Aus der Klinik für Psychiatrie und Psychotherapie
Geschäftsführender Direktor: Prof. Dr. T. Kircher

des Fachbereiches Medizin der Philipps-Universität Marburg

Connectivity models in the neural face perception domain – interfaces to understand the human brain in health and disease?

Inaugural-Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften
dem Fachbereich Medizin der Philipps-Universität Marburg
vorgelegt von

Roman Keßler
aus Wolfach

Marburg, 2022

Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg am 27.07.2022.

Gedruckt mit Genehmigung des Fachbereichs

- Dekanin: Frau Prof. Dr. D. Hilfiker-Kleiner
- Referent: Herr Prof. Dr. A. Jansen
- 1. Korreferentin: Frau PD Dr. M. Bopp

Publications

Journal articles (selection)

1. **Long-term Neuroanatomical Consequences of Childhood Maltreatment: Reduced Amygdala Inhibition by Medial Prefrontal Cortex**
   Roman Kessler, Simon Schmitt, Torsten Sauder, Frederike Stein, Dilara Yüksel, Dominik Grotegerd, Udo Dannlowski, Tim Hahn, Astrid Dempfle, Jens Sommer, Olaf Steinsträter, Igor Nenadic, Tilo Kircher & Andreas Jansen

   Published in Frontiers in Systems Neuroscience (2020, Impact Factor: 3.293)

   [Kessler et al., 2020]

2. **Revisiting the effective connectivity within the distributed cortical network for face perception**
   Roman Kessler, Kristin M. Rusch, Kim C. Wende, Verena Schuster & Andreas Jansen

   Published in NeuroImage: Reports (2021)

   [Kessler et al., 2021b]

3. **Function is bound by structure in effective connectivity models**
   Roman Kessler & Andreas Jansen

   manuscript

   [Kessler and Jansen, 2022]

Additional articles

[Sahraei et al., 2021] Sahraei, I., Hildesheim, F.E., Thome, I., Kessler, R., Rusch, K.M., Sommer, J., ... & Jansen, A. (2021). Developmental changes within the extended face processing network: A cross-sectional functional magnetic resonance imaging study. Developmental Neurobiology.

[Thome et al., 2021] Thome, I., Hohmann, D.M., Zimmermann, K.M., Smith, M.L., Kessler, R., & Jansen, A. (2021). "I Spy with my Little Eye, Something that is a Face...": A Brain Network for Illusory Face Detection. Cerebral Cortex.

[Kessler et al., 2021a] Kessler, R., Henniger, O., & Busch, C. (2021). Fingerprints, forever young? 25th International Conference on Pattern Recognition (ICPR) (pp. 8647-8654). IEEE.

[Hildesheim et al., 2020] Hildesheim, F.E., Debus, I., Kessler, R., Thome, I., Zimmermann, K.M., Steinsträter, O., ... & Jansen, A. (2020). The trajectory of hemispheric lateralization in the core system of face processing: a cross-sectional functional magnetic resonance imaging pilot study. Frontiers in Psychology, 11.

[Kessler et al., 2016] Kessler, R., Bach, M., & Heinrich, S.P. (2016). Two-tactor vibrotactile navigation information for the blind: Directional resolution and intuitive interpretation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(3), 279-286.

Abstract

The recognition and processing of faces is a core competence of our human brain, in which many neuronal areas are involved. Faces are not only a means to recognize and distinguish between individuals, but also a means to convey emotions, intentions, or trustworthiness of our counterpart. The processing of faces is an orchestrated interaction of a multitude of neuronal regions. This interplay can be quantified at the neuronal level using so-called effective connectivity analysis. The most common effective connectivity analysis, which is also used in the present work, is called Dynamic Causal Modeling. With its help, interregional interactions are modelled at the neuronal level, and at the measurable level – such as with functional magnetic resonance imaging – evidence is found for the probability of the presence of neuronal connections and also their quantitative expression. Effective connectivity analyses can thus reveal the couplings between brain areas during specific cognitive processes, such as face perception.

The way we process faces also changes when, for example, mental illness is present. Thus, negative emotions such as fear may be perceived disproportionately more intense, or positive emotions such as joy less intense. The evaluation of neuronal parameters in face processing could be used in clinical practice, e.g. for the early detection of mental illnesses or the quantification of therapy success.

A prerequisite for clinical application is the reliability of the modeling method. Thus, results of models should be generalizable and not depend on certain nuances of the modeling. Furthermore, the interpretability of many model parameters turns out to be difficult. However, this is necessary to be able to describe causal relationships.

In the present dissertation, so-called Dynamic Causal Models are applied in the field of neural face processing. In a first study a clinical context is used. Here, neural models of emotion regulation in face processing were used to identify potential consequences of risk factors for the development of mental illness. In another study, the generalizability of neural network models was tested in a healthy population. Here, many limitations of the method as a whole were revealed. In a final study, both observed and simulated data were used to uncover more limitations in the interpretation of model parameters.

Keywords: face perception, neural processing of faces, functional magnetic resonance imaging, effective connectivity, Dynamic Causal Modeling

Zusammenfassung

Die Erkennung und Verarbeitung von Gesichtern ist eine Kernkompetenz unseres menschlichen Gehirns, an welcher viele neuronale Areale beteiligt sind. Gesichter dienen nicht nur zur Erkennung und Unterscheidung zwischen Individuen, sondern transportieren zum Beispiel auch Emotionen, Absichten, oder Vertrauenswürdigkeit unseres Gegenübers. Dabei ist die Verarbeitung von Gesichtern ein orchestriertes Zusammenspiel einer Vielzahl von neuronalen Regionen. Dieses Zusammenspiel kann mittels der sogenannten effektiven Konnektivitätsanalyse auf neuronaler Ebene quantifiziert werden. Die häufigste, und auch in der vorliegenden Arbeit verwendete Konnektivitätsanalyse trägt den Namen Dynamic Causal Modeling. Mit ihrer Hilfe modelliert man interregionale Interaktionen auf neuronaler Ebene, und findet auf messbarer Ebene – wie z.B. mit funktioneller Magnetresonanztomographie – Hinweise für die Wahrscheinlichkeit neuronaler Verbindungen und auch deren quantitative Ausprägung. Mit Hilfe von effektiven Konnektivitätsanalysen können somit die Kopplungen zwischen Hirnarealen bei bestimmten kognitiven Vorgängen, wie z.B. der Gesichterwahrnehmung, aufgedeckt werden.

Die Art und Weise, wie wir Gesichter verarbeiten, ändert sich beispielsweise, wenn z.B. psychische Erkrankungen vorliegen. So können negative Emotionen wie Furcht unproportional stärker wahrgenommen werden, oder positive Emotionen wie Freude weniger stark. Die Auswertung neuronaler Kennwerte bei der Gesichterverarbeitung könnte perspektivisch im klinischen Alltag zum Einsatz kommen, z.B. zur Früherkennung von psychischen Erkrankungen, oder der Quantifizierung von Therpieerfolg.

Voraussetzung für einen klinischen Einsatz ist jedoch eine Verlässlichkeit der Modellierungsmethode. So sollten Ergebnisse von Modellen generalisierbar sein, und nicht von bestimmten Nuancen der Modellierung abhängen. Weiterhin stellt sich die Interpretatierbarkeit vieler Modellparameter als schwierig heraus. Diese ist jedoch notwendig, um ursächliche Zusammenhänge beschreiben zu können.

In der vorliegenden Dissertation werden sogenannte Dynamic Causal Models im Bereich der neuronalen Gesichterverarbeitung eingesetzt. In einer ersten Studie wird ein klinischer Kontext herangezogen. Hier wurden anhand neuronaler Modelle der Emotionsregulation in der Gesichterverarbeitung Auswirkungen von möglichen Risikofaktoren zur Entwicklung psychischer Erkrankungen auf die Hirnkonnektivität erkannt. In einer weiteren Studie wird die Generalisierbarkeit neuronaler Netzwerkmodelle an einer gesