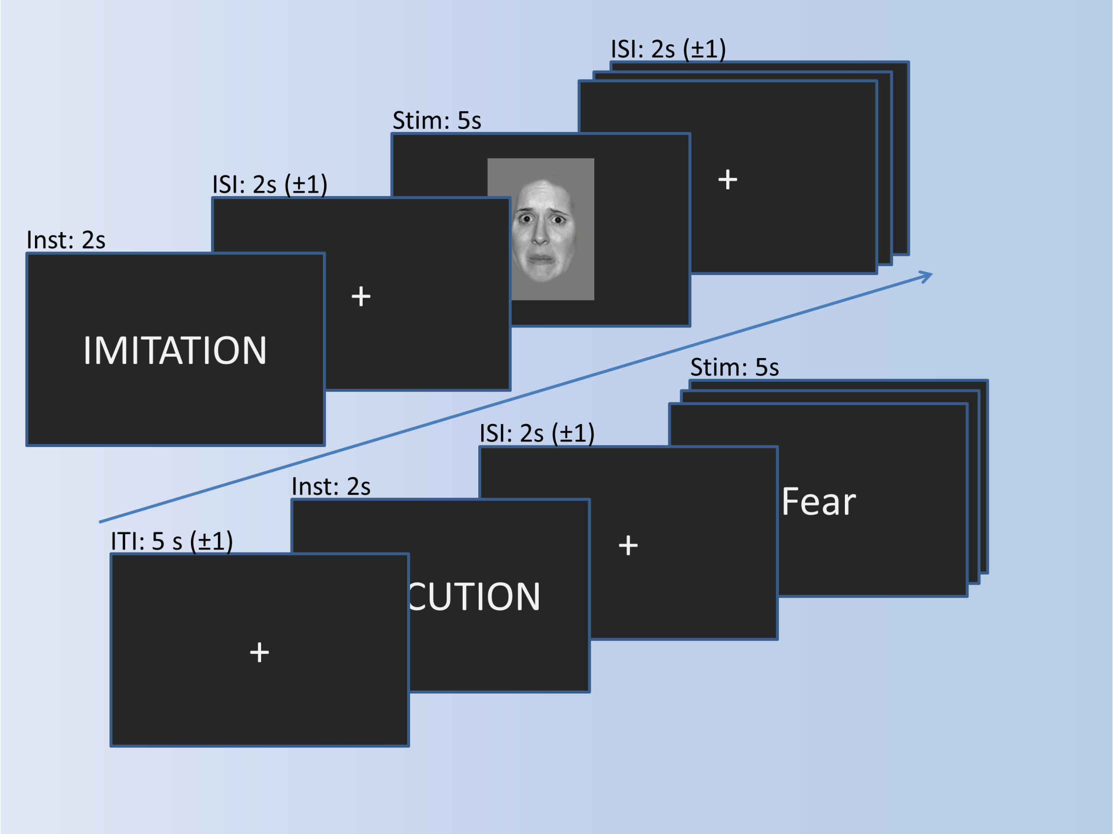


Figure 3.2.: Imitation paradigm with timing of trials. A trial for imitation and a trial for execution is shown exemplarily.

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. Echo-planar imaging ran with 32 descending $3 \times 3 \times 3$ mm slices with 1mm gap, TR = 2000 ms, TE = 30 ms; flip angle = 80°, field of view = 192 mm; matrix = 64×64. Prior to the experiments, an anatomical sequence was recorded (TR = 1570 ms, TE = 2.75 ms; flip angle = 15°, field of view = 256 mm; matrix = 256×256; voxel size $1 \times 1 \times 1$ mm).

Preprocessing consisted of slice time correction, realignment to the mean image, normalization and resampling with $3 \times 3 \times 3$ mm 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 (Imitation, Observation, Execution and Control) and the 6 movement parameters from the realignment procedure as covariates. First eigenvariates of the time series of imitation > control, were extracted with $p < 0.5$ without a cluster size threshold while adjusting for the activation during imitation from the regions of interest (ROI's). The ROI's were the main regions associated with the human MNS: BA 44, IPL and STS. 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 STS mask was based on activation in a study on social cognition and has been used as the region of interest in earlier publications [66, 67].

Synthetic data

To validate the results from the empirical dataset (comprising 42 participants), we generated a synthetic dataset for which the network architecture and parameter values were known. We generated the synthetic fMRI data using the standard bilinear equation (equation 3.1) and the usual hemodynamic equations [34]. Here we use a typical connectivity model from a three-area model (Fig. 3.8B). Its network structure consists of one driving input into the first region and feedforward connections from the first region to the second and third regions, as well as a forward connection from the third region to the second region. There is also a contextual input on the forward connections from region 1 to region 3 and from region 3 to region 2. The generating parameters were also sampled from the estimated posterior values (the mean of expected values) of the previous section's empirical data to ensure the synthetic data is realistic (reported in Fig. 3.8B). We simulated the BOLD signal from this model and then generated the synthetic data by adding ten realizations of normally distributed random noise for each SNR. In this way, we have simulated ten artificial time series for each region with different signal-to-noise ratios. Then we did the parameter estimation for both Bilinear and W-C DCMs for these synthetic data with different SNRs. In this way, we could test the robustness of the analysis for varying levels of noise. We compare the two estimates using the percentage of the observed time series variance explained by the time series predicted by DCM [68]. SNR values range between zero and 0.5, and the repetition time (TR) equals two seconds.

Usually, one uses synthetic data to establish the face validity of DCM in terms of Bayesian model comparison. In other words, one would generate data under a variety of models and then assess the evidence for the different datasets under the models used to generate the data. This creates a confusion matrix of model evidences that can be used to establish that the model generating data was recovered via Bayesian model selection. In one sense, we have already established face validity at the level of model comparison (see above).

The use of synthetic data in this section differs slightly and speaks to the robustness of model inversion instead of validity. In some circumstances - due to the nonlinearity of DCM's - there may be a failure of convergence to the global minimum of free energy. In other words, the scheme gets trapped in local minima; usually, that random fluctuations can explain all the data. This means that the predicted data responses 'flat-line'. Therefore, we assessed the ability of the W-C DCM to elude local minima by showing that inversion under different levels of noise reduces the instances of 'flat-lining' - as scored with the accuracy or variance explained.

3.3. Results

In the following sections, we will test whether the W-C based equations can improve the predictions of fMRI data using both empirical and synthetic data.

3.3.1. Validation of established data set

In a first step, we investigate the validity of the W-C DCM using a well-studied data set on visual attention [22]. As shown in Fig. 3.3A, activity is modeled in three regions, V1, V5, and superior parietal cortex (SPC), with sensory input to V1 and motion and attention as modulatory inputs on connections [22]. Previous DCM researches have established a connection scheme between these regions [51, 55], so we can use this data set as a test case of our extended method.

We used the two models that had the most substantial evidence according to previous research using the Bayesian model comparison (Fig. 3.3A) [51, 62]. The results of this comparison are represented in Fig. 3.3B/C in terms of the relative log-evidence and posterior probability for both single- and two-state DCM. As can be seen, model 2, in which the attention input modulates the forward connection from V1 to V5, has more robust evidence in bilinear and nonlinear DCM, consistent with previous findings [51, 55]. Furthermore, there is much stronger evidence (posterior probability) for both models in favor of Wilson-Cowan-based DCM. Thus, the modified DCM framework provides a better explanation for the data while preserving the original distinction between the two connection schemes. Please note that the dataset contains only one subject, so we performed FFX BMS, as RFX analysis can only be performed in a group analysis.

3.3.2. Validation of novel empirical data

Next, we apply W-C DCM to a novel data set using an imitation task to probe the human MNS, including the three regions BA44, IPL, and STS (see methods for details). For this task, it is known that the visual input goes to the STS region [23, 69], and we use this hypothesis to build the model space, including 540 different models to test all the possible combinations of the forward and backward connections with their modulatory elements. We constructed the model space accordingly: From STS, the input would propagate to the IPL and BA44. The effective connectivity between the two regions can be both feed-forward or reciprocal. So in our case, we have three nodes, and these nodes can maximally have six connections in case of mutual connectivity. Furthermore, we have considered all possible modulatory inputs on the connections. In this way, 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 total, for a network of three nodes and one experimental condition, one can build 5832 (all possible models) models [70] (to get this number, we used the second equation in the section 'Combinatorial Explosion' in the paper, n=3, m=1). In this way, the experimental condition (imitation in our case) can integrate into each of 3 regions (one region, two regions,

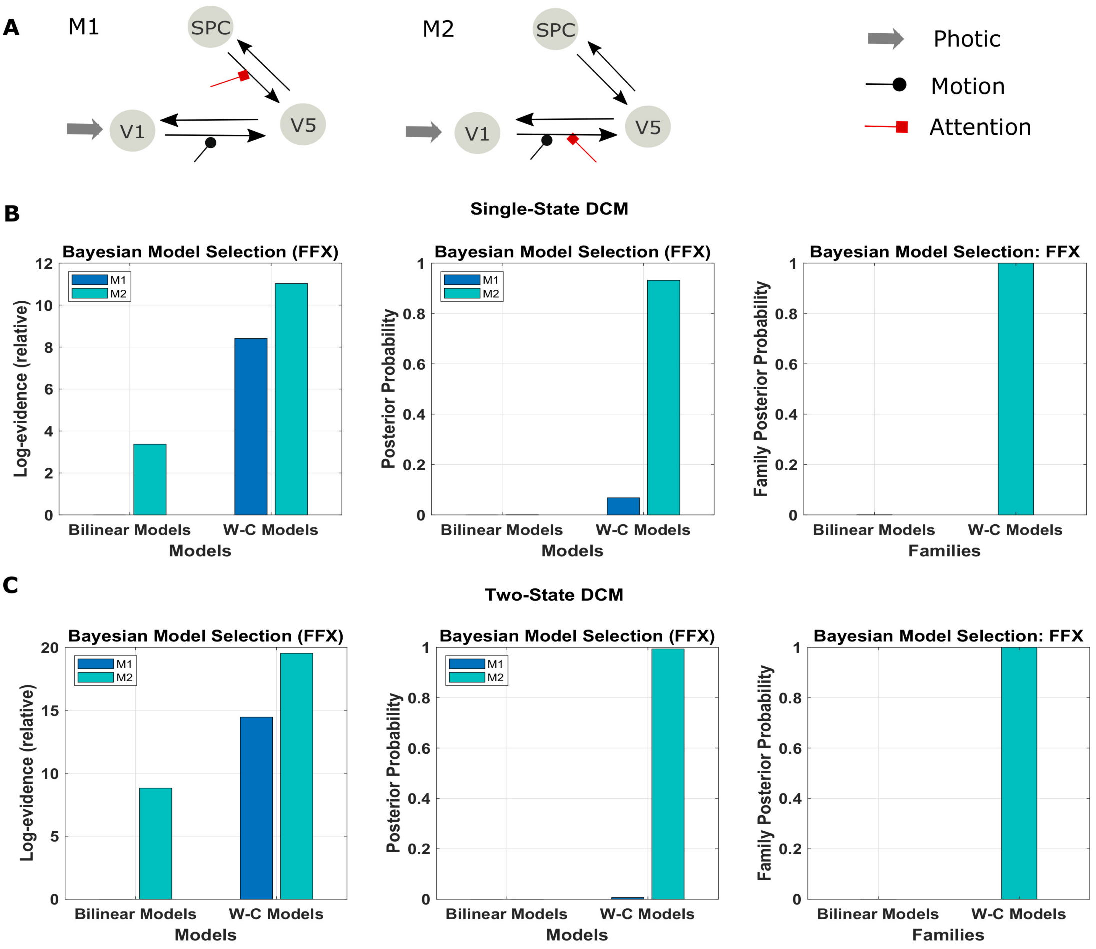


Figure 3.3.: Results of the FFX Bayesian model comparison of two models of the forward and backward attention modulation (M1 and M2) for the bilinear models, and the W-C models. (A) Illustration of the two best models (see the text). Comparison of the two models in row A for (B) single-state and (C) two-state DCM with W-C and bilinear neuronal equations (left and middle panel) and comparison of the two equation types with the two models combined in one family (right panel). The results show strong evidence (both single- and two-state) for the W-C models in all cases (family posterior probability one for W-C models and zero for standard bilinear).

or all three regions simultaneously; 3!=6 different variants) and modulate the connections. However, with our hypothesis, we restricted the external input only into the STS region and could build 540 models.

We have tested all the 540 models in the family-level to compare the modified DCM to the standard bilinear DCM. As shown in Fig. 3.4 and Table 3.1, the result of the Bayesian model selection for both fixed effect and random effects shows that the modified version of DCM has a probability of one and zero for the Bilinear models. For RFX BMS, the results are presented with expected, exceedance, and protected exceedance probability in table 3.1. We illustrate the exceedance and protected exceedance probability in Fig. 3.4, together with the posterior probability for FFX BMX. As can be seen, W-C models have a strong probability of one in all cases.

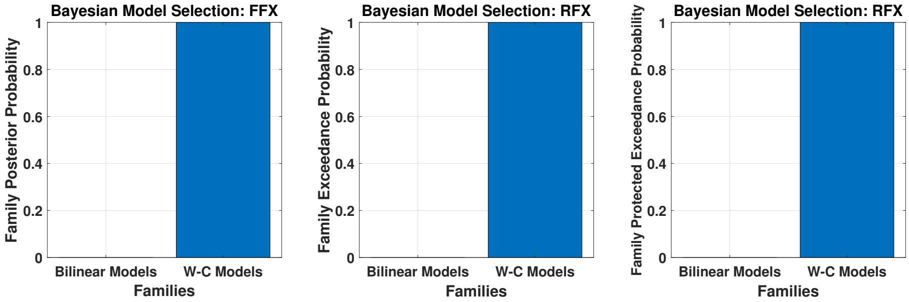


Figure 3.4.: Results of Bayesian model comparisons for all possible models in the family-level with standard bilinear equations and W-C equations. This comparison is done with both FFX and RFX BMS.

Table 3.1.: RFX BMS results of the single-state and two-state models for both bilinear and W-C models and also the comparison between the single-state and the two-state W-C model (Fig. 3.5B).

RFX BMS	**Single-State Model**		**Two-State Model**		**Wilson-Cowan Model**	
	Bilinear	**W-C**	**Bilinear**	**W-C**	**Single-State**	**Two-State**
Expected Probability	0.06	0.94	0.15	0.85	0.57	0.43
Exceedance Probability	0	1	0	1	0.80	0.20
Protected Exceedance Probability	0	1	0	1	0.82	0.18

To assess the flexibility of the modification introduced above, we also applied it to two-state DCM. Figure 3.5 (Table 3.1) shows the Bayesian model comparison for two-state DCM with bilinear and W-C equations. In Fig. 3.5A, by using random effects BMS, we have compared all the models on the family-level as before for the two-state DCM. As can be seen, it gives a probability of one for the modified version. In Fig. 3.5B, we also compare the modified single- and two-state DCM with each other. In contrast to the original paper [55], the single state has a higher probability with our data and thus provides a better fit to data than the two-state model. This is an important result because it shows that the best model is not necessarily the most complex model. In other words, the simpler one-state model had more evidence than the more complex two-state model (that can fit the data more accurately). We will return to this issue in the discussion.

An additional benefit of our modification is that the W-C DCM produces meaningful results for participants with weak activation (e.g. showing activation for the more lenient threshold p=0.5, but no activation for the stricter, standard threshold p=0.05). We observed convergence to a local minimum at low activation (i.e., flat-lining) in 18 subjects out of 42 with the standard DCM, while the W-C DCM produced non-flat time series for almost all of these subjects (17 subjects out of 18). Figure 3.6 shows the differences between the outputs of DCM analysis for the two kinds of models of a typical single subject in addition to the connectivity model used to plot these graphs. Next to each connection, one can find the estimated parameters (the mean of expected values) from the Bilinear and W-C

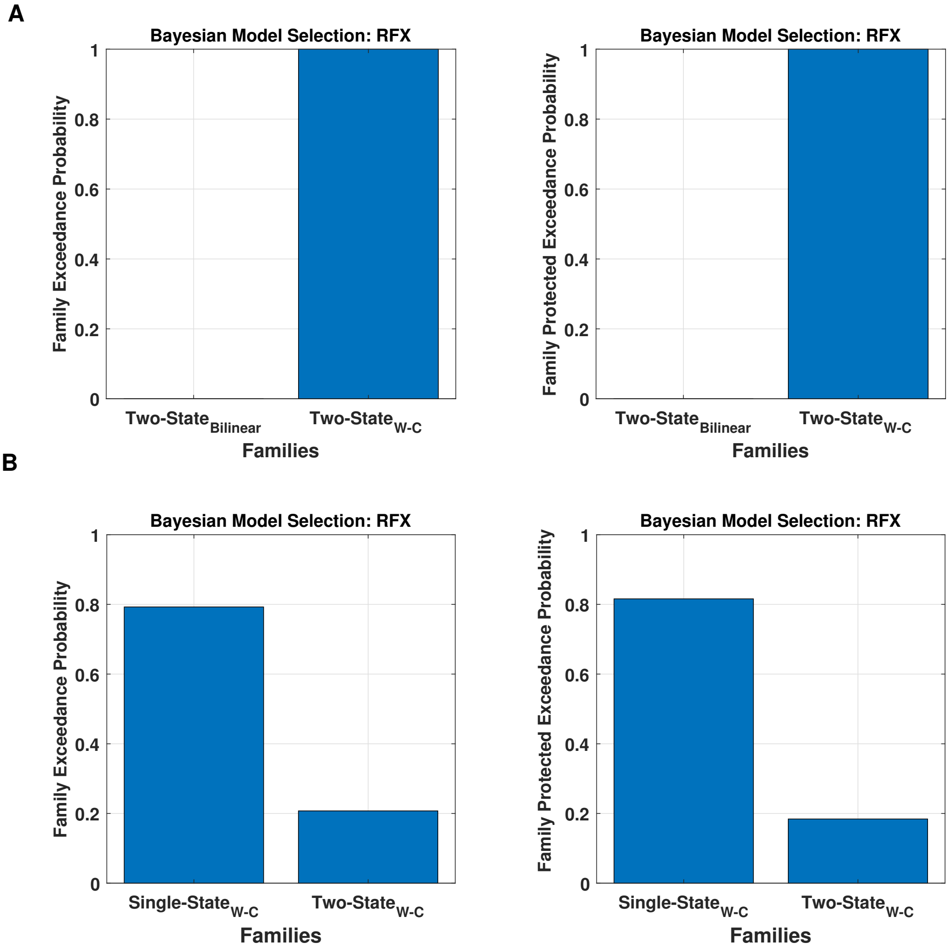


Figure 3.5.: Bayesian model comparison (RFX) for the two-state model at the family-level. (A) comparison between two-state bilinear models and two-state W-C models, (B) comparison between single and two-state W-C models.

DCM. As can be seen, the W-C model's estimated connections are much stronger than the standard bilinear model.

Finally, we investigate how the bilinear and W-C models inform us about the microcircuitry of the MNS. For this, we perform a BMA analysis for both Bilinear and W-C single-state models over all model space, 540 models. Inspection of Fig. 3.7 and Tables 4.2 and 4.3 show the parameter estimations for the two models are very different from each other. In Fig. 3.7, only the parameters with a posterior probability of greater than 0.95 ($P > 0.95$) are shown. For both models, the self-inhibitory connections and the forward connections from STS to BA44 and IPL are in common; however, the W-C model connections appear stronger. Regarding the model differences, for the W-C model, there are reciprocal connections between BA44 and IPL with modulatory inputs on all significant intrinsic connections except for IPL→ BA44 (however, its probability is very close to 0.95). In addition, the bilinear model predicts the BA44→ IPL as weak inhibitory connection and the W-C model as an excitatory and robust connection.

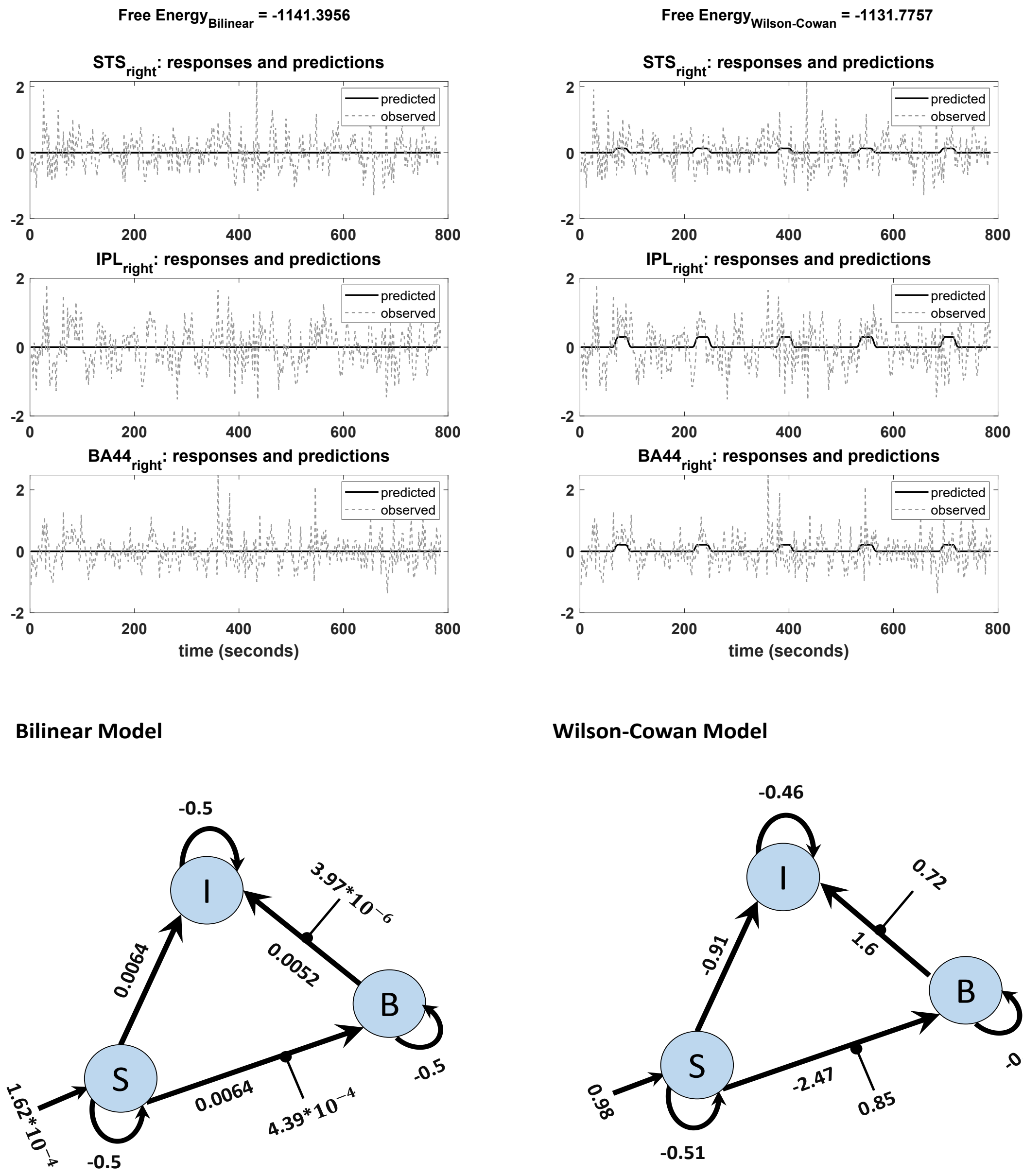


Figure 3.6.: Observed and predicted time-series in DCM analysis of bilinear (left) and W-C (right) models for one subject. The left graph shows a flat time series, while in the right graph, predicted activity reacts to the inputs. The network structures with estimated parameters (the mean of expected values) of each model are illustrated below of each graph.

3.3.3. Validation of synthetic data

Here, we are interested in investigating the robustness of the modified single-state DCM with a synthetic data set for which the properties are known. We generated the synthetic fMRI data using the standard bilinear equation for a three-area model achieved by analyzing real data and adding random noise with different signal-to-noise ratios (see 3.2). In this way, we could check how well the modified DCM

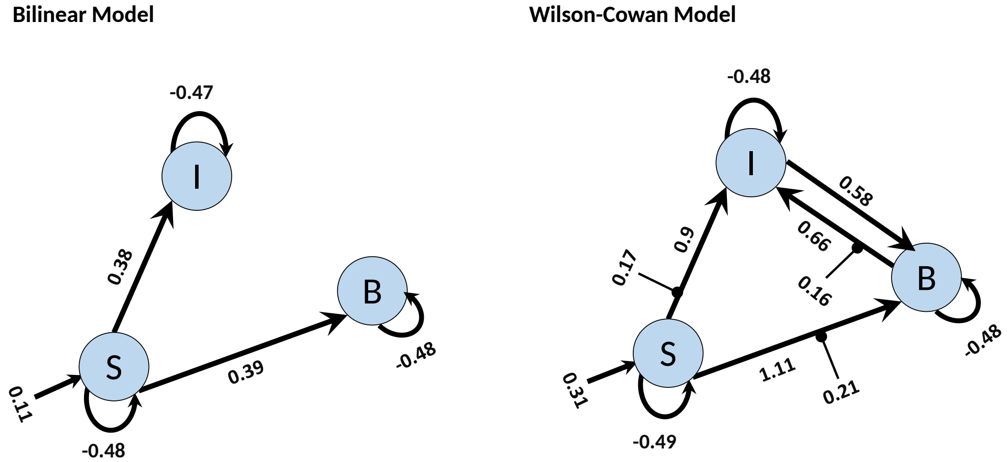


Figure 3.7.: BMA results for Bilinear and W-C models. Here we illustrate only the parameters which are significantly different from zero. The values next to each connection are the expected value (mean) of each parameter and the values for external inputs (Matrix C). All parameters are in Hz.

Table 3.2.: BMA results. The expected values (mean) for endogenous connectivity (Matrix A) (in Hz). In parentheses, the posterior probability is shown for each parameter to be different from zero.

Bilinear From To	**STS**	**IPL**	**BA44**
STS	-0.4772 (1.00)	-0.0671 (0.67)	-0.1750 (0.88)
IPL	0.3841 (0.99)	-0.4731 (1.00)	-0.0113 (0.53)
BA44	0.3953 (0.99)	0.0452 (0.62)	-0.4776 (1.00)
Wilson-Cowan			
STS	-0.4919 (1.00)	-0.1748 (0.82)	-0.1431 (0.77)
IPL	0.9006 (1.00)	-0.4795 (1.00)	0.6591 (0.99)
BA44	1.1071 (1.00)	0.5788 (0.99)	-0.4798 (1.00)

could identify the underlying 'ground truth' based on synthetic data with different noise levels. In this regard, we compared the W-C model with the bilinear model by checking the percentage of variance explained by the model for different SNRs. As can be seen in Fig. 3.8A, even for very small values of SNR (zooming portion), the W-C model fits the data better than the bilinear model in a significant way ($P - value = 0.013$). With increasing the SNR values, W-C models still fit the data better (but not significantly, $P - value = 0.42$) until explained variance values merge again for larger SNR. This shows how the W-C model enables an inversion scheme to escape the local minima in a low to a higher SNR range. In Fig. 3.8A, we also add error bars to show that the explained variance is actually

Table 3.3.: BMA results. The expected values (mean) for modulatory connectivity (Matrix B). In parentheses, the posterior probability is shown for each parameter to be different from zero.

Bilinear From To	**STS**	**IPL**	**BA44**
STS	-	-0.0005 (0.50)	-0.0144 (0.59)
IPL	0.0853 (0.85)	-	0.0386 (0.70)
BA44	0.0989 (0.87)	0.0433 (0.73)	-
Wilson-Cowan			
STS	-	0.0094 (0.54)	0.0095 (0.54)
IPL	0.1711 (0.96)	-	0.1609 (0.96)
BA44	0.2059 (0.97)	0.1515 (0.947)	-

(significantly) different for intermediate SNRs. We roughly estimate the SNR for the novel empirical data used in the previous section to be 0.02 [71], being at the lower end of the spectrum.

As any sensitive measure can be prone to produce false positive results, we assess the probabilities for each connection in the reconstructed model and compare the results with the 'ground truth' network structure that was used to generate the synthetic data (Fig. 3.8B). In particular, we vary the threshold p the probability for a connection needs to exceed to predict that particular connection to exist. If such a predicted connection does not exist in the generating model, it is counted as a false positive. Conversely, any missing connection in the reconstructed model that is present in the generating model is considered a false negative. Plotting the percentage of true positives (1 minus false negatives) against the percentage of false positives yields a receiver operator characteristic (ROC) curve, and the area under this curve (AUC) is a measure of the diagnostic ability of the model independent of the choice of the threshold. Figure 3.8C shows the AUC of the ROC curves for different signal-to-noise ratios. It is apparent that the W-C model is more efficient in correctly detecting connections between the areas than the bilinear model for small SNR values (below 0.3), while the detection performance of the two models converges for larger SNR values. In particular, false positive rates are comparable for both models at all SNR values, while false negative rates are lower for the W-C model at low SNR values.

3.4. Discussion

In this paper, we present a new DCM variant for fMRI on the level of neuronal states in which the equations have a sigmoidal form for the latent variables. Using our measurements as well as synthetic and established real data, we show by Bayesian model comparison that the modified model explains better with the observed data than the standard bilinear equation and allows us to detect smaller effects. Furthermore, our results support current theories on information flow in the MNS.

In particular, Bayesian model selection showed that the W-C DCMs with the

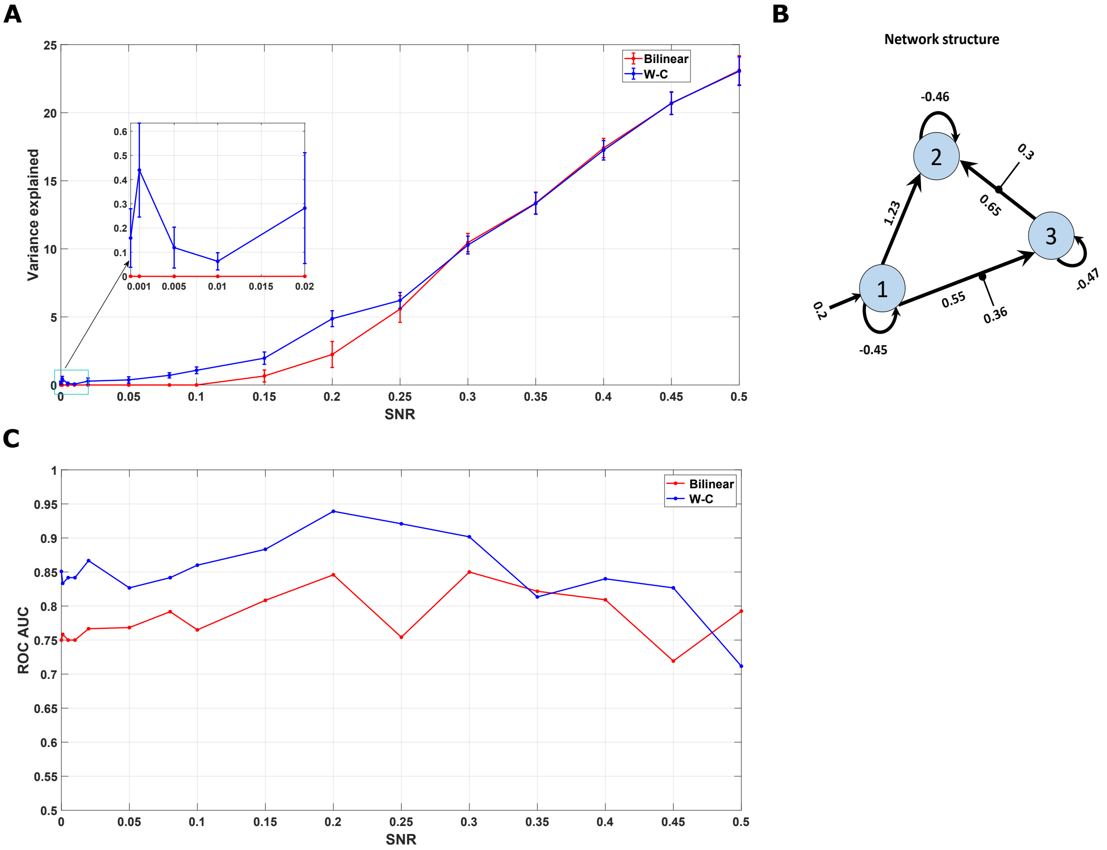


Figure 3.8.: Synthetic data result. (A) The mean value of the variance explained by the model with different Signal to Noise Ratios for both W-C and bilinear models (ten realizations of noise for each SNR). Error bars are standard errors. (B) The underlying model's network structure, which is used to generate synthetic data with the estimated parameters (the mean of expected values) from the novel empirical data. (C) Area under the curve (AUC) of the receiver operator characteristic (ROC) curves of detecting the existence of a connection between two areas as a function of the SNR value.

sigmoidal form have more significant expected and exceedance probabilities than the standard bilinear equation models for single state DCM. Moreover, at the microcircuitry level, it informs us more about the network connections responsible for imitation in the MNS. Thus, we have shown that DCM with a slightly more complex neuronal model outperforms the simple linear model, and additionally allows us to make predictions about local neuronal circuits, namely on the slope of the input-output relation. Furthermore, 18 participants out of all 42 subjects gave a flatline in the predicted time series with the bilinear models. With the W-C models, 17 of these participants could be rescued for analyses. Our significance threshold ($p < 0.5$) for time series extraction was untypically lenient for DCM analyses. The W-C modification might be a way to fix the limitation of DCM to results with large effect sizes. While the possibility of false-positive results due to the more lenient threshold should be considered, results from synthetic data suggest that false positive rates are similar for W-C and bilinear DCM at all levels of noise.

We also implemented the modified model in the two-state DCM, where each region consists of excitatory (glutamatergic) and inhibitory (GABAergic) subpopulations [55], just like the model proposed by Wilson and Cowan (1972) [20]. By performing the Bayesian model selection for the empirical data on imitation and comparing it with a single-state model, we found that the single-state model fits better to data, in opposition to the report in the original paper [55]. However, in both cases (single- and two-state), the W-C model reached a higher probability than the standard bilinear models. One may claim that in our modification, the number of parameters has increased and so it would be trivial to have a better fit in the case of higher complexity. However, as we showed in Fig. 3.5, we received a worse probability in the two-state analysis. Here the number of estimated parameters increases, as we also estimate all the parameters for connecting inhibitory and excitatory neurons in the two-state model. Thus, increasing the number of parameters alone does not explain the better model fit in data sets.

3.4.1. Comparison to other DCM extensions

Since its introduction in 2003, DCM for fMRI has received several extensions and methodological refinements [30, 72, 73]. These extensions include (i) two-state DCMs [55], with separate excitatory and inhibitory populations within each brain region, (ii) nonlinear DCMs with a quadratic state allowing the activity of one region to modulate the connectivity between two other regions [56], (iii) stochastic DCMs which account for random endogenous fluctuations in the neuronal states and inputs [45, 74]. and (iv) spectral DCMs for modeling resting-state fMRI data, which estimate the covariance of the hidden states instead of the states itself [54, 75]. Furthermore, for the large scale brain regions, a linear DCM in the frequency domain using regression DCM [76] has been developed for the task-related fMRI, and in this way, compute for hundreds of regions is now possible [77].

There have been some advances in hemodynamic transfer function (HRF) of the DCM for fMRI that resolves the limitation for the decoupling of BOLD signal and Cerebral blood volume (CBV) [78], which cannot be achieved with the standard hemodynamic model in DCM [34]- and also DCMs with laminar resolution measured with high-resolution fMRI [79]. However, in this study, we use the original hemodynamic model [34] for consistency with the framework mostly used in DCM studies.

Common to all these DCM variants for fMRI is the relative simplicity of the neuronal state equations that utilize parameters that are far from the underlying biology and difficult to match with parameters used in other fields of computational modeling of the brain. On the other end of the spectrum, DCM variants employed for EEG/MEG [31–33] and very recently also for fMRI [58, 60] include complex neuronal state equations consisting of currents and membrane potentials instead of firing rates [58, 60, 80] with up to 4 interacting subpopulations in each region representing different cell type [58,81]. The purpose of this kind of DCM is to make use of both EEG and fMRI data by feeding the fMRI inversion with posteriors estimated from EEG data [59, 60]. However, considering these complex models only for fMRI data may come at the cost that the numerous parameters of these models are much harder to fit the limited amount of data. While it has been argued that the limited temporal resolution of fMRI data can be compensated by spectral information [58], such estimations may also be limited by the strong and diverse fluctuations in the data that cannot easily be attributed to neuronal and hemodynamics sources. The present DCM variant with the W-C equation can compromise between very simple and very complex neuronal state equations. It is easy to implement and opens the possibility of inferring the global input-output properties of the neurons within each region. These properties can be compared to the f-I curves that are measured in electrophysiological recordings of single neurons via data-driven network simulations of a single region [19]. Moreover, the W-C form of DCM has increased the interpretability of the connection matrices (A, B, and C), and now they can be directly interpreted. Below, we elaborate on how the DCM results can be used to constrain local networks, and thus allow inferring about the properties of the involved neurons.

3.4.2. The role of the nonlinearity

The currently established version of the nonlinear DCM [56] was achieved with the further Taylor extension of equation 1 to the quadradic term and defining an additional matrix D for the modulatory inputs from regions. However, the version presented in this study is a formally motivated approach to nonlinearity. Rather than approximating the nonlinear activation function by a Taylor expansion, the W-C equation assumes a particular form for this function, namely a sigmoid function shown in Fig. 3.1. This function is chosen to mimic the input-output relation in real neurons and cannot be emulated by a quadratic term. Additionally, the W-C equation adds a dynamic component in the form of relaxation to a steady-state given by the output of the sigmoid with a specific time constant. However, we found that including this time constant did not contribute much to the model evidence, probably due to the limited time resolution, so we removed this parameter and concentrated on the nonlinearity.

If the input to a given region is close to zero, the sigmoid function is almost linear (Fig. 3.1). In this regime, the extended DCM does not react qualitatively different from the bilinear model. One could argue that the slope parameter is redundant in this regime, as any change in excitability could be compensated by the inverse change in the connection matrix A. However, as the (absolute) input increases (i.e. during the experimental conditions), the nonlinearity of the sigmoid function extends its influence, ultimately leveling off the impact of the input and saturating the output into a limited maximum. This kind of nonlinearity mimics the limited dynamic range of neurons, physiologically incorporated in the form of a depolarization block [82]. In the model, the dynamic range of a region is governed by the sigmoid slope: A large slope implies a narrow range, a smaller slope widens this range (Fig. 3.1). Introducing a limited dynamic range adds an important degree of freedom to the DCM: In a linear neuronal state equation, a large input can lead to a destabilization of the entire system; thus, the strength of the connections is strongly constrained. For this reason, the behavior of the system must be constrained, e.g., by using shrinkage priors or limitations on the sign of the connectivity matrix. Furthermore, the limited input strength can potentially prevent sufficient differentiation between resting and active states. On the other hand, in a Wilson-Cowan-type model, the effect of the input is intrinsically limited by the nonlinearity, so even strong inputs will not destabilize the system, allowing for much more flexibility e.g., choosing priors that reflect knowledge about the connectivity structure. In summary, the linear model approximates more complex dynamics close to a fixed point (given by the resting state) and can thus be destabilized if it is driven away from this fixed point. In contrast, in the W-C model, external input creates a new stable fixed point at the higher activity. In the two-state DCM, the W-C model exhibits an even richer dynamic repertoire, including the possibility of persistent activity and oscillations at various frequencies [83]. Thus, the present extension of DCM considerably extends the range of dynamic behaviors with only a small increase in model complexity.

3.4.3. Implications for the mirror neuron system

Mirror neurons have first been found in the monkey brain and have been repeatedly shown to respond to both executed and observed actions [3,84]. The problem that we encounter is that while mirror neurons are a highly promising candidate to allow interpersonal understanding, we can hardly measure them in the human brain. Thus, studies in humans mainly rely on fMRI or other indirect techniques. For observation and execution of actions, including imitation of facial expressions [66, 69,85,86], studies show activity in regions that are homologs of the primate brain areas linked to mirror neurons, namely in inferior prefrontal gyrus (IFG), inferior parietal cortex (IPL), as well as in superior temporal sulcus (STS), a region of highest importance for action perception, but without own motor neurons. There is only one study directly measuring neurons with such a mirror property [13], and the patients examined had surgeries mainly outside those central regions of the MNS. While the authors showed the activity of neurons in the temporal lobes and further brain regions that are active during observation and execution of actions, more studies are needed to get a deeper understanding of the physiology and functioning of mirror neurons in the human brain. To date, these properties of the mirror neurons are missing even in the monkey literature. Using a two-stage modelling approach (see below) is a way to get closer to the neuronal activity underlying the BOLD signal linked to such a mirroring process in humans.

In a first step, we showed through the empirical fMRI data that the modified model has a higher probability and validate this with synthetic data based on the real data for different values of the signal to noise ratio. The empirical fMRI data on the imitation of emotional facial expressions shows connectivity between STS, IPL, and BA 44. Based on prior knowledge, the driving input was fixed to the STS [23, 69]. The model presented in Fig. 3.7, from the BMA result on W-C DCM, showed a connection from STS to BA 44 and IPL with mutual coupling between BA 44 and IPL. It can be assumed that STS provides visual input, the IPL codes the exact motor action while BA44 codes the motor goal. In the context of predictive coding, this feedback from BA 44 to IPL is used for updating the motor action with the motor goals [23]. Thus, our results are in good agreement with the assumed function of STS, BA 44, and IPL for motor imitation and add empirical evidence for effective connectivity between these regions for imitating facial expressions.

3.4.4. Constraints of local neural networks

While the focus of this paper is the introduction of a nonlinear DCM variant, the ultimate goal of the underlying project is a two-stage modelling approach which uses the results from DCM to constrain an existing, completely data-driven spiking network model [19] to construct a computational model of the human MNS. The spiking network model has previously been shown to be a statistically accurate description of rodent neural activity in vivo. However, the parameters of the model were adjusted exclusively by in vitro anatomical and electrophysiological data [19]. In this way, the nonlinear extension of the DCM technique presented here allows combining local modelling with constraints from animal experiments and global modelling with constraints from fMRI data. To account for the potential differences in species and brain regions, we introduced a number of global scaling parameters for the neural and synaptic properties, which are being adjusted to the DCM results [87]. These adjustments will lead to a model with unprecedented predictive power about the physiological properties and the temporal dynamics of the human MNS, as it accounts both for the global dynamics in humans obtained from DCM and for the detailed neuronal machinery on the level of local circuits, which are likely to be conserved across species. In contrast to existing, more abstract models of the MNS [7, 8, 88–90], this model holds the promise to capture detailed neuronal phenomena such as the suppression of the mu rhythm [91, 92] or the modulation of the MNS by neurotransmitters [93–95].

4. Effective connectivity of the human mirror neuron system during social cognition

The human mirror neuron system (MNS)[1] can be considered as 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 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 different 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.

4.1. Introduction

Mirror neurons are considered essential building blocks to present the neuronal basis of social cognition [5, 96]. It is assumed that humans obtain an immediate understanding of others' emotions, desires, and intentions by representing their motor states in their motor system [5, 97, 98]. A large body of fMRI and EEG studies have provided indirect evidence for the involvement of mirror neurons in social-cognitive processes (see [99] for a meta-analysis), including imitation [100, 101], action learning [102], emotion recognition [103], theory of mind (ToM) [4,66], and empathy [92, 104]. However, research is needed to explain how regions of the MNS interact to understand others' emotions and intentions. Studies using DCM or related methods for inference about effective connectivity can help to get a deeper understanding of sensory-motor processing in the MNS (see [105] for a review) and inform computational models that model the physiological processes in the MNS.

---

[1] This chapter is submitted to Social Cognitive and Affective Neuroscience (in revision).

MNs were discovered by di Pellegrino and colleagues in 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 [3, 106]. Due to ethical reasons, direct examination of mirror neuron 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 mirror neuron system. These studies in humans identified several brain regions with mirror properties [9]. Among these, the regions that are closely linked to primate data of mirror neurons and that are the basis for models explaining mirror neuron function are Broca's area (BA44) located in inferior frontal gyrus (IFG) with adjacent ventral premotor cortex that corresponds to area F5 of the primate brain [107], as well as inferior parietal lobule (IPL) [106,108]. In addition, the posterior superior temporal sulcus (STS) has been suggested as the region that conveys the visual input to the mirror neuron system [109]. These regions of the human MNS have been mainly identified in studies with visual images of actions and execution of motor actions [9, 110]. However, a special interest in mirror neurons exists due to their proposed role in social cognition [5, 66]. Recently, we demonstrated common activation for imitation, empathy, and affective ToM across and within participants in STS, IPL, and BA44 during social cognition tasks [14]. 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 MNS for mental disorders, such as in autism [111,112], psychopathy [12], schizophrenia [67] and borderline personality disorder [11]. Thus, a deeper understanding of the human mirror neuron system 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 mirror neurons (for a review, see [113]). 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 [114, 115]. Especially in humans, direct connections from STS to IFG have also been found [116, 117]. Based on this connectivity profile and the mirror properties of IPL and IFG, common assumptions of the models of the MNS are that a) visual information enters the MNS via the STS, b) motor information is transferred from IFG directly or via the parietal cortex to STS, and in several cases, c) 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., [118]), neural fields (e.g., [119]), layers of deep neural networks (e.g., [120]) 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), [121]) or Bayesian predictors (where IFG and IPL act as empirical priors for STS, [122, 123]). 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 ([118–120, 122, 123], but see [124] 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 [16]. The few existing DCM studies on the MNS have largely confirmed the connectivity profile outlined above [89, 125–127], including the direct pathway from STS to IFG [89]. 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 [128], 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 [129]. To date, only parts of the MNS have been studied with DCM while watching emotional facial expressions [130, 131] or social scenes [132]. 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 [53]). 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. Out of the three tasks, only imitation included actual movements. All tasks were based on pictures of facial expressions. Activation patterns of these tasks are published in Schmidt et al., (2021) [14]. By applying 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, it was demonstrated that 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 [53] was applied. We designed 540 models divided into four families and used Bayesian model comparison and Bayesian model averaging 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.

4.2. Materials & Methods

4.2.1. Data sets

Participants

We invited 80 participants to two appointments. Of these, we excluded one due to excessive head movement (more than 2 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. 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 [53] 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, 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) [14].

Imitation The imitation task contained four conditions (Fig. 4.1A). 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 each 5 seconds (Fig. 4.1B). Control blocks contained two stimuli each 3 seconds and were presented interleaved with the experimental blocks. A fixation cross with jittered duration was presented between trials (for 1-3 seconds) and between blocks (for 4-6 seconds). 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 minutes.

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 4 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 empathize with the presented person?' (Affective Empathy), or 'How big is the circle?' (Control condition)(Fig. 4.2A). 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'; Fig. 4.2B).

Analogous to the imitation task, we chose a design with experimental blocks of 4 stimuli each 3 seconds alternating with a control block of 2 stimuli each 3 seconds. A fixation cross with jittered duration was presented between trials (1-3 seconds) and between blocks (4-6 seconds). Again, there were 20 trials for each experimental block and 30 total control trials. The task duration was 17 minutes.

Theory of Mind (ToM) The ToM task had three experimental conditions Affective ToM, Emotion Recognition, Neutral Face Processing, and a control condition (Fig. 4.3A). Conditions were presented in pseudo-randomized order in an event-related design (Fig. 4.3B). 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

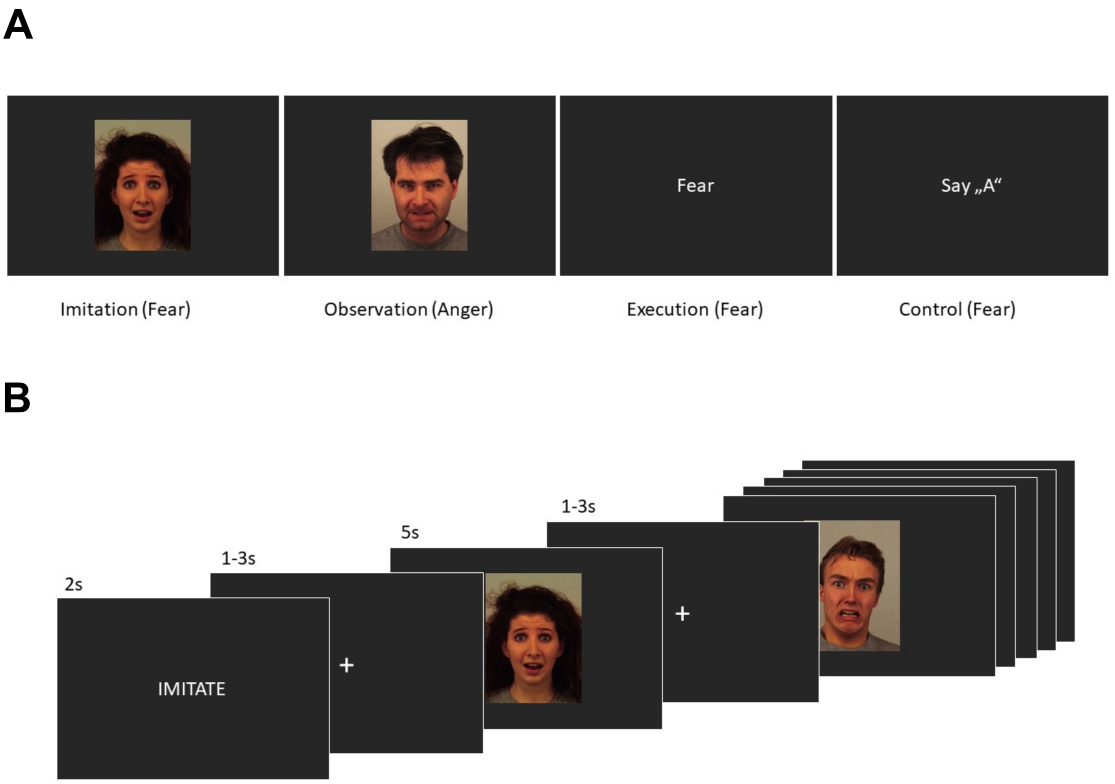


Figure 4.1.: Imitation task. A) Overview over the four conditions imitation, observation, execution, and control, with exemplary stimuli. In all conditions except control, half of the stimuli showed angry, the other half fearful facial expressions or word cues. The control condition served as a motor 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, 4 stimuli with a duration of 5s were presented, in the control blocks, 2 stimuli with 3s duration.

statement and face was presented for 2s, and the inter-stimulus interval lasted between 1 to 3s. The task duration was 8 minutes.

Data acquisition

fMRI data was 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 $1 \times 1 \times 1\ mm^3$). For recording of task activation, echo-planar imaging was set to 32 descending $3 \times 3 \times 3\ mm^3$ 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 minus 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 3mm translation or 3° rotation, by replacing the scan with the mean of the below-threshold scans surrounding it.

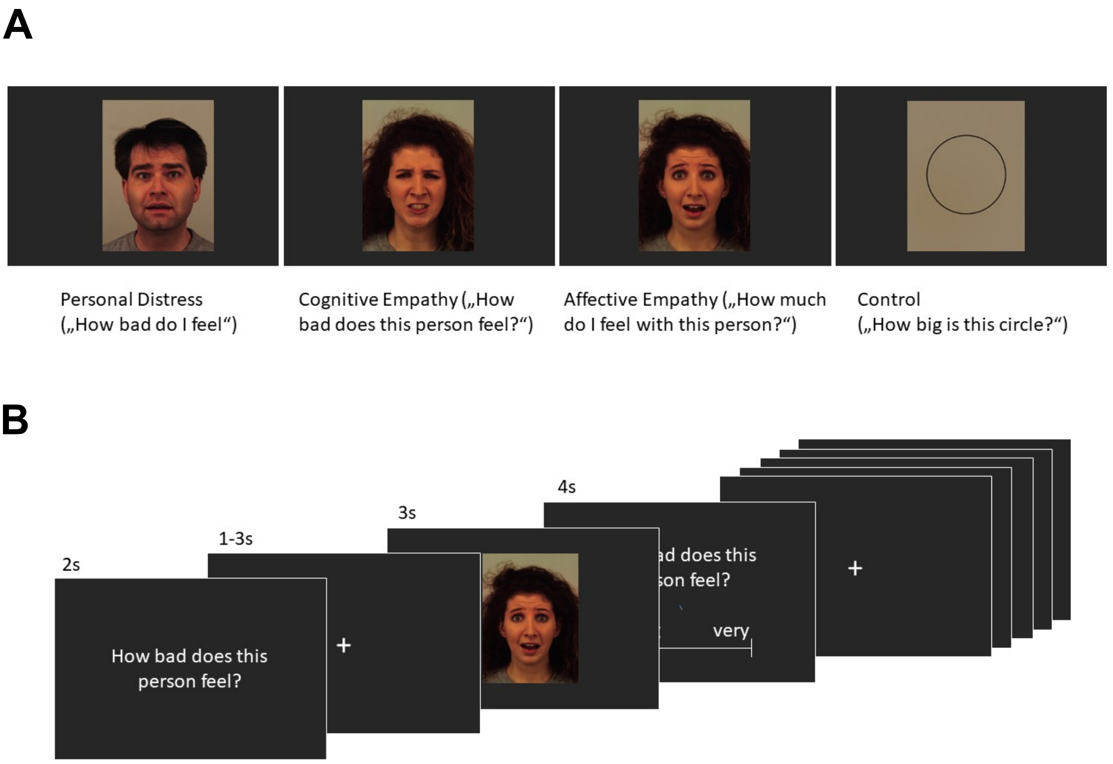

A
Personal Distress („How bad do I feel“)
Cognitive Empathy („How bad does this person feel?“)
Affective Empathy („How much do I feel with this person?“)
Control („How big is this circle?“)
B
2s
1-3s
3s
4s
How bad does this person feel?
+
ad does this
son feel?
very
+

Figure 4.2.: 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 analogue scale. B) Task flow with presentation times. At the beginning of each block, the cue question was presented. In the experimental blocks, 4 stimuli with a duration of 3s were presented, in the control blocks, 2 stimuli with 3s duration.

Time series extraction

Data were preprocessed and analyzed with Statistical Parametric Mapping 8 (SPM8, `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`). Preprocessing consisted of slice time correction, realignment to the mean image, normalization with segmentation and co-registration to the individual MPRage, and resampling with $3 \times 3 \times 3\ mm^3$ 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. 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

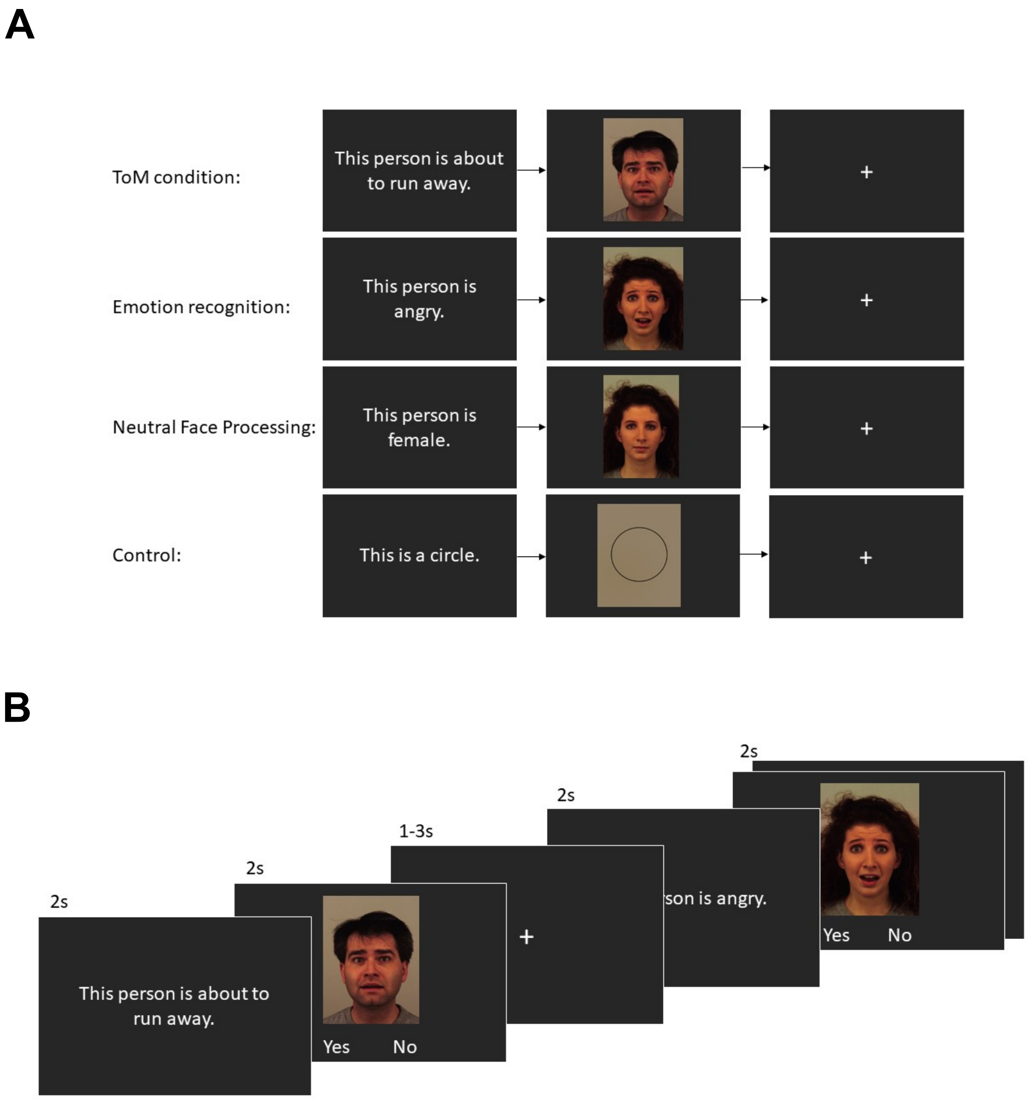


Figure 4.3.: ToM task. A) Overview over the four conditions ToM, emotion recognition, neutral face processing, and control, with exemplary stimuli. Faces in the ToM and emotion recognition condition showed angry and fearful expressions, in the neutral face processing condition neutral expressions, and 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.

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 region of interest (ROI) in previous publications [66, 67].

4.2.2. 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 [16].

We performed the DCM10 (r6313) in SPM8, which we have modified before, by replacing the standard linear equations with Wilson-Cowan-based models [53]. This model shows the changes in neural activity according to:

$$\begin{aligned} \dot{z}_t &= -z_t + S(x) \\ S(x) &= \frac{1}{1+\exp(-\alpha * x)} - \frac{1}{2} \\ x &= (A + \sum_{j=1}^{m} u_j B^j) z_t + Cu \end{aligned} \tag{4.1}$$

$\dot{z}_t$ denotes the time derivative of neuronal activity, and function $S(x)$ shows a sigmoid function in which the parameter α determines the slope of it. 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 [53], we showed that the modified DCM allows a significantly better fit to the empirical data than the standard bilinear model. 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), (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 [23, 69, 133]. 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 Fig. 4.4, a feed-forward connection from STS to BA44 and IPL is always available for family one. 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 Fig. 4.4 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 two, the activation propagates from STS to IPL and then from IPL to BA44, and for family three, from STS to BA44 and then to the IPL region. In family four, 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 one 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) [17] among all single models with inference over families of models [24]. To account for the heterogeneity of the model structure across subjects, we used random effects (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

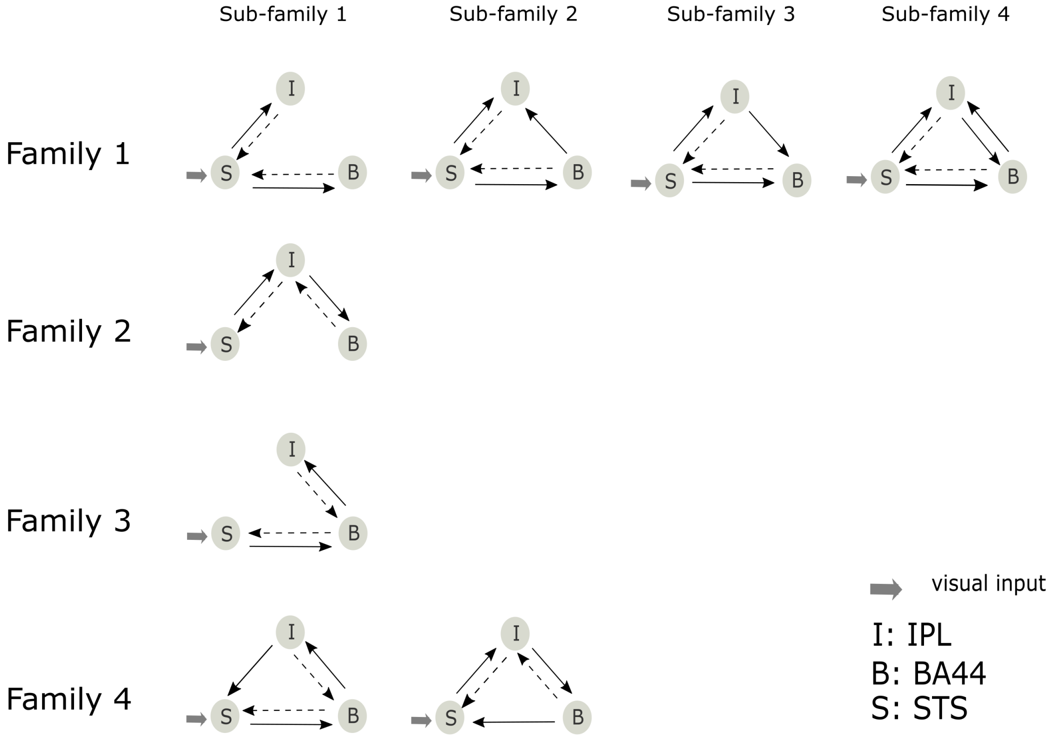


Figure 4.4.: Model space. Schemata of parameters that made up the models included in four families. Solid lines show the connections that always are present, and dashed lines the connections that can be present or not. The modulatory input can be exerted on these interregional connections. Family 1 consists 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.

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 [134] (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) [24].

Still, finding sufficient evidence for one model or family of models being optimal is not always possible. Bayesian model averaging (BMA) helps resolve 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 [24]. 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 Fig. 4.4), 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.

4.3. Results

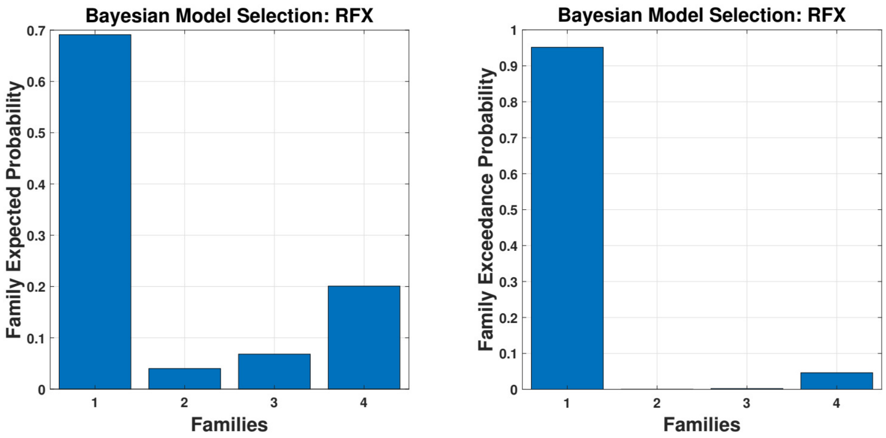
Using the Bayesian model comparison, we first used the family level inference to find which set of models in Fig. 4.4 (divided into four families) is selected over other families. Results indicated that family 1, in which input from the STS region propagates to BA44 and IPL with feed-forward connections, is the most probable structure for all three experiments. The exceedance probability for all experiments is larger than 0.9 (Fig. 4.5) 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 most significant probability for each of the experimental tasks (Table 4.1). 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 4.1), 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.

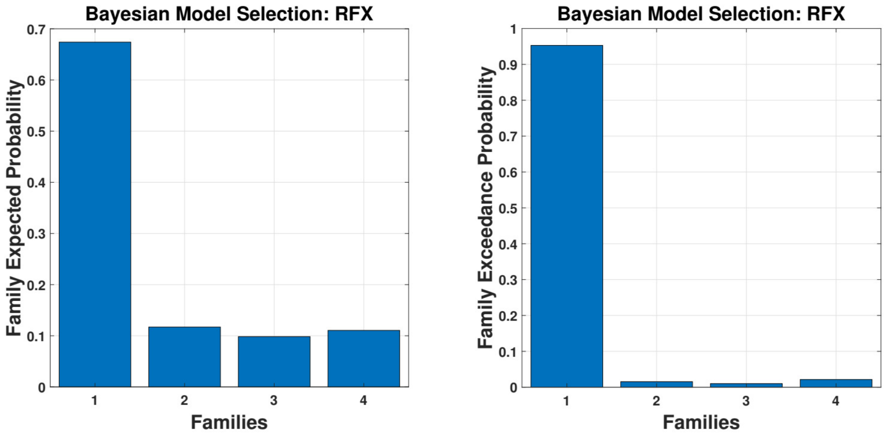
Table 4.1.: RFX BMS results on models within family 1 for all experiments.

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

Imitation



Empathy



Theory of Mind

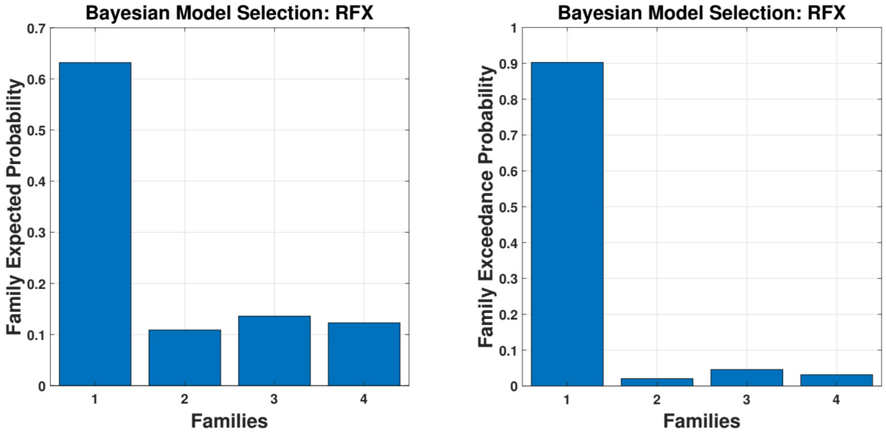


Figure 4.5.: Family level inference performed on models within the families in Fig. 4.4. All experiments demonstrate that family 1 has highest expected and exceedance probability, in which the models within this family have forward connections from STS to IPL and BA44.

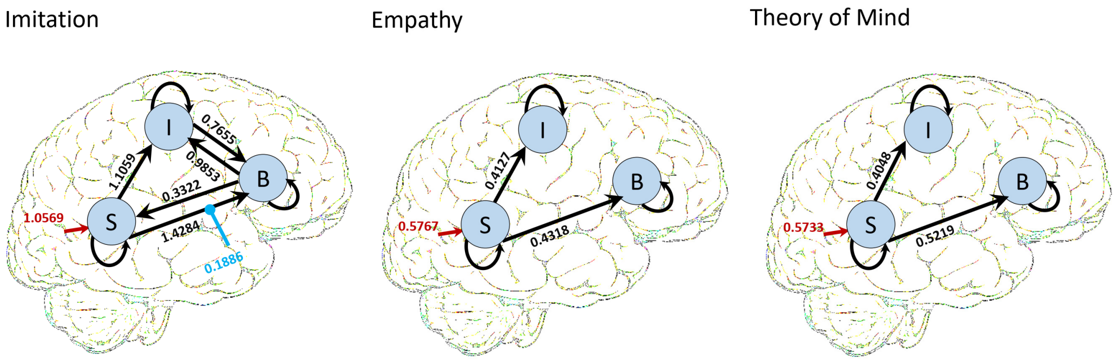


Figure 4.6.: BMA results for all three tasks for the winning family one. Here we illustrate only the parameters which are significantly > 0. The values for external inputs (Matrix C) which are not reported in Tables 4.2 and 4.3, are shown here (all parameters are in Hz).

Table 4.2.: BMA results for endogenous connectivity (Matrix A) (in Hz). Next to each parameter is the posterior probability which is different to 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

However, as shown in Table 4.1, for empathy and ToM, the exceedance probabilities of the winning sub-families do not provide definitive evidence (0.49 max) for them. Therefore, as a third step, we use Bayesian model averaging (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 4.2, 4.3, and Fig. 4.6 for all three experimental tasks. These results for empathy and ToM are on all models within family one and for imitation only within the winning sub-family 4. In Fig. 4.6, 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

Table 4.3.: BMA results for modulatory connectivity (Matrix B)(in Hz). Next to each parameter is the posterior probability which is different to 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	1.1059, 1.00	-	0.9853, 1.00
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	-

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.

4.4. 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 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 [113, 114], termed an inverse model in the context of controller architectures [121]. However, contrary to most MNS models [114], a forward model from motor to sensory areas is missing, as well as a clear hierarchy of areas within the MNS [118–120,122–124]. 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 that information about exact motor states 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 [66]. Overall, our results support the notion of a different processing route for emotional facial expression than the one that has been suggested by connectivity data in monkeys [129] that revealed 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 [135–137] and the prefrontal cortex [132,138,139], both in agreement with the results from monkey studies [129]. 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 and colleagues (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 pSTS region [140]. Furthermore, both regions share effective connectivity with the anterior cingulate cortex (ACC) in this study, which has a prominent role in attention [141].

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 [142]. Future studies are needed to disentangle these possible processes, e.g., by using an attentional condition and a social-cognitive condition, as in Schuwerk and colleagues (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., contraction of 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 [143]. 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 [114, 121].

4.4.1. 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 [105, 144]. 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 activation and connectivity patterns suggest an active imitation of the participants [14], 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 [5, 145], to involve information flow from STS to IPL and BA44. Further, the results of the modified DCM analysis [53] 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 [19, 146]. 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 mirror neurons, 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 [5]. 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 [147–149].

5. Detailed spiking network model of the human mirror neuron system

One way to learn more about the human mirror neuron system without directly measuring individual neurons' activity consists of the theoretical modeling of the involved cell assemblies. The mathematical description of neuron networks' activity and the simulation of the dynamics makes it possible to calculate the non-invasive measurement methods' indicators and compare them with the measured values. Using statistical optimization methods, the free parameters of the model can be fitted to the experimental data. Thus, statements about the cell assemblies' physiology are possible since the model's parameters are directly related to biophysical properties, e.g., cellular and synaptic conductivities and resting potentials.

Here, we use a highly detailed network model of the prefrontal cortex (PFC). In this network, all neurons and synapses parameters are determined by anatomical and in vitro electrophysiological data, which has previously been shown to statistically reproduce a wide range of measures from in vivo prefrontal data [19]. We adapted this spiking network model to the fMRI data; in the first step, the effective connectivity between the activated regions has been identified with DCM by comparing 540 models (chapter 4). Then some modifications were done on the DCM approach to use the (nonlinear) Wilson-Cowan-type model instead of the standard DCM schemes (chapter 3). The global connectivity was inferred from the W-C DCM, and the connections between regions were made according to the best model, which was found from DCM analysis. The neurons' input-output functions in the firing-rate model are matched with the predicted data from DCM analysis by comparing the resulting outputs, thus realizing the transfer from the macro- to the micro-level. This model can be used to predict the task performance and make predictions about a completely different set of data to make statements about the physiological properties of the human mirror neuron system.

Here I want to give a brief introduction about the detailed spiking network model of the PFC (full description in detail can be found here [19]) and then a detailed explanation about how the implementation of the DCM model on the spiking network model has been done.

5.1. Spiking network model of the prefrontal cortex

The detailed Spiking network model of prefrontal cortex (PFC) is based on a simple computationally tractable single neuron model (simpAdEX) [150], with all parameters derived from in vitro electrophysiological and anatomical data [19]. The simulated model consists of 1000 Neurons divided into two laminar network structures, namely the superficial layer L2/3 and deep layer L5, organized in a single column. Each layer includes five cell types, one type excitatory (pyramidal cells (PC)) and four types inhibitory (fast-spiking (FS), bitufted (BT), Martinotti, and large basket interneurons) neurons. Specifically, the four different types of interneurons are defined as local interneurons (IN-L, fast-spiking interneurons) project within the same layer and column, cross-layer interneurons (IN-CL, bitufted cells), cross-column cells (IN-CC, large basket cells), and far-reaching interneurons (IN-F, Martinotti cells) with projections both across layers and columns.

These neurons are connected randomly with the different connection probabilities and through conductance base (AMPA and NMDA (excitatory), and $GABA_A$ (inhibitory)) synapses, equipped with short-term plasticity (STP) dynamics [151,152] and synaptic delays. There are three STP classes for either excitatory or inhibitory connections; facilitating, depressing, or a combination (early facilitation and late depression). The cell types of the pre- and post-synaptic neurons (pyramidal cells or interneurons) determine which classes are used for each combination.

All neurons are driven by constant DC currents of 250 pA to all pyramidal cells and 200 pA to all interneurons for both layers. These currents represent the synaptic connections from outside of the network, and they are the only parameters that are not directly obtained from experimental data. These background currents are treated as free parameters and are estimated from the simulated network activity itself (see [19] for details).

This model has been validated by reproducing a wide range of in vivo statistics, including single-cell spiking trains, the local field potential data from awake rodents, and the fluctuations of the membrane potential [19]. Furthermore, it is used to generate persistent activity of single-cell assemblies within the spiking network model to present a working memory model [146]. It is investigated how the persistent activity is preserved with the homogeneous excitability of the interneurons, as well as a homogeneous distribution of synaptic inputs and short-term plasticity.

5.2. Implementation of the DCM model on the spiking network model

In chapter 3, I modified the DCM approach to use the (non-linear) W-C model instead of the more simple, strictly linear type of model in the standard DCM schemes. The main reason for this modification is that we want to use the results from the macroscopic model (e.g., the model used to fit the fMRI data) to constrain parameters of the detailed network model of a single brain region. This is only possible if the macroscopic model has a direct relation to the detailed, microscopic model. In the W-C model, one can interpret the sigmoid transfer function as an averaged f-I curve of many neurons in a single region. Thus, estimating the parameters of the model (ideally individually for the pyramidal cells and interneurons of each region) allows us to constrain the neuronal properties of the detailed model, along with the synaptic parameters. In contrast, there is no obvious relation to a microscopic network model in the bilinear model, and even the synaptic parameters (A, B, and C matrices) may be different from those in the more realistic Wilson-Cowan model.

In this chapter, I adjust the firing-rate model to the fMRI data by Inferring the global connectivity (i.e., the scaling factors used for the connection probabilities and synaptic weights) input-output functions of the neurons in the spiking model from the firing-rate model and compare the resulting outputs. This is where the transfer between the macro- and the micro-level occurs.

The general concept is to set up three networks of spiking neurons (1000 neurons each) using the firing rate model. Each network represents one of the three regions included in the DCM analysis and constructs a firing rate signal by averaging the firing rates of all the spiking neurons in each of the pyramidal cells and interneurons. The connectivity structure of this network is simulated according to the network structure achieved from W-C DCM. The overall connection strength is governed by a scaling factor, which can be set to different values reflecting the connections found based on the fMRI data. Furthermore, neuron parameters should be adjusted to reflect the input-output relations found in the Wilson-Cowan model, on average. Changing the f-I curve of these neurons shows how this behavior changes, and I try to match it with a W-C model's behavior. In this way, the spiking network is constrained by the data found in the fMRI data.

The following steps should be done in order to fit the spiking model to the W-C DCM results:

1. Computing the sigmoid function parameters for L2/3 pyramidal cells and interneurons separately.

2. Using the two-state W-C model to fit the fMRI data, with the sigmoid parameters from the spiking network as priors for the sigmoid parameters.

3. Modifying the cell parameters of the pyramidal cells and interneurons so that the sigmoid function parameters from the two-state W-C model are fitted.

4. Connecting three copies of the current network with the modified cell parameters and the connectivity from the W-C model and comparing the resulting firing rate dynamics with the W-C DCM.

The result is a spiking network adjusted to the W-C model and, thus, to the fMRI data. This process will be the basis for all further experiments. In the following, I explain these steps in more detail.

The first step towards this goal is to establish the connection between the mean firing rate dynamics and the neuron parameters to show how a given input-output relation in the W-C model can be mimicked in the spiking network. There is no one-to-one relation to those dynamics between the W-C model and the spiking network model, so I try to establish this connection experimentally. More precisely to say, the key variable of the W-C model is firing rate (both input and output), while in the spiking model, it is spike times (output) and synaptic currents (input). Thus, a connection between these variables needs to be established. In the spiking network, the spike times can be converted into firing rates (using kernel density estimation (KDE), see [153, 154]), and firing rates can be converted into synaptic currents (by applying synaptic dynamics to a Poisson spike train with that rate). In this way, the spiking network can be probed (to start with a single area) by applying a spike train with a given firing rate and measuring the resulting firing rate. An input-output curve, equivalent to the W-C model, is obtained by varying the input firing rate. Thus, one can directly compare the W-C network's sigmoid function and the input-output function of the spiking network and effectively manipulate parameters in the spiking network to mimic a particular sigmoid function.

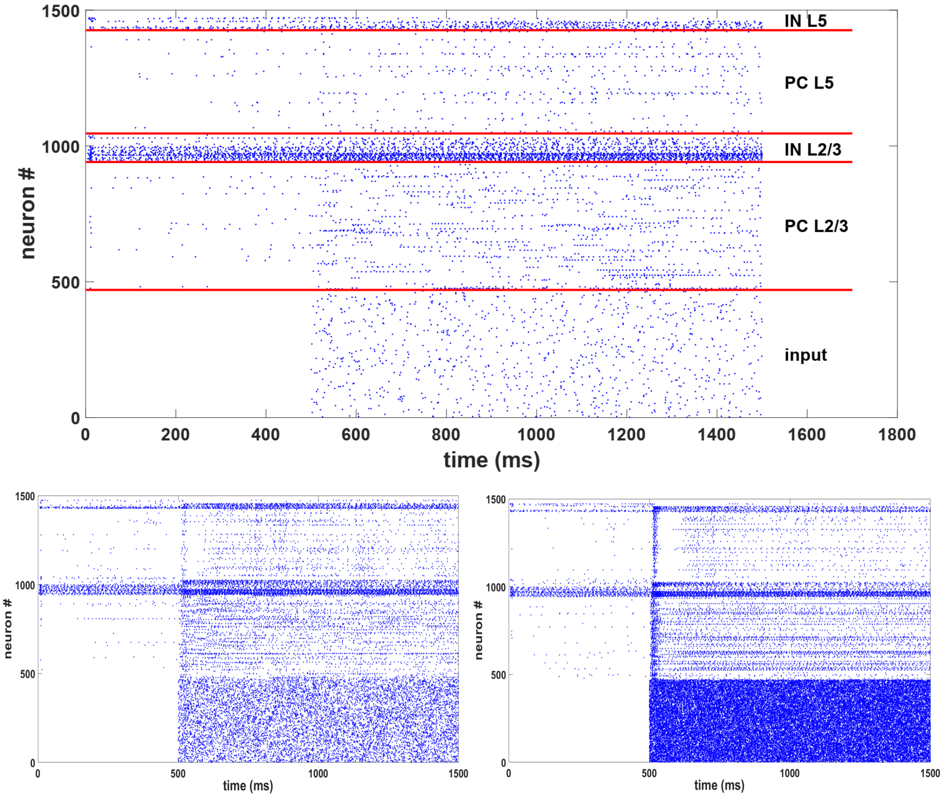


Figure 5.1.: Raster plot of the spike times in the network in response to the spiking input into the pyramidal cells L2/3. The pyramidal cells (PC) and the interneurons (IN) of each layer (L2/3 and L5) are separated by red lines in the first plot. There are three different raster plots for three different firing rates of the input from low to high frequency. The neurons' responses to the stimulus are clear from the raster plots for the transient (500-1000 ms) and the steady-state (1000-1500 ms). Here, I let the simulation run for 500 ms and then give the spiking input to the region.

The spiking input into the excitatory neurons (PC) is defined as one input neuron with an individual Poisson spike train for each of the 470 pyramidal cells in L2/3. Extending the input to the entire simulation time (after a rest period of 500 ms, see Fig. 5.1) and varying the input firing rate, one can derive the network's input-output curve in terms of firing rates.

Here I fit an actual sigmoid function to the f-I curve and compare its parameters

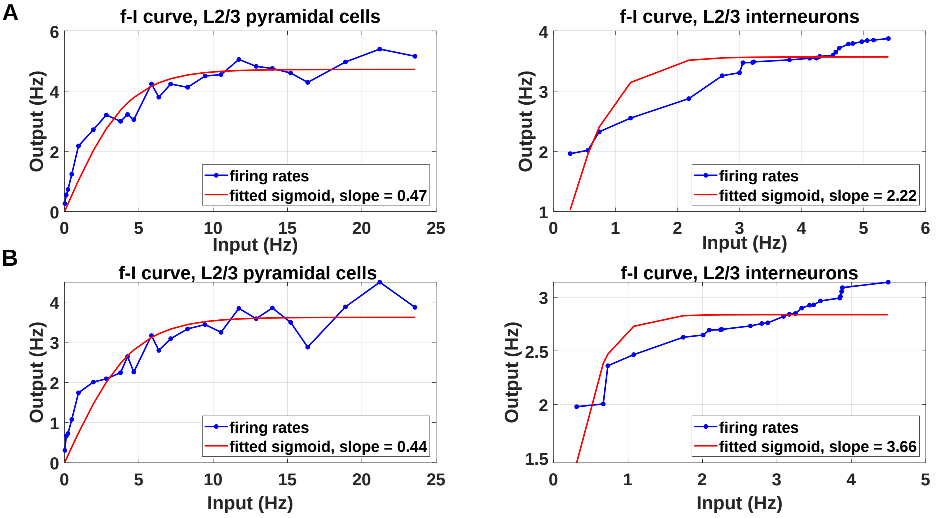


Figure 5.2.: Spiking network model f-I curve for a single region. The plots show the firing rates with the fitted sigmoid function for L2/3 pyramidal cells and interneurons as A) transient curves (short period right after the stimulus) and B) steady-state (the end of the simulation). For the excitatory neurons (PC), the input is the firing rates of the Poisson spike trains, and the output is the averaged firing rates of the pyramidal cells. For the inhibitory neurons (IN), the input is the averaged firing rates of the PC neurons, and the output is the averaged firing rates of the interneurons.

to the W-C DCM. As can be seen in Fig. 5.2, the offset of the sigmoid is close to zero, which is also used in the W-C model in chapter 3. To sharpen this observation, I let the simulation run for a zero input rate and sample also a few minimal input rates. The output rate data points are the averaged firing rate over a short period (500 ms) right after the stimulus and at the end of the simulation (last 500 ms). These two regimes denote the transient (Fig. 5.2A) and the steady-state (Fig. 5.2B) response to the stimulus, which are different if one looks at the f-I curve and fitted sigmoid, and also the raster plot of the spiking times in Fig. 5.1. As shown in Fig. 5.1, I let the simulation run for 500 ms and then give the spiking input to the network to avoid transient effects at the beginning of the simulation. The fitted sigmoid slopes of each f-I curve are also reported in Fig. 5.2. One crucial thing in Fig. 5.2 is that, when calculating the input-output curve of the model, to relate the firing rate of the input neurons to the firing rate of the model neurons. It is correct for the pyramidal cells in L2/3, which receive this input directly but not for the L2/3 interneurons, which indirectly receive the input via the L2/3 pyramidal cells. Thus, I compute the input-output curve with the pyramidal cells' firing rates on the x-axis and not the input firing rate. As can be seen, this curve has a steeper slope than the pyramidal cells, as this is the case for the individual neurons [150].

The next step (step 2 above) is to use these parameters (computed sigmoid slope of excitatory and inhibitory neurons) as priors in the W-C two-state DCM (Chapter 3) and fit our best model (chapter 4, Fig. 5.3A) to the fMRI data. Then, I use the posterior values (estimated parameters) of the following parameters (Table 5.1) from the modified two-state DCM in the spiking network model: self-excitation (excitatory, SE), self-inhibition (inhibitory, SI), excitatory to inhibitory (EI), inhibitory to excitatory (IE), and excitatory to excitatory connections between regions (EE), in addition to modulatory and external inputs. These parameters are shown in Fig. 5.3B for the two-state DCM.

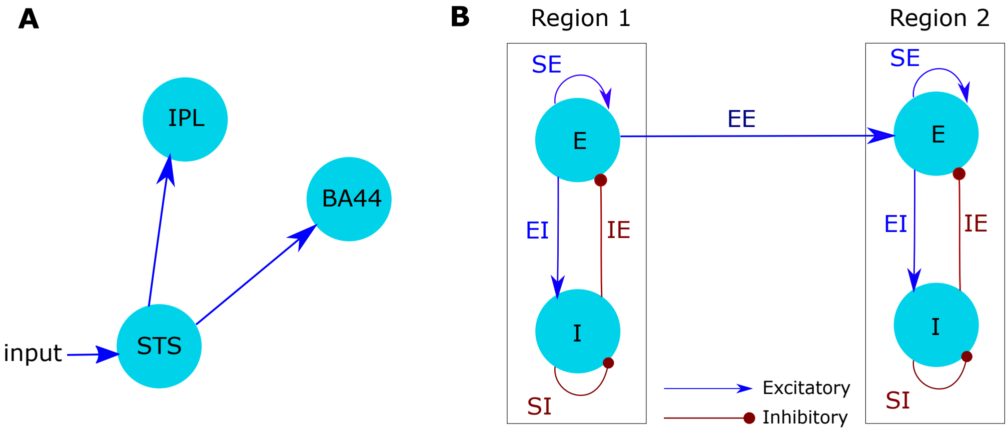


Figure 5.3.: A) Network structure achieved from the previous chapter for the global network of the human mirror neuron system (see chapter 4). B) Illustration of the two-state neuronal model implemented in the DCM for two distinct regions. Each region consists of the neuronal populations E (excitatory) and I (inhibitory). The estimated parameters used in the spiking network model are as follows: SE=self-excitation, SI=self-inhibition, EE=excitatory to excitatory, EI=excitatory to inhibitory, IE=inhibitory to excitatory. Please note here that only EE parameters are between regions, and the rest are related to the parameters within one region (for more details, see chapter 3).

For this purpose, I need to connect three stripes or columns (each 1000 neurons) of the spiking network model in the same as the global network of the mirror neurons found in the previous chapter (Fig. 5.3A) with the parameters achieved from DCM analysis (Table 5.1) as explained above. The time series of the predicted data from W-C DCM and the firing rates from the spiking network model should eventually match. Thus, the two time-series are scaled in the same way to compare them and conform to the respective regions' amplitudes. There are some adjustments in the DCM estimation in SPM software; The signal is first mean-corrected, then scaling is applied to ensure the maximum signal change is 4%. Furthermore, as the experimenter can not control the input frequency for fMRI data (since the input comes from adjacent cortical areas), I treat the external input (firing rates of the Poisson spike trains) as a free parameter for the spiking network model.

Figure 5.4 shows the spiking times for three stripes (regions) connected as Fig. 5.3A and compares the simulated time series with the predicted data from DCM analysis. The simulation's stimuli timing is based on the block design fMRI data measurements introduced in chapter 3. To measure the matching between the two time series, I have used the linear correlation coefficient between them for each region, as shown in Fig. 5.4B. The firing rates of the spiking times in Fig. 5.4B are represented using kernel density estimation [153, 154] with a large kernel

Table 5.1.: The posterior expected values (mean) and the probability of each value for the parameters shown in Fig. 5.3B, and the sigmoid slopes of excitatory (E) and inhibitory (I) neurons achieved from DCM analysis.

Parameters	**Expected Values (mean)**	**Probabilities**
SE	0.88	0.9966
SI	1.06	0.9834
EI	0.78	0.9435
IE	0.78	0.9435
EE (STS to IPL)	1.24	1.00
EE (STS to BA44)	1.096	1.00
Slope-E	1.6	1.00
Slope-I	2.17	1.00
External Input	0.45	0.9979

bandwidth to have a smooth firing rate to resemble the time series of the predicted data.

The time scales for the spiking network model and the fMRI data are milliseconds and seconds, respectively. Because of this difference between the time scales, the firing rate maximum of the spiking model may not be reached. While it is not efficient to run the simulation for hundreds of seconds, I have used a factor of 10 ms instead (10 ms of the simulation is equivalent to 1 second in the fMRI data) and scaled the stimuli accordingly to reach the maximum value of the firing rates in the spiking network model (longer time means more inputs, and more inputs increase the firing rate). In this way, the peaks of the two time series (predicted and simulated) are more similar to each other. The result depicted in Fig. 5.4B (correlations between the two time series) shows a pretty good matching between the time series. However, the simulated time series still need to match quantitatively (up to some degree of noise) with the predicted time series achieved from the W-C DCM. In the following steps, I try to match these two time-series better with the interplay between the f-I curve and the connectivity. Then, I attempt to improve the fit by changing the neurons' f-I curves according to the DCM results via varying the neurons' parameters.

Regarding the interplay between the f-I curve and the connectivity, I measure the f-I curve for the three stripes network and compute the fitted slope of the sigmoid function in the same way as the single stripe of neurons to see if the f-I curve has changed with three stripes (note that the inputs that generate the f-I curve are inserted on top of the synaptic inputs from the other regions and the stimuli mimicking conditions from the fMRI experiments). More precisely, according to the four steps above, I compute the changed f-I curve from the spiking network and use its parameters as priors in DCM analysis to calculate (potentially different) new connectivities. I put these connectivities into the spiking network again to calculate a new f-I curve and iterate this until the difference between the two subsequent f-I curves is small. Figure 5.5 shows the variant runs of the f-I curve and the fitted sigmoid for the excitatory and inhibitory neurons of layer 2/3 in the

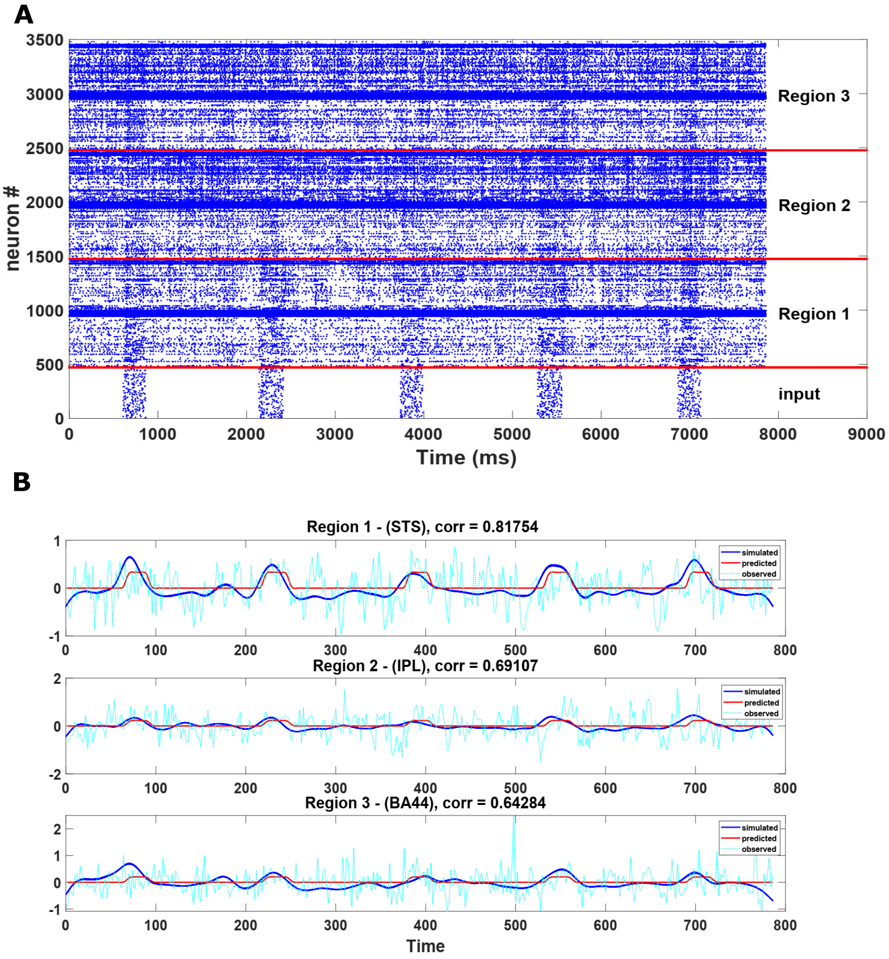


Figure 5.4.: A) Raster plot of the spike times for a network of three regions (each 1000 neurons) in response to the spiking stimuli (470 neurons) into the pyramidal cells L2/3. The time intervals of the inputs are determined as the fMRI data measurements. B) Time series of the observed (cyan dotted line), predicted (red solid line), and simulated (blue solid line) data sets for the network of three regions connected as Fig. 5.3A. The correlation of the simulated and predicted data is shown for each region. The time scale for the spiking network is ms and for the fMRI data in seconds. Here I have rescaled the time for simulated data (10 ms of the simulation is equivalent to 1 second in the fMRI data).

first region. In the same process as the single stripe, I use the slope parameters shown in Fig. 5.5 as priors in the W-C DCM and use the estimated parameters in the three regions' spiking network.

Here I consider only the f-I curve parameters of the first region and transient state. In order to have a smaller complexity in the model evidence, it is optimal to have a minimum number of parameters. In this way, I consider only the first region and its value of fitted slope as priors for all three regions in the W-C DCM.

After six runs, I could have a relatively small difference between two subsequent f-I curves for the excitatory and inhibitory neurons (Fig. 5.5). The final parameters are set according to the last f-I curve in Fig. 5.5. As stated by the predicted parameters reported in Table 5.2, the f-I curve is also altered in the DCM analysis,

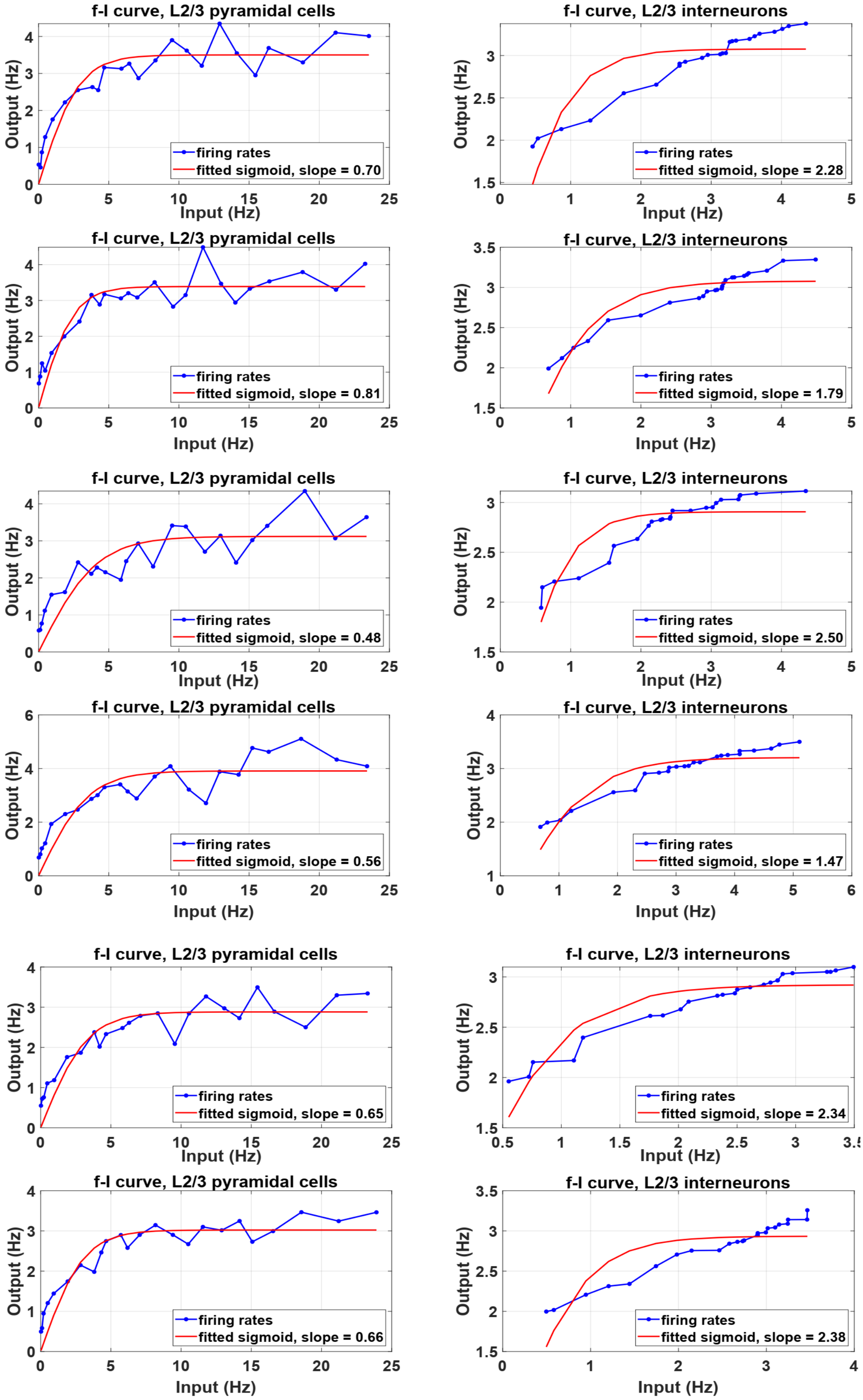


Figure 5.5.: Spiking network model f-I curve for a network of three regions connected as Fig. 5.3A for six different runs. Each row (run) shows the firing rates and the fitted sigmoid function for L2/3 pyramidal cells and interneurons of the first region that receives the external input. The inputs and outputs are the same as the single region in Fig. 5.2, and here I consider only the transient state.

Table 5.2.: The posterior expected values (mean) and the probability of each value for the parameters shown in Fig. 5.3B, and the sigmoid slopes of excitatory (E) and inhibitory (I) neurons achieved from DCM analysis. These parameters are conducted from the prior values shown in Fig. 5.5 for six runs and the slopes implied by DCM.

Parameters	**Expected Values (mean)**						
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Implied by DCM
SE	0.88	0.886	0.885	0.894	0.90	0.876	0.68
SI	1.06	1.05	1.07	1.04	1.06	1.07	1.05
EI	0.78	0.80	0.77	0.82	0.77	0.77	0.86
IE	0.78	0.80	0.77	0.82	0.77	0.77	0.86
EE (STS to IPL)	1.24	1.23	1.23	1.24	1.24	1.24	1.1
EE (STS to BA44)	1.09	1.07	1.09	1.08	1.08	1.09	0.91
Slope-E	1.64	1.70	1.61	1.64	1.67	1.62	2.82
Slope-I	2.23	1.73	2.45	1.39	2.19	2.33	2.15
External Input	0.445	0.41	0.45	0.42	0.42	0.46	0.21

Parameters	**Probabilities**						
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Implied by DCM
SE	0.996	0.997	0.997	0.997	0.997	0.996	0.995
SI	0.983	0.982	0.984	0.981	0.983	0.984	0.983
EI	0.943	0.948	0.94	0.952	0.942	0.942	0.965
IE	0.943	0.948	0.94	0.952	0.942	0.942	0.965
EE (STS to IPL)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EE (STS to BA44)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Slope-E	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Slope-I	1.00	0.999	1.00	0.997	0.999	1.00	1.00
External Input	0.998	0.998	0.998	0.998	0.998	0.998	0.997

i.e., the prior and posterior values of the slope parameters are different comparing the fitted slopes in Fig. 5.5 and values in Table 5.2. In this way, I also need to have a tiny difference between the slope parameters' prior and posterior values. To fix this issue, I use the last estimated slope values from Table 5.2 as the prior values in the DCM analysis and repeat this process (use posterior values as priors) until the gap between them is minimal. The values I achieve from this process are 2.74 and 2.18 for the excitatory and inhibitory slope, respectively. Then, I should move the f-I curve of the spiking network model in the right direction by comparing the parameters with these values implied by DCM analysis. For this purpose, I need to modify the neuron parameters to move the f-I curves, as I explain in the following.

In the next step, I try to improve the matching between predicted and simulated time series by changing the spiking neurons' f-I curves based on the DCM results via a change of the neuron parameters. According to the literature, the slopes of the f-I curves are mainly determined by the neuron parameters b, τ_w, and V_r in the simpAdEX model [150].

Here I vary only the parameters for the L2/3 pyramidal cells, as those are the main players for the interaction between layers. However, I also tried to vary inhibitory neuron parameters, but I couldn't find any substantial effect on the f-I curves' slopes.

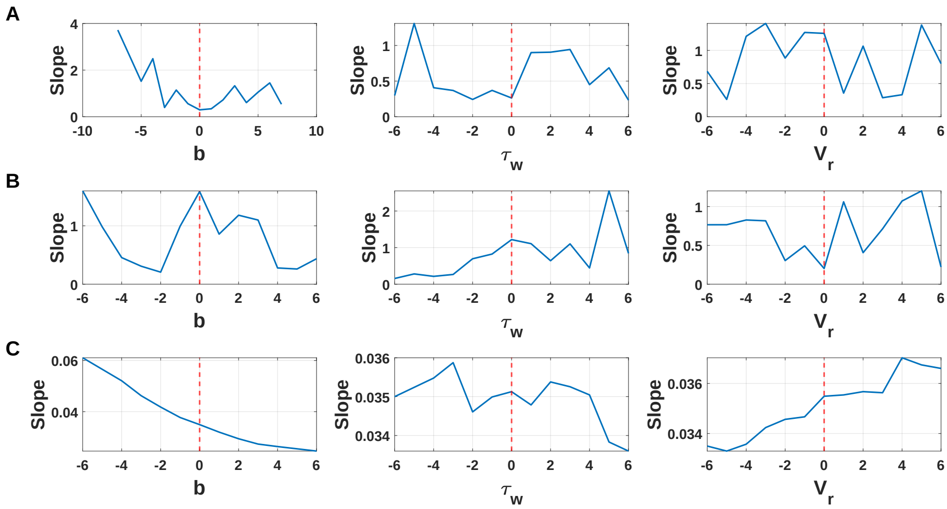


Figure 5.6.: The f-I curves' slopes vs. varying three neuron parameters for A) L2/3 excitatory neurons, B) L2/3 inhibitory neurons, and C) L2/3 excitatory neurons when they have no synaptic connections.

In Fig. 5.6, I have shown the effects of the changed parameters on the f-I curve for excitatory and inhibitory neurons. Furthermore, I have tested the increasing and decreasing of the L2/3 excitatory neurons' parameters when there are no synaptic connections to see better how the changed parameters affect the f-I curve. As can be seen, the results for manipulating the parameter b for excitatory neurons (Fig. 5.6A, C) look reasonable - e.g., in Fig. 5.6C for parameter b, the change in the f-I curves' slopes can be fitted by a linear function, the gradient of which changes about two-fold with b, which is a strong modulation. Thus, I use only this parameter for modifying the f-I curves.

Fig. 5.7 shows the best match between the predicted and simulated time series, which is achieved with the interplay between the f-I curve and the connectivity and varying the parameter b in L2/3 PC neurons for all three stripes. As shown compared to Fig. 5.4, here, the simulated data's amplitude perfectly matches the predicted data, and they are highly correlated for all three regions. The final posterior parameters are reported in Table 5.2 (implied by DCM), and I increased the parameter b by five units (b+5) for all three stripes and the L2/3 excitatory neurons. The model is now ready to predict task performance of social cognition and to be validated on a different data set.

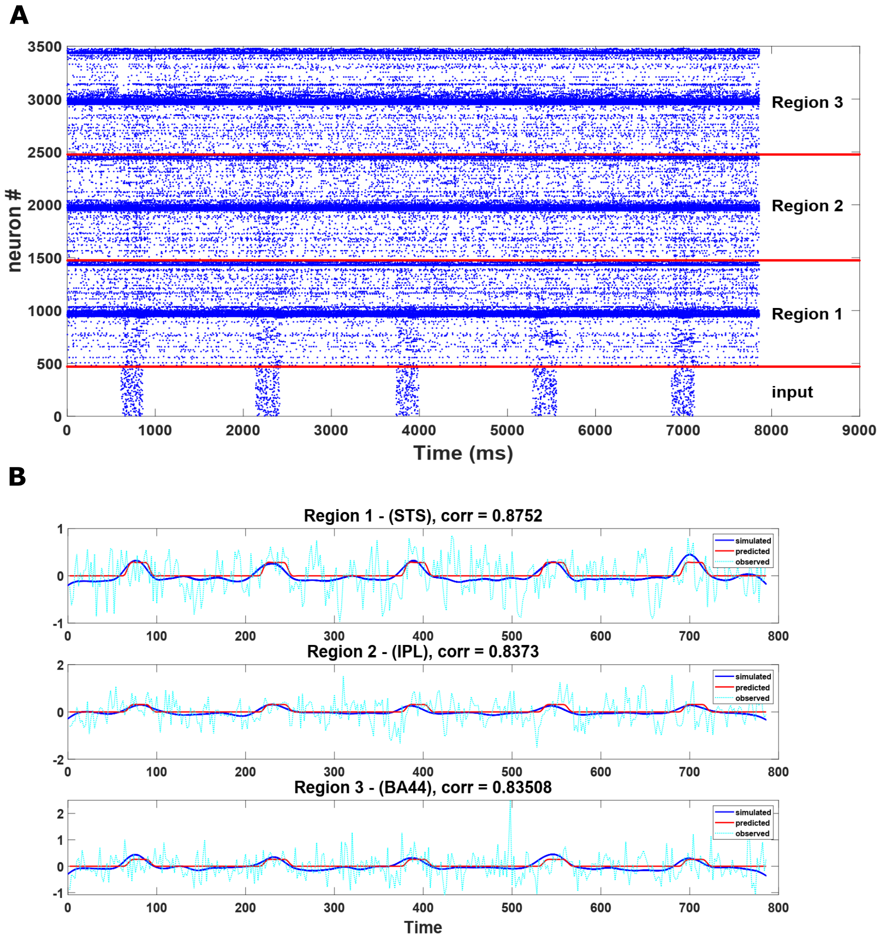


Figure 5.7.: The same explanation as Fig. 5.4. Here, we can see a better match and correlation between the predicted and simulated time series after applying the above-explained process.

6. General discussion of possible extensions & Outlook

In the present dissertation a replacement to the neuronal equation in the DCM for fMRI was developed and a global network for the human MNS during social cognition was proposed. Then, I used these modification and findings to simulate the human MNS with the spiking network model.

The model is now ready to predict task performance of social cognition and to be validated on a different data set. In particular, a second series of experiments aimed to unravel the role of motivational factors on emotion representations in the human MNS. In particular, my colleagues found experimentally that tasks involving a higher level of motivation resulted in a stronger BOLD response in the fMRI recordings. Under the hypothesis that the level of motivation is encoded in the level of dopamine, the next steps will consist of simulations involving dopaminergic modulation of the spiking network model including the effects of dopamine (DA) receptors type 1 and 2 (D1 and D2) [94] on the human MNS. Furthermore, for this project, a set of genotyping for the same group of subjects has been performed with respect to the dopaminergic and oxytocinergic neurotransmitter systems. Here I present my preliminary results on these measurements as a source for future research.

6.1. Dopaminergic modulation of the spiking network model

The activity of the mirror neurons is significantly influenced by motivational and intentional factors. Such psychological factors were associated physiologically with a tonic elevation of the dopamine level [155] as well as an increased power in the gamma spectrum of cortical rhythms [156]. We will, therefore, simulate these changes in the theoretical methods which provides a much deeper understanding of the mechanisms underlying the observed phenomenon. The modeling allows insights into the temporal dynamics of the mirror neurons and their modulation, which would not be accessible by non-invasive measurement methods.

There are (at least) two DA receptors[1], D1 and D2, with somewhat oppositional effects on synapse and neuron parameters. The synaptic changes of the two receptors are taken from the literature [157, 158], and the neuronal changes are taken from our own experiments (these changes will be published in future studies). As we do not know to which extend D1 and D2 receptors are being activated by a tonic change in DA levels, I implement both sets of changes in a gradual manner,

---

[1] It is known that there are more DA receptors, but D1 and D2 are the most abundant ones.

e.g., define a parameter for both receptors that ranges from zero to one, where one represents the full changes and zero no change at all (unmodified parameters). Then, I first explore the three extreme cases - full D1 receptor activation only, full D2 activation only, and full activation of both receptors - and compare the results with the baseline simulation (e.g., none of the two receptors active). Here only the results of the full activation of both receptors are presented.

In the first step, I want to see whether dopamine affects the temporal dynamics of the activity. The hypothesis is that DA modulation (D1, for the most part) stabilizes activity in response to a brief stimulus [93], allowing higher and more prolonged responses, which are more likely to be detected by fMRI. To check the effects of DA modulation on the spiking network model, I run the simulation for 12000 ms and compute the firing rate time series for layer 2/3 and Layer 5 in response to spiking inputs in the middle of the simulation.

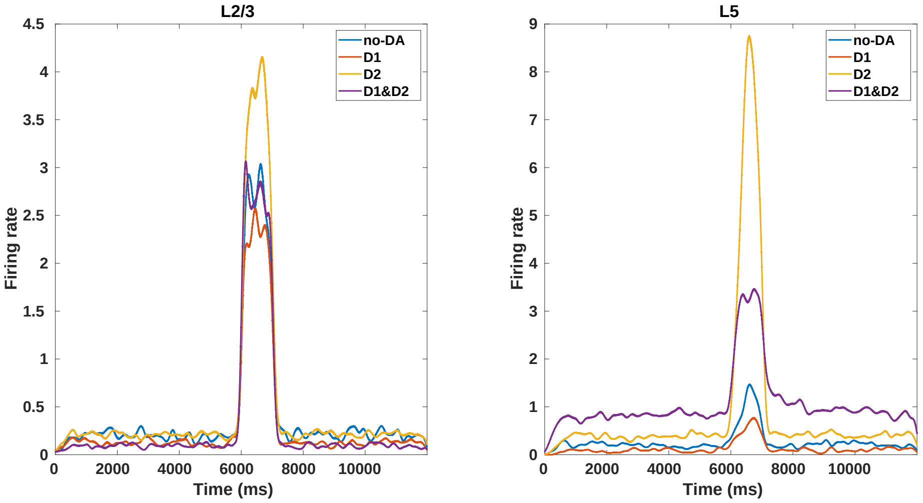


Figure 6.1.: The firing rate time series of the spiking network model, layer 2/3 and layer 5, in response to spiking inputs at the interval between 6000 and 7000 ms, when there is no DA modulation and D1, D2, D1&D2 receptors are activated.

As shown in Fig. 6.1, there is not much of a difference before and after the stimulus, and DA does not induce persistent activity per se. In order to have a persistent activity, I need to implement cell assemblies into the simulation, subsets of pyramidal cells in L2/3 that are wired more strongly than the others. Based on several previous studies (most notably [93]), DA, particularly D1, would be expected to enhance the persistent activity, e.g., make it more robust against disturbance from the outside, and therefore more long-lived. For this purpose, I use a version of the spiking network model, which produces persistent activity with some degree of noisy background activity [146] and apply the DA modulation on this network.

Figure 6.2 shows the raster plots of the network with one cell assembly in L2/3 when there is no DA modulation, and D1, D2, D1&D2 receptors are activated. d_{PA} is a measure, which is defined as the difference between the normalized activity in the cell assembly after and before the stimulus, ranging between zero and one (it is zero when the difference is negative). Conventionally when it is more than 0.3, we

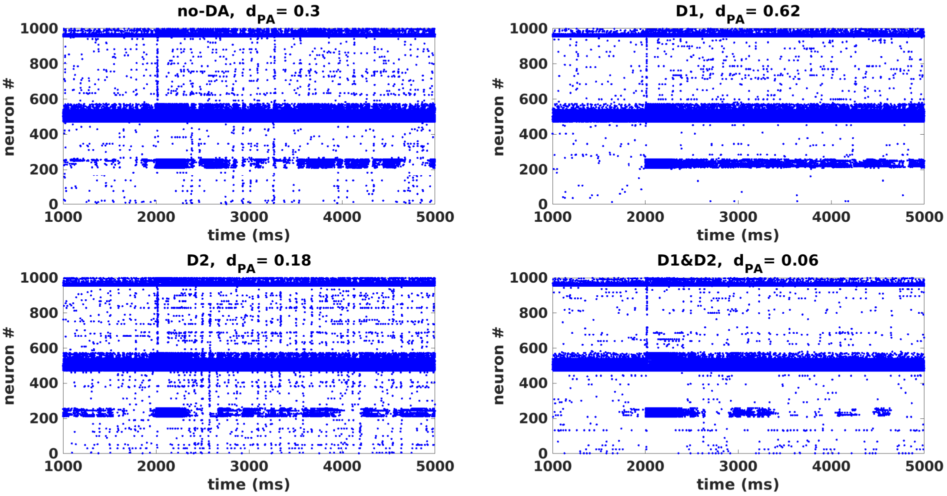


Figure 6.2.: Raster plots of the spike times in the network of 1000 neurons in response to the external input activated at the time 2000 ms when there is no DA modulation and when D1, D2, D1&D2 receptors are activated. One cell assembly is implemented in Layer 2/3, and a single input neuron is randomly connected to L2/3 cells. d_{PA} shows the persistent activity measure [146].

conclude a persistent activity in the network (see [146] for more details). As shown in Fig. 6.2, the rates are consistently higher than the period before the stimulus. The persistent activity is visible in the raster plot after the stimulus within the cell assembly (the 80 middle neurons of L2/3) for the panels on the right - the ones on the left show spontaneous persistent activity, independent of the stimulus. The situation without DA modulation shows spontaneous persistent activity. When the D1 receptor is activated, the persistent activity is enhanced ($d_{PA} > 0.3$), and when the D2 receptor is activated, the persistent activity is decreased ($d_{PA} < 0.3$). When both receptors are activated, the cell assembly activity dies out since these two receptors have oppositional effects. However, this is not necessarily always the case; If both receptors exactly cancel each other, persistent activity would work well as the situation without DA modulation.

I also test the effects of DA on the network with two cell assemblies, one activated early and the other one later. If activity in the first assembly is stable, it should not be affected by the second (and the second may die out because of overall increased inhibition). The hypothesis would be that this stability is enhanced by DA. Figure 6.3 shows this process with two cell assembly. Without DA modulation, the activity in the first assembly completely dies out when the second stimulus activates the second cell assembly. For the D1 receptor, as expected, the activity in the assembly is not affected by the second one, and it shows a persistent activity with a very high d_{PA} number. On the other hand, for the D2 receptor, the activity of the first assembly is affected by the second one, and it is more persistent than the first assembly. When both D1 and D2 receptors are activated, the neuronal spiking dies out for both assemblies, and in this case, two receptors neutralize each other's effects.

These results are examples, and many repetitions are necessary to confirm these

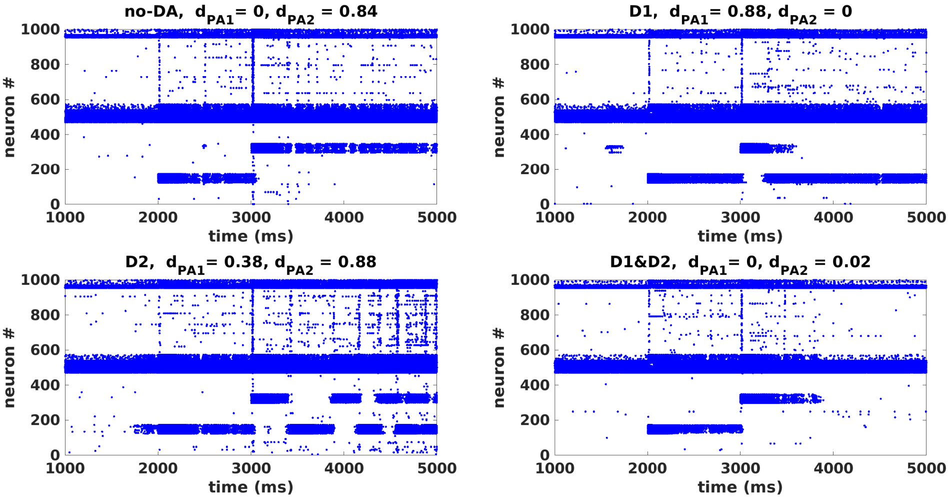


Figure 6.3.: Raster plots of the spike times in the network of 1000 neurons in response to the external input when there is no DA modulation and D1, D2, D1&D2 receptors are activated. Two cell assemblies are implemented in Layer 2/3. A single input neuron is randomly connected to the first assembly at the time 2000 ms and another one at the time 3000 ms to the second assembly. d_{PA1} and d_{PA2} show the persistent activity measures of the first and second assembly.

effects statistically. In the future, we add the changes from DCM into the network and investigate this process for more samples.

6.2. PEB analysis on subjects according to their genotypes

Dopamine is an essential candidate for investigating the neurotransmitter systems involved in the functionality of the MNS and plays a crucial role in the functioning of the motor system [159]. In this way, to investigate the relevance of DA for social cognition, a set of genotyping concerning the dopaminergic for the same group of subjects (chapter 4) has been performed. The single-nucleotide polymorphism (SNP) rs4680 is the best-studied genetic variant for the dopamine system in the COMT gene [160]. The COMT gene codes for the COMT enzyme, which degrades dopamine (DA) in the prefrontal cortex [161]. For this SNP, there are three different groups/genotypes in which our subjects are categorized according to these genotypes. The genotype GG denotes the subjects with low DA level, AA high DA level, and GA intermediate level of DA.

Parametric Empirical Bayes (PEB) is a hierarchical Bayesian model that uses both non-linear (at first level) and linear (at the second level) analyses [162, 163]. The main advantage of using PEB analysis is that it identifies the commonalities and differences across groups in the effective connectivity and takes into account the estimated covariance between parameters.

In this approach, the fully connected DCM (with hypothesized constraints) is estimated for each subject, and then Bayesian model reduction (BMR) and BMA are performed. BMR is a particularly efficient form of BMS that invert the reduced models using only the posterior densities of a full model. The important aspect of BMR is that models differ only in their priors, and the posterior of a reduced model can be derived from the posterior of the full model.

There are two ways of using BMR. We can pre-define the reduced models with strong hypotheses about between-subject effects on connectivity and infer the posterior densities of the reduced models with the posterior density of the full model or perform an automatic search over the reduced models [162, 164, 165]. For the automatic search, BMR performs a greedy search, automatically compares the full model with the reduced models by iteratively pruning out the connections which have the least evidence and do not contribute to the model evidence. A Bayesian model average is then calculated over the 256 models from the final iteration of the greedy search [163].

Here I have the same subjects as the chapter 4 and the same regions of interest for social-cognition tasks are considered. For now, I only present the preliminary result of the imitation task. To perform the PEB analysis, I define between-subjects differences as a design matrix X, which rows are the number of subjects and columns are the regressors for defining the group mean (first column) and differences (second column and more). In general, for n groups, there are $n-1$ group differences, which leads to a maximum of $n-1$ columns in the design matrix (as well as the overall mean). First, I should define a full model with all parameters in the matrices (A, B, and C). In our full model, as we have three regions (STS, IPL, and BA44), matrix A has 9 parameters, B has 6 (no modulation of self-connection), and matrix C one parameter (with the hypothesis input to STS), which are estimated for all subjects. After estimating the full model for all subjects, PEB runs an automatic search on all possible models and shows the reduced model by eliminating the parameters that have no role in the free energy, i.e., the presence or absence of these parameters has no effect on free energy. It will prune the parameters from the PEB model that do not contribute to the model evidence.

Here I use a linear effect of dopamine for modeling three groups in the SNP rs4680. A linear effect of groups - i.e., group 1 $<$ group 2 $<$ group 3 (low-level, intermediate, and high level of dopamine)- can be modeled using two regressors: 1) the overall mean and 2) the difference between groups 1 and 3 (this is shown in the Fig. 6.4).

The automatic search is performed on matrix A and Matrix B separately. The numbers below are the number of parameters (shown in Fig. 6.4), e.g., the parameter number 4 in the A matrix is the connection from IPL to STS (the direction of connections are from regions on the row to regions on the column):

$$\mathrm{A} = \begin{array}{c} \\ \scriptstyle STS \\ \scriptstyle IPL \\ \scriptstyle BA44 \end{array} \begin{array}{c} \begin{array}{ccc} \scriptstyle STS & \scriptstyle IPL & \scriptstyle BA44 \end{array} \\ \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} \end{array}$$

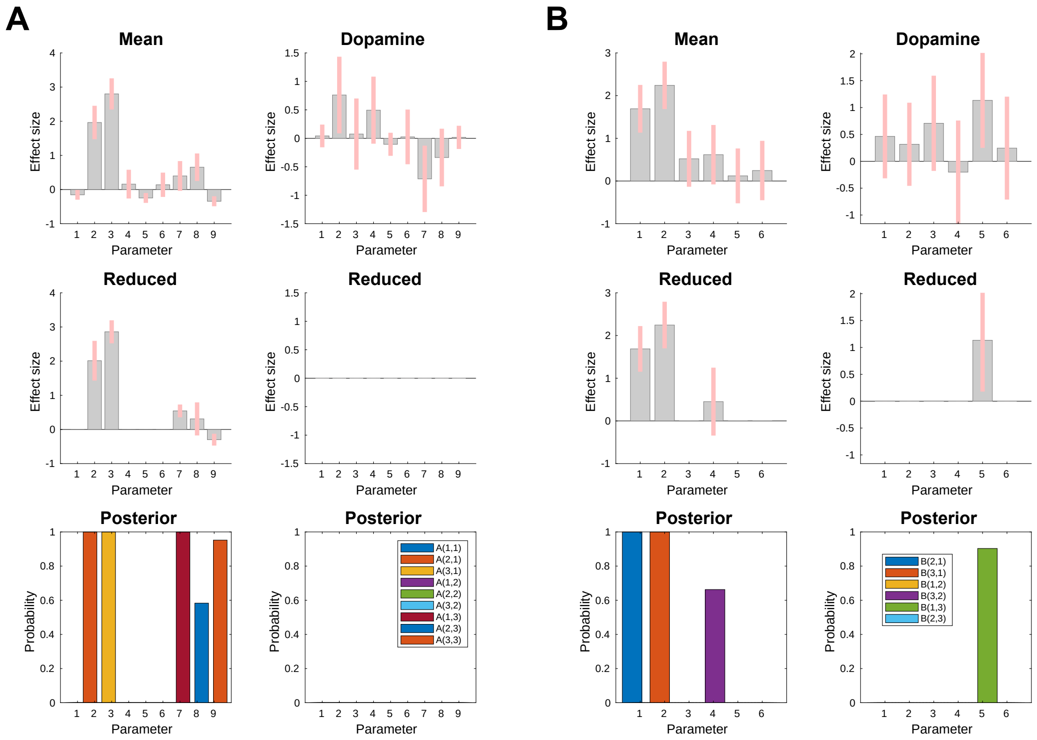


Figure 6.4.: Posterior parameter estimates of the matrices A (9 parameters) and B (6 parameters) based on the BMA performed on the final 256 models of an automatic parameter search for the imitation task. Each grey bar corresponds to one DCM connection (numbers shown in the above matrices), and the error bars are 90% credible intervals, derived from the posterior variance of each parameter. The first columns for each matrices A and B are the commonalities across subjects (group average). The second column shows the PEB parameters relating to the group difference (Dopamine level). The first row shows the parameters from the BMA before the search, and the second row shows the parameters after the model search (reduced model). The bottom row shows the posterior probability for each parameter in the reduced model.

$$\mathrm{B} = \begin{array}{c} \\ \mathit{STS} \\ \mathit{IPL} \\ \mathit{BA44} \end{array} \begin{array}{c} \begin{array}{ccc} \mathit{STS} & \mathit{IPL} & \mathit{BA44} \end{array} \\ \begin{pmatrix} - & 3 & 5 \\ 1 & - & 6 \\ 2 & 4 & - \end{pmatrix} \end{array}$$

Figure 6.4 shows the automatic search on matrix A (the intrinsic connections, 9 parameters) and B (modulatory parameters, 6 parameters) separately for SNP rs4680, imitation task. In this figure, the first column (for each A and B) shows the parameters in common for all groups. The second row shows the surviving parameters including for matrix A the parameter 2 (STS→IPL, P=1), 3 (STS→BA44, P=1), 7 (BA44→STS, P=1), 8 (BA44→IPL, P=0.6), 9 (BA44→BA44, P=0.95) and matrix B , parameter 1 (STS→IPL, P=1), 2 (STS→BA44, P=1), 4 (IPL→BA44, p=0.6). Clearly, many parameters have been pruned away because they did not contribute to the model evidence (free energy). The probability larger than 75% is considered a positive parameter and larger than 95 percent as a strong parameter (Table 2.1).

To compare these connections with those from chapter 4, the feed-forward information flows from STS to IPL and BA44 are also present. These connections are modulated with the contextual input. In contrast to what we found in chapter 4 for the imitation task, the connection between BA44 and IPL is weak (P=0.6) in this automatic search.

Regarding the group differences (second column in Fig. 6.4, here dopamine level), it is clear from the middle plot (second row) no effect of dopamine has survived in the intrinsic parameters (matrix A). For modulatory parameters, only parameter 5 (BA44→STS) with a probability of about 90% is survived. This result shows the dopamine would increase only this modulatory input with a strength of about 1.1 with a probability of 0.9.

6.3. Concluding remarks

In my Ph.D. thesis, I attempt to gain a better understanding of the human MNS by finding the effective connectivity and information flow between the activated regions during social cognition tasks. I use this network structure revealed by DCM for a mathematical network model that can reproduce, for the first time, the temporal dynamics and physiological properties of the MNS. Furthermore, an improved methodology was introduced in the DCM framework for fMRI data.

In the first step, I developed a modification of the neuronal equation in the DCM framework and substituted the bilinear form with a Wilson-Cowan-based equation, which has a sigmoidal form and allows a more direct comparison of the results from DCM with local neural activity. The main reason for this modification was to use the results from the DCM to constrain parameters of the detailed network model of a single brain region in chapter 5. Our results validated the superiority of the Wilson-Cowan-based models with the novel, established, and simulated data sets. Furthermore, this modification can be helpful to the large community of scientists working on DCM, as this new formulation is both more powerful and more realistic. It infers the sigmoid transfer function as an averaged f-I curve of brain regions and can adopt generative models for fMRI time series to be informed by physiological principles.

In the next step, I performed the optimized version of DCM to find the effective connectivity between the activated brain regions in the MNS during imitation, empathy, and theory of mind tasks. The results showed the feedforward connections from STS to IPL and BA44 for all social-cognitive processes and additional mutual connections between STS and BA44, as well as BA44 and IPL for the imitation task only. These results suggest an inverse internal model in which BA44 and IPL receive sensory information from the STS, contradicting standard theories of MNS function involving hand motion. On the other hand, the pattern during imitation suggests a closed loop with an exchange of sensory and motor information, probably due to the motion component of the task, which sets it apart from the other tasks. These results can significantly contribute to understanding the information flow of human MNS during social cognition.

In the final step, I used the network structure and the estimated posterior parameters achieved from chapters 3 and 4 to model the MNS mathematically. The neuron and synapse properties of this model were first obtained from animal models [19], and I use it here for transferring from macro to the micro-level by matching the predicted data from DCM analysis with the simulated time-series from the spiking network model. For future research, this model can be enrolled to investigate genetic variations and TMS manipulation mechanisms. In perspective, the high temporal accuracy of this model will also allow matching with EEG data to better understand, for example, the dynamics of mu rhythm suppression in the MNS.

To sum up, in the present thesis, we developed a two-step model approach, which allows conclusions to be drawn from the experimental data on the anatomy and physiology of the mirror neuron system. Together with this two-step approach, including identifying the network structure by an improved DCM methodology and fitting the local physiological parameters by the DCM data, a widely applicable methodology for more accurate modeling of human brain functions beyond the understanding of the MNS has also been created.

In the end, I want to briefly mention our future research to have a better understanding of the properties of the human MNS. My colleagues have also recorded the EEG data (together with fMRI data) during the social cognition tasks within the same subject population to combine the temporal and spatial resolution of both methods and find the intersection of brain activation with EEG-fMRI combination. In the EEG, suppression of the Mu rhythm is an indicator of movement observation and execution, and then a signature of the activity of the human MNS [92]. We also perform the DCM analysis for the EEG data, and the theoretical modeling provides the development and modulation of the Mu rhythm in the EEG and the temporal dynamics of the activity.

Furthermore, within the same subject population, the modulation of the mirror neuronal activity was investigated by a virtual lesion (induced by inhibitory transcranial magnetic stimulation (TMS) over the right BA44). Here, we can find the effective connectivity between the activated regions in the same way as the fMRI data. The TMS inhibition can also be directly implemented in the spiking network model and allows a mechanistic understanding of its effect on mirror neuron activity.

In addition, we suspect a modulation of the activity and connectivity of the MNS by the genetic polymorphisms. As mentioned before, all subjects are also typed for the genes COMT and OXTR. I presented the preliminary results concerning the effective connectivity for the SNP rs4680 in the COMT gene. Furthermore, mainly two other SNPs of interest are investigated, namely rs1344706, which has been identified to be associated with schizophrenia [166], and rs1800497, which is known for its influence on the dopamine D2 receptor (DRD2) gene. We use the PEB analysis to study the effect of these neurotransmitters on the strength of the connections between the activated regions in the social cognition tasks. The theoretical modeling presented in chapter 5 also provides for the first time the possibility to understand the physiological mechanisms of the mirror neuron activity, in particular the modeling of the dopaminergic and oxytocinergic transmitter system.

In this way, the computational modeling of a multimodal recording of indicators of mirror neuron activity (behavioral measures, fMRI and EEG) during core processes of social cognition (imitation of emotional facial expressions, empathy, emotion recognition, and ToM) and also modeling of the factors influencing mirror neuron activity (deactivation of brain areas with TMS as well as genotyping concerning the dopaminergic and oxytocinergic neurotransmitter system) will help for the first time to simulate the indicators of the non-invasive measurement methods and compare them with the measured values. Therefore the statements about the physiology of the cell assemblies are possible since the parameters of the model are directly related to biophysical properties, e.g., cellular and synaptic conductivities and resting potentials.

Appendices

A. Maximum likelihood estimation

In maximum likelihood estimation (MLE), we estimate the parameters that maximize a likelihood function. It is defined as below:

$$\hat{\theta}_{\mathrm{MLE}}(y) = \arg\max_{\theta} p(y \mid \theta), \tag{A.1}$$

where $p(y|\theta)$ is the likelihood function, defined as the probability of the observed data y given parameters θ. For numerical reasons, it is often more convenient to use the log-likelihood. This is equivalent because the logarithm is a strictly monotonic function, i.e., maximizing the log-likelihood also maximizes the likelihood function.

B. Fisher scoring algorithm

Fisher scoring algorithms is a kind of Newton's method to solve maximum likelihood equations numerically. It is defined as below:

$$\theta_{m+1} = \theta_m + \mathcal{I}^{-1}(\theta_m)V(\theta_m), \tag{B.1}$$

where $V(\theta)$ is the score function defined as the gradient of the log-likelihood function with respect to the parameter vector θ. $\mathcal{I}(\theta)$ is the expected value of the negative of the second derivative of the log-likelihood called the fisher information.

In this way, the scoring algorithm can be written as below which is the form used in the chapter 2:

$$\theta_{m+1} = \theta_m - \left\langle \frac{\partial^2 I}{\partial \theta^2} \right\rangle^{-1} \frac{\partial I}{\partial \theta}, \tag{B.2}$$

where I is the log-likelihood function.

Bibliography

[1] R. Hari and M. V. Kujala, “Brain basis of human social interaction: from concepts to brain imaging”, Physiological reviews **89**, 453 (2009).

[2] G. Rizzolatti and C. Sinigaglia, “The mirror mechanism: a basic principle of brain function”, Nature Reviews Neuroscience **17**, 757 (2016).

[3] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, “Understanding motor events: a neurophysiological study”, Experimental brain research **91**, 176 (1992).

[4] V. Gallese and A. Goldman, “Mirror neurons and the simulation theory of mind-reading”, Trends in cognitive sciences **2**, 493 (1998).

[5] V. Gallese, “Before and below ‘theory of mind’: embodied simulation and the neural correlates of social cognition”, Philosophical Transactions of the Royal Society of London B: Biological Sciences **362**, 659 (2007).

[6] C. Keysers and V. Gazzola, “Towards a unifying neural theory of social cognition”, Progress in brain research **156**, 379 (2006).

[7] E. Oztop, M. Kawato, and M. Arbib, “Mirror neurons and imitation: A computationally guided review”, Neural Networks **19**, 254 (2006).

[8] S. Thill, D. Caligiore, A. M. Borghi, T. Ziemke, and G. Baldassarre, “Theories and computational models of affordance and mirror systems: an integrative review”, Neuroscience & Biobehavioral Reviews **37**, 491 (2013).

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

[10] M. Dapretto, M. S. Davies, J. H. Pfeifer, A. A. Scott, M. Sigman, S. Y. Bookheimer, and M. Iacoboni, “Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders”, Nature neuroscience **9**, 28 (2006).

[11] D. Mier, S. Lis, C. Esslinger, C. Sauer, M. Hagenhoff, J. Ulferts, B. Gallhofer, and P. Kirsch, “Neuronal correlates of social cognition in borderline personality disorder”, Social Cognitive and Affective Neuroscience **8**, 531 (2012).

[12] D. Mier, L. Haddad, K. Diers, H. Dressing, A. Meyer-Lindenberg, and P. Kirsch, “Reduced embodied simulation in psychopathy”, The World Journal of Biological Psychiatry **15**, 479 (2014).

[13] R. Mukamel, A. D. Ekstrom, J. Kaplan, M. Iacoboni, and I. Fried, "Single-neuron responses in humans during execution and observation of actions", Current biology **20**, 750 (2010).

[14] S. N. Schmidt, J. Hass, P. Kirsch, and D. Mier, "The human mirror neuron system–A common neural basis for social cognition?", Psychophysiology , e13781 (2021).

[15] S. N. L. Schmidt, "Neural mechanisms of social cognition–the mirror neuron system and beyond", PhD thesis, University of Heidelberg, 2020.

[16] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling", Neuroimage **19**, 1273 (2003).

[17] K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston, "Bayesian model selection for group studies", Neuroimage **46**, 1004 (2009).

[18] L. Rigoux, K. E. Stephan, K. J. Friston, and J. Daunizeau, "Bayesian model selection for group studies - revisited", Neuroimage **84**, 971 (2014).

[19] J. Hass, L. Hertäg, and D. Durstewitz, "A detailed data-driven network model of prefrontal cortex reproduces key features of in vivo activity", PLoS computational biology **12**, e1004930 (2016).

[20] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons", Biophysical journal **12**, 1 (1972).

[21] H. R. Wilson and J. D. Cowan, "A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue", Kybernetik **13**, 55 (1973).

[22] C. Büchel and K. J. Friston, "Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI.", Cerebral cortex (New York, NY: 1991) **7**, 768 (1997).

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

[24] W. D. Penny, K. E. Stephan, J. Daunizeau, M. J. Rosa, K. J. Friston, T. M. Schofield, and A. P. Leff, "Comparing families of dynamic causal models", PLoS computational biology **6**, e1000709 (2010).

[25] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[26] N. K. Logothetis and B. A. Wandell, "Interpreting the BOLD signal", Annu. Rev. Physiol. **66**, 735 (2004).

[27] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price, "Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics", NeuroImage **12**, 466 (2000).

[28] M. A. Lindquist et al., "The statistical analysis of fMRI data", Statistical science **23**, 439 (2008).

[29] K. J. Friston, “Functional and effective connectivity: a review”, Brain connectivity **1**, 13 (2011).

[30] A. Razi and K. J. Friston, “The connected brain: causality, models, and intrinsic dynamics”, IEEE Signal Processing Magazine **33**, 14 (2016).

[31] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston, “Dynamic causal modeling of evoked responses in EEG and MEG”, NeuroImage **30**, 1255 (2006).

[32] S. J. Kiebel, M. I. Garrido, R. J. Moran, and K. J. Friston, “Dynamic causal modelling for EEG and MEG”, Cognitive neurodynamics **2**, 121 (2008).

[33] R. J. Moran, D. A. Pinotsis, and K. J. Friston, “Neural masses and fields in dynamic causal modeling”, Frontiers in computational neuroscience **7**, 57 (2013).

[34] K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston, “Comparing hemodynamic models with DCM”, Neuroimage **38**, 387 (2007).

[35] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, “Classical and Bayesian inference in neuroimaging: theory”, NeuroImage **16**, 465 (2002).

[36] K. J. Friston, “Bayesian estimation of dynamical systems: an application to fMRI”, NeuroImage **16**, 513 (2002).

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society: Series B (Methodological) **39**, 1 (1977).

[38] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, “Variational free energy and the Laplace approximation”, Neuroimage **34**, 220 (2007).

[39] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*, Elsevier, 2011.

[40] R. M. Neal and G. E. Hinton, “A view of the EM algorithm that justifies incremental, sparse, and other variants”, in *Learning in graphical models*, pages 355–368, Springer, 1998.

[41] K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain”, Journal of physiology-Paris **100**, 70 (2006).

[42] K. Friston, “The free-energy principle: a rough guide to the brain?”, Trends in cognitive sciences **13**, 293 (2009).

[43] M. J. Beal, *Variational algorithms for approximate Bayesian inference*, PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[44] T. Tanaka, “A theory of mean field approximation”, in *Advances in Neural Information Processing Systems*, pages 351–360, 1999.

[45] J. Daunizeau, K. J. Friston, and S. J. Kiebel, “Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models”, Physica D: nonlinear phenomena **238**, 2089 (2009).

[46] K. J. Friston, B. Li, J. Daunizeau, and K. E. Stephan, “Network discovery with DCM”, Neuroimage **56**, 1202 (2011).

[47] G. Deco, V. K. Jirsa, P. A. Robinson, M. Breakspear, and K. Friston, “The dynamic brain: from spiking neurons to neural masses and cortical fields”, PLoS Comput Biol **4**, e1000092 (2008).

[48] A. C. Marreiros, K. E. Stephan, and K. J. Friston, “Dynamic causal modeling”, Scholarpedia **5**, 9568 (2010).

[49] R. E. Kass and A. E. Raftery, “Bayes factors”, Journal of the american statistical association **90**, 773 (1995).

[50] E. Raftery Adrian, “Bayesian model selection in social research”, Sociological methodology **25**, 111 (1995).

[51] W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston, “Comparing dynamic causal models”, Neuroimage **22**, 1157 (2004).

[52] K. E. Stephan, J. C. Marshall, W. D. Penny, K. J. Friston, and G. R. Fink, “Interhemispheric integration of visual processing during task-driven lateralization”, Journal of Neuroscience **27**, 3512 (2007).

[53] S. Sadeghi, D. Mier, M. F. Gerchen, S. N. Schmidt, and J. Hass, “Dynamic Causal Modeling for fMRI With Wilson-Cowan-Based Neuronal Equations.”, Frontiers in Neuroscience **14**, 593867 (2020).

[54] K. J. Friston, J. Kahan, B. Biswal, and A. Razi, “A DCM for resting state fMRI”, Neuroimage **94**, 396 (2014).

[55] A. C. Marreiros, S. J. Kiebel, and K. J. Friston, “Dynamic causal modelling for fMRI: a two-state model”, Neuroimage **39**, 269 (2008).

[56] K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. den Ouden, M. Breakspear, and K. J. Friston, “Nonlinear dynamic causal models for fMRI”, Neuroimage **42**, 649 (2008).

[57] O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis, “Identifying neural drivers with functional MRI: an electrophysiological validation”, PLoS biology **6**, e315 (2008).

[58] K. J. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, and P. Zeidman, "Dynamic causal modelling revisited", Neuroimage **199**, 730 (2019).

[59] A. Jafarian, V. Litvak, H. Cagnan, K. J. Friston, and P. Zeidman, "Comparing dynamic causal models of neurovascular coupling with fMRI and EEG/MEG", NeuroImage **216**, 116734 (2020).

[60] H. Wei, A. Jafarian, P. Zeidman, V. Litvak, A. Razi, D. Hu, and K. J. Friston, "Bayesian fusion and multimodal DCM for EEG and fMRI", NeuroImage **211**, 116595 (2020).

[61] E. Marder and A. A. Prinz, "Modeling stability in neuron and network function: the role of activity in homeostasis", Bioessays **24**, 1145 (2002).

[62] J. Ashburner et al., "SPM12 manual", Wellcome Trust Centre for Neuroimaging, London, UK **2464** (2014).

[63] J. Daunizeau, V. Adam, and L. Rigoux, "VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data", PLoS Computational Biology **10**, e1003441 (2014).

[64] L. Harrison, W. D. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series", Neuroimage **19**, 1477 (2003).

[65] W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston, "Modelling functional integration: a comparison of structural equation and dynamic causal models", Neuroimage **23**, S264 (2004).

[66] D. Mier, S. Lis, K. Neuthe, C. Sauer, C. Esslinger, B. Gallhofer, and P. Kirsch, "The involvement of emotion recognition in affective theory of mind", Psychophysiology **47**, 1028 (2010).

[67] D. Mier, C. Sauer, S. Lis, C. Esslinger, J. Wilhelm, B. Gallhofer, and P. Kirsch, "Neuronal correlates of affective theory of mind in schizophrenia out-patients: evidence for a baseline deficit", Psychological Medicine **40**, 1607 (2010).

[68] P. Zeidman, A. Jafarian, N. Corbin, M. L. Seghier, A. Razi, C. J. Price, and K. J. Friston, "A guide to group effective connectivity analysis, part 1: First level analysis with DCM for fMRI", Neuroimage **200**, 174 (2019).

[69] M. Iacoboni, L. M. Koski, M. Brass, H. Bekkering, R. P. Woods, M.-C. Dubeau, J. C. Mazziotta, and G. Rizzolatti, "Reafferent copies of imitated actions in the right superior temporal cortex", Proceedings of the national academy of sciences **98**, 13995 (2001).

[70] G. Lohmann, K. Erfurth, K. Müller, and R. Turner, "Critical comments on dynamic causal modelling", Neuroimage **59**, 2322 (2012).

[71] S. M. Hadi, “Estimating effective connectivity within brain emotional circuitry using dynamic causal modeling and fMRI”, PhD thesis, Oakland University, 2014.

[72] J. Daunizeau, O. David, and K. E. Stephan, “Dynamic causal modelling: a critical review of the biophysical and statistical foundations”, Neuroimage **58**, 312 (2011).

[73] K. E. Stephan and A. Roebroeck, “A short history of causal modeling of fMRI data”, Neuroimage **62**, 856 (2012).

[74] B. Li, J. Daunizeau, K. E. Stephan, W. Penny, D. Hu, and K. Friston, “Generalised filtering and stochastic DCM for fMRI”, neuroimage **58**, 442 (2011).

[75] A. Razi, J. Kahan, G. Rees, and K. J. Friston, “Construct validation of a DCM for resting state fMRI”, Neuroimage **106**, 1 (2015).

[76] S. Frässle, E. I. Lomakina, A. Razi, K. J. Friston, J. M. Buhmann, and K. E. Stephan, “Regression DCM for fMRI”, Neuroimage **155**, 406 (2017).

[77] S. Frässle, E. I. Lomakina, L. Kasper, Z. M. Manjaly, A. Leff, K. P. Pruessmann, J. M. Buhmann, and K. E. Stephan, “A generative model of whole-brain effective connectivity”, Neuroimage **179**, 505 (2018).

[78] M. Havlicek, A. Roebroeck, K. Friston, A. Gardumi, D. Ivanov, and K. Uludag, “Physiologically informed dynamic causal modeling of fMRI data”, Neuroimage **122**, 355 (2015).

[79] J. Heinzle, P. J. Koopmans, H. E. den Ouden, S. Raman, and K. E. Stephan, “A hemodynamic model for layered BOLD signals”, Neuroimage **125**, 556 (2016).

[80] A. C. Marreiros, J. Daunizeau, S. J. Kiebel, and K. J. Friston, “Population dynamics: variance and the sigmoid activation function”, Neuroimage **42**, 147 (2008).

[81] A. M. Bastos, V. Litvak, R. Moran, C. A. Bosman, P. Fries, and K. J. Friston, “A DCM study of spectral asymmetries in feedforward and feedback connections between visual areas V1 and V4 in the monkey”, Neuroimage **108**, 460 (2015).

[82] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*, The MIT Press, 2001.

[83] E. Wallace, M. Benayoun, W. Van Drongelen, and J. D. Cowan, “Emergent oscillations in networks of stochastic spiking neurons”, Plos one **6**, e14804 (2011).

[84] G. Rizzolatti and L. Fogassi, “The mirror mechanism: recent findings and perspectives”, Philosophical Transactions of the Royal Society B: Biological Sciences **369**, 20130420 (2014).

[85] L. Carr, M. Iacoboni, M.-C. Dubeau, J. C. Mazziotta, and G. L. Lenzi, "Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas", Proceedings of the national Academy of Sciences **100**, 5497 (2003).

[86] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti, "Grasping the intentions of others with one's own mirror neuron system", PLoS biology **3**, e79 (2005).

[87] S. Sadeghi, D. Mier, and J. Hass, "Detailed spiking network model of the human mirror neuron system", Bernstein Conference 2017, 2017.

[88] D. Caligiore, A. M. Borghi, D. Parisi, and G. Baldassarre, "TRoPICALS: A computational embodied neuroscience model of compatibility effects.", Psychological Review **117**, 1188 (2010).

[89] A. T. Sasaki, T. Kochiyama, M. Sugiura, H. C. Tanabe, and N. Sadato, "Neural networks for action representation: a functional magnetic-resonance imaging and dynamic causal modeling study", Frontiers in human neuroscience **6**, 236 (2012).

[90] J. Triesch, H. Jasso, and G. O. Deák, "Emergence of mirror neurons in a model of gaze following", Adaptive Behavior **15**, 149 (2007).

[91] H. M. Hobson and D. V. Bishop, "Mu suppression–a good measure of the human mirror neuron system?", cortex **82**, 290 (2016).

[92] A. Moore, I. Gorodnitsky, and J. Pineda, "EEG mu component responses to viewing emotional faces", Behavioural brain research **226**, 309 (2012).

[93] D. Durstewitz, J. K. Seamans, and T. J. Sejnowski, "Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex", Journal of neurophysiology **83**, 1733 (2000).

[94] J. Hass and D. Durstewitz, "Models of dopaminergic modulation", Scholarpedia **6**, 4215 (2011).

[95] A. Meyer-Lindenberg, G. Domes, P. Kirsch, and M. Heinrichs, "Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine", Nature Reviews Neuroscience **12**, 524 (2011).

[96] C. D. Frith and U. Frith, "Social cognition in humans", Current Biology **17**, R724 (2007).

[97] J. M. Kilner and R. N. Lemon, "What we know currently about mirror neurons", Current biology **23**, R1057 (2013).

[98] L. M. Oberman, J. A. Pineda, and V. S. Ramachandran, "The human mirror neuron system: a link between action observation and social skills", Social cognitive and affective neuroscience **2**, 62 (2007).

[99] S. Bekkali, G. J. Youssef, P. H. Donaldson, N. Albein-Urios, C. Hyde, and P. G. Enticott, “Is the putative mirror neuron system associated with empathy? A systematic review and meta-analysis”, Neuropsychology Review , 1 (2020).

[100] M. Iacoboni, “Neural mechanisms of imitation”, Current opinion in neurobiology **15**, 632 (2005).

[101] P. Molenberghs, R. Cunnington, and J. B. Mattingley, “Is the mirror neuron system involved in imitation? A short review and meta-analysis”, Neuroscience & biobehavioral reviews **33**, 975 (2009).

[102] R. Cook, G. Bird, C. Catmur, C. Press, and C. Heyes, “Mirror neurons: from origin to function.”, Behavioral and Brain Sciences **37**, 177 (2014).

[103] M. C. Keuken, A. Hardie, B. Dorn, S. Dev, M. Paulus, K. Jonas, W. Van Den Wildenberg, and J. Pineda, “The role of the left inferior frontal gyrus in social perception: an rTMS study”, Brain research **1383**, 196 (2011).

[104] M. Iacoboni, “Imitation, empathy, and mirror neurons”, Annual review of psychology **60**, 653 (2009).

[105] M. Schurz, L. Maliske, and P. Kanske, “Cross-network interactions in social cognition: A review of findings on task related brain activation and connectivity”, cortex **130**, 142 (2020).

[106] G. Rizzolatti and L. Craighero, “The mirror-neuron system”, Annu. Rev. Neurosci. **27**, 169 (2004).

[107] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, “Premotor cortex and the recognition of motor actions”, Cognitive brain research **3**, 131 (1996).

[108] G. Rizzolatti, “The mirror neuron system and its function in humans”, Anatomy and embryology **210**, 419 (2005).

[109] F. Van Overwalle, “Social cognition and the brain: a meta-analysis”, Human brain mapping **30**, 829 (2009).

[110] M. Iacoboni and M. Dapretto, “The mirror neuron system and the consequences of its dysfunction”, Nature Reviews Neuroscience **7**, 942 (2006).

[111] A. F. d. C. Hamilton, “Reflecting on the mirror neuron system in autism: a systematic review of current theories”, Developmental cognitive neuroscience **3**, 91 (2013).

[112] G. Rizzolatti and M. Fabbri-Destro, “Mirror neurons: from discovery to autism”, Experimental brain research **200**, 223 (2010).

[113] M. A. Giese and G. Rizzolatti, “Neural and computational mechanisms of action processing: Interaction between visual and motor representations”, Neuron **88**, 167 (2015).

[114] J. M. Kilner, “More than one pathway to action understanding”, Trends in cognitive sciences **15**, 352 (2011).

[115] K. Nelissen, E. Borra, M. Gerbella, S. Rozzi, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. A. Orban, “Action observation circuits in the macaque monkey cortex”, Journal of Neuroscience **31**, 3743 (2011).

[116] M. Catani, D. K. Jones, and D. H. Ffytche, “Perisylvian language networks of the human brain”, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society **57**, 8 (2005).

[117] J. K. Rilling, M. F. Glasser, T. M. Preuss, X. Ma, T. Zhao, X. Hu, and T. E. Behrens, “The evolution of the arcuate fasciculus revealed with comparative DTI”, Nature neuroscience **11**, 426 (2008).

[118] Y. Yamashita and J. Tani, “Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment”, PLoS Comput Biol **4**, e1000220 (2008).

[119] W. Erlhagen, A. Mukovskiy, and E. Bicho, “A dynamic model for action understanding and goal-directed imitation”, Brain research **1083**, 174 (2006).

[120] F. Fleischer, V. Caggiano, P. Thier, and M. A. Giese, “Physiologically inspired model for the visual recognition of transitive hand actions”, Journal of Neuroscience **33**, 6563 (2013).

[121] D. M. Wolpert, K. Doya, and M. Kawato, “A unifying computational framework for motor control and social interaction”, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **358**, 593 (2003).

[122] J. M. Kilner, K. J. Friston, and C. D. Frith, “Predictive coding: an account of the mirror neuron system”, Cognitive processing **8**, 159 (2007).

[123] K. Friston, J. Mattout, and J. Kilner, “Action understanding and active inference”, Biological cybernetics **104**, 137 (2011).

[124] S. T. Grafton and A. F. d. C. Hamilton, “Evidence for a distributed hierarchy of action representation in the brain”, Human movement science **26**, 590 (2007).

[125] M. Lebreton, S. Kawa, B. F. d’Arc, J. Daunizeau, and M. Pessiglione, “Your goal is mine: unraveling mimetic desires in the human brain”, Journal of Neuroscience **32**, 7146 (2012).

[126] M. Thioux and C. Keysers, “Object visibility alters the relative contribution of ventral visual stream and mirror neuron system to goal anticipation during action observation”, NeuroImage **105**, 380 (2015).

[127] B. A. Urgen and A. P. Saygin, “Predictive processing account of action perception: Evidence from effective connectivity in the action observation network”, Cortex **128**, 132 (2020).

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

[129] P. Ferrari, M. Gerbella, G. Coudé, and S. Rozzi, “Two different mirror neuron networks: the sensorimotor (hand) and limbic (face) pathways”, Neuroscience **358**, 300 (2017).

[130] W. Sato, T. Kochiyama, and S. Uono, “Spatiotemporal neural network dynamics for the processing of dynamic facial expressions”, Scientific reports **5**, 1 (2015).

[131] W. Sato, T. Kochiyama, S. Uono, S. Yoshikawa, and M. Toichi, “Direction of amygdala–neocortex interaction during dynamic facial expression processing”, Cerebral Cortex **27**, 1878 (2017).

[132] M. Arioli, D. Perani, S. Cappa, A. M. Proverbio, A. Zani, A. Falini, and N. Canessa, “Affective and cooperative social interactions modulate effective connectivity within and between the mirror and mentalizing systems”, Human brain mapping **39**, 1412 (2018).

[133] N. E. Barraclough*, D. Xiao*, C. I. Baker, M. W. Oram, and D. I. Perrett, “Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions”, Journal of Cognitive Neuroscience **17**, 377 (2005).

[134] K. E. Stephan, W. D. Penny, R. J. Moran, H. E. den Ouden, J. Daunizeau, and K. J. Friston, “Ten simple rules for dynamic causal modeling”, Neuroimage **49**, 3099 (2010).

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

[136] E. G. Bruneau, N. Jacoby, and R. Saxe, “Empathic control through coordinated interaction of amygdala, theory of mind and extended pain matrix brain regions”, Neuroimage **114**, 105 (2015).

[137] J.-W. Seok and C. Cheong, “Dynamic Causal Modeling of Effective Connectivity During Anger Experience in Healthy Young Men: 7T Magnetic Resonance Imaging Study”, Advances in Cognitive Psychology **15**, 52 (2019).

[138] T. Schuwerk, K. Döhnel, B. Sodian, I. R. Keck, R. Rupprecht, and M. Sommer, “Functional activity and effective connectivity of the posterior medial prefrontal cortex during processing of incongruent mental states”, Human Brain Mapping **35**, 2950 (2014).

[139] S. Esménio, J. M. Soares, P. Oliveira-Silva, P. Zeidman, A. Razi, Ó. F. Gonçalves, K. Friston, and J. Coutinho, “Using resting-state DMN effective

connectivity to characterize the neurofunctional architecture of empathy", Scientific reports **9**, 1 (2019).

[140] T. Schuwerk, M. Schurz, F. Müller, R. Rupprecht, and M. Sommer, "The rTPJ's overarching cognitive function in networks for attention and theory of mind", Social cognitive and affective neuroscience **12**, 157 (2017).

[141] K. D. Davis, W. D. Hutchison, A. M. Lozano, R. R. Tasker, and J. O. Dostrovsky, "Human anterior cingulate cortex neurons modulated by attention-demanding tasks", Journal of neurophysiology **83**, 3575 (2000).

[142] S. N. Schmidt, C. A. Sojer, J. Hass, P. Kirsch, and D. Mier, "fMRI adaptation reveals: The human mirror neuron system discriminates emotional valence", Cortex **128**, 270 (2020).

[143] M. Isoda, "Understanding intentional actions from observers? viewpoints: a social neuroscience perspective", Neuroscience research **112**, 1 (2016).

[144] M. Schurz, J. Radua, M. G. Tholen, L. Maliske, D. S. Margulies, R. B. Mars, J. Sallet, and P. Kanske, "Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind.", Psychological Bulletin (2020).

[145] F. Donnarumma, M. Costantini, E. Ambrosini, K. Friston, and G. Pezzulo, "Action perception as hypothesis testing", Cortex **89**, 45 (2017).

[146] J. Hass, S. Ardid, J. Sherfey, and N. Kopell, "Constraints on Persistent Activity in a Biologically Detailed Network Model of the Prefrontal Cortex with Heterogeneities", bioRxiv , 645663 (2019).

[147] T.-W. Lee, O. Josephs, R. J. Dolan, and H. D. Critchley, "Imitating expressions: emotion-specific neural substrates in facial mimicry", Social cognitive and affective neuroscience **1**, 122 (2006).

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

[149] E. Prochazkova and M. E. Kret, "Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion", Neuroscience & Biobehavioral Reviews **80**, 99 (2017).

[150] L. Hertäg, J. Hass, T. Golovko, and D. Durstewitz, "An approximation to the adaptive exponential integrate-and-fire neuron model allows fast and predictive fitting to physiological data", Frontiers in computational neuroscience **6**, 62 (2012).

[151] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons", Proceedings of the National Academy of Sciences **95**, 5323 (1998).

[152] W. Maass and H. Markram, “Synapses as dynamic memory buffers”, Neural Networks **15**, 155 (2002).

[153] O. Kornienko, “Neural representations and decoding with optimized kernel density estimates”, PhD thesis, University of Heidelberg, 2015.

[154] D. Durstewitz, *Advanced Data Analysis in Neuroscience: Integrating Statistical and Computational Models*, Springer, 2017.

[155] S. Ikemoto, C. Yang, and A. Tan, “Basal ganglia circuit loops, dopamine and motivation: a review and enquiry”, Behavioural brain research **290**, 17 (2015).

[156] M. Gergelyfi, B. Jacob, E. Olivier, and A. Zénon, “Dissociation between mental fatigue and motivational state during prolonged mental activity”, Frontiers in behavioral neuroscience **9**, 176 (2015).

[157] J. K. Seamans and C. R. Yang, “The principal features and mechanisms of dopamine modulation in the prefrontal cortex”, Progress in neurobiology **74**, 1 (2004).

[158] C. C. Lapish, S. Kroener, D. Durstewitz, A. Lavin, and J. K. Seamans, “The ability of the mesocortical dopamine system to operate in distinct temporal modes”, Psychopharmacology **191**, 609 (2007).

[159] M. Joshua, A. Adler, and H. Bergman, “The dynamics of dopamine in control of motor behavior”, Current opinion in neurobiology **19**, 615 (2009).

[160] A. Meyer-Lindenberg, T. Nichols, J. Callicott, J. Ding, B. Kolachana, J. Buckholtz, V. Mattay, M. Egan, and D. Weinberger, “Impact of complex genetic variation in COMT on human brain function”, Molecular psychiatry **11**, 867 (2006).

[161] M. Akil, B. S. Kolachana, D. A. Rothmond, T. M. Hyde, D. R. Weinberger, and J. E. Kleinman, “Catechol-O-methyltransferase genotype and dopamine regulation in the human brain”, Journal of Neuroscience **23**, 2008 (2003).

[162] K. J. Friston, V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. Van Wijk, G. Ziegler, and P. Zeidman, “Bayesian model reduction and empirical Bayes for group (DCM) studies”, Neuroimage **128**, 413 (2016).

[163] P. Zeidman, A. Jafarian, M. L. Seghier, V. Litvak, H. Cagnan, C. J. Price, and K. J. Friston, “A guide to group effective connectivity analysis, part 2: Second level analysis with PEB”, Neuroimage **200**, 12 (2019).

[164] K. Friston and W. Penny, “Post hoc Bayesian model selection”, Neuroimage **56**, 2089 (2011).

[165] M. Rosa, K. Friston, and W. Penny, “Post-hoc selection of dynamic causal models”, Journal of neuroscience methods **208**, 66 (2012).

[166] Z. Yan, S. N. Schmidt, J. Frank, S. H. Witt, J. Hass, P. Kirsch, and D. Mier, "Hyperfunctioning of the right posterior superior temporal sulcus in response to neutral facial expressions presents an endophenotype of schizophrenia", Neuropsychopharmacology **45**, 1346 (2020).

List of Figures

1.1. The mirror neuron system. 18

2.1. The decomposition of the log evidence into negative free energy and a KL divergence. 28

3.1. A schematic illustration of two sigmoid functions with different slopes and the corresponding linear functions. 39
3.2. Imitation paradigm with timing of trials. 43
3.3. Results of the FFX Bayesian model comparison of two models of the forward and backward attention modulation (M1 and M2) for the bilinear models, and the W-C models. 46
3.4. Results of Bayesian model comparisons for all possible models in the family-level with standard bilinear equations and W-C equations. 47
3.5. Bayesian model comparison (RFX) for the two-state model at the family-level. 48
3.6. Observed and predicted time-series in DCM analysis of bilinear (left) and W-C (right) models for one subject. 49
3.7. BMA results for Bilinear and W-C models. 50
3.8. Synthetic data result. 52

4.1. Imitation task. Overview over the four conditions and task flow with presentation times 62
4.2. Empathy task. Overview over the four conditions and task flow with presentation times 63
4.3. ToM task. Overview over the four conditions and task flow with presentation times 64
4.4. Model space. Schemata of parameters that made up the models included in four families. 67
4.5. Family level inference performed on models within the families in Fig. 4.4 69
4.6. BMA results for all three tasks for the winning family one. 70

5.1. Raster plot of the spike times in the network in response to the spiking input into the pyramidal cells L2/3. 78
5.2. Spiking network model f-I curve for a single region. 79
5.3. Network structure of the best model and the illustration of the two-state neuronal model implemented in DCM for two distinct regions. 80
5.4. Raster plot of the spike times and the time series of the observed, predicted, and simulated data sets for the network of three regions. 82
5.5. Spiking network model f-I curve for a network of three regions connected as Fig. 5.3A for six different runs. 83

5.6. The f-I curves' slopes vs. varying three neuron parameters. 85
5.7. Raster plot of the spike times and the time series of the observed, predicted, and simulated data sets for the network of three regions after the f-I curve and the connectivity interplay and changing the neurons' parameters. 86

6.1. The firing rate time series of the spiking network model, layer 2/3 and layer 5, in response to spiking inputs at the interval between 6000 and 7000 ms, when there is no DA modulation and D1, D2, D1&D2 receptors are activated. 88
6.2. Raster plots of the spike times in the network of 1000 neurons during DA modulation with one cell assembly implemented in Layer 2/3. . 89
6.3. Raster plots of the spike times in the network of 1000 neurons during DA modulation with two cell assemblies implemented in Layer 2/3. 90
6.4. Posterior parameter estimates of the matrices A (9 parameters) and B (6 parameters) based on the BMA performed on the final 256 models of an automatic parameter search for the imitation task. . . 92

List of Tables

2.1. Interpretation of Bayes factors. 31

3.1. RFX BMS results of the single-state and two-state models for both bilinear and W-C models and also the comparison between the single-state and the two-state W-C model (Fig. 3.5B). 47
3.2. BMA results (Matrix A). 50
3.3. BMA results (Matrix B). 51

4.1. RFX BMS results on models within family 1 for all experiments. . . 68
4.2. BMA results for endogenous connectivity (Matrix A) (in Hz). . . . 70
4.3. BMA results for modulatory connectivity (Matrix B)(in Hz). . . . 71

5.1. The posterior expected values (mean) and the probability of each value for the parameters. 81
5.2. The posterior expected values (mean) and the probability of each value for the parameters of six different runs. 84

Curriculum Vitae

Sadjad Sadeghi,
born on January 29th, 1987 in Jahrom, Iran

2016-2022	Ph.D. degree in physics, Heidelberg University, Heidelberg, Germany
2010- 2013	M.Sc. degree in physics, Institute for Advanced Studies in Basic Science (IASBS), Zanjan, Iran
2006- 2010	BSc degree in physics, Razi University, Kermanshah, Iran

List of Publications

The following list contains authored scientific publications which were submitted during the time as a Master and Ph.D. student:

- **S. Sadeghi**, S. N. L. Schmidt,D. Mier, and J. Hass, *Effective connectivity of the human mirror neuron system during social cognition*, Social Cognitive and Affective Neuroscience **Revise**, (2021).

  This study is included in Chapter 4.

- **S. Sadeghi**, D. Mier, M. F. Gerchen, S. N. L. Schmidt, and J. Hass, *Dynamic Causal Modeling for fMRI With Wilson-Cowan-Based Neuronal Equations*, Frontiers in Neuroscience **14**, 593867 (2020).

  This study is included in Chapter 3.

- **S. Sadeghi**, A. Valizadeh, *Synchronization of delayed coupled neurons in presence of inhomogeneity*, Journal of Computational Neuroscience **36**, 55-66 (2014).

Contributions to Scientific Conferences & Grants

The following list contains leading-author contributions to scientific conferences during the time as a Ph.D. student:

- **S. Sadeghi**, D. Mier, J. Hass, *A global network of human mirror neuron system: An fMRI DCM study* (Poster), Bernstein Conference, Berlin (2018).

- **S. Sadeghi**, D. Mier, J. Hass, *A detailed spiking network model of the human mirror neuron system* (Poster), Bernstein Conference, Göttingen (2017).
- **S. Sadeghi**, D. Mier, J. Hass, *An fMRI Dynamical Causal Modeling study with Wilson-Cowan based neuronal equations* (Poster), Bernstein Conference , Berlin (2016).

The following list contains my received Grants during the time as a Ph.D. student:

- The Completion grant funded through the Landesgraduiertenförderung (LGF) program (2020), The Graduate Academy, Heidelberg University, Germany.
- Travel grant for participation in the Bernstein Conference, Berlin (2016).

Acknowledgements

Firstly, I would like to thank my supervisor, Joachim Hass, from the bottom of my heart, for allowing me to start my doctorate study at Heidelberg University and also for his continuous support of my Ph.D. study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D. study.

My sincere thanks also go to Daniela Mier, my advisor, and Stephanie, who provided me with the measurements. Whenever I had any questions about the data sets and the experimental parts of the study, they promptly provided me with explanations. They always brought energy and thoughtful insight into our discussions, and their inspirational comments on my research were priceless.

Besides, I would like to thank Daniel Durstewitz for letting me be a part of his group at the Central Institute of Mental Health (CIMH) as a guest researcher and providing me access to the laboratory and research facilities. Without his precious support, it would not be possible to conduct this research.

I would like to thank Mark. E. Ladd, who graciously agreed to review this thesis, and Björn Malte Schäfer and Klaus Reygers for accepting to be my examiners at my defense session.

Amongst many others, I would like to thank the following people from the theoretical neuroscience group and CIMH: Christine, Eleonora, Dominik, Hazem, Grant, Thomas, Sven, Georgia, Fungi, Carla, Florian, Loreen, Sebas, Olga, Leonard, Janine, Zahra, Max, Philine, Manuel, Simone, and Janet. Of these, Fungi, whose own research was an essential inspiration for this thesis, deserve special mention.

I thank the Heidelberg Academy of Sciences and Humanities, who funded the project and my Ph.D. study. I also thank the Graduate Academy of Heidelberg University for awarding me a dissertation completion grant, providing me with the financial means to complete this thesis.

Last but not least, I would like to thank my family, the most important people of my life: my wife, Lida, to who I have dedicated my thesis; thanks for all your support, without which I would have stopped these studies a long time ago. I thank my parents, brother, and sister for supporting me spiritually throughout writing this thesis and my life in general. I would also like to thank my friends, Shabnam and Elham, for encouraging and supporting me whenever I needed them.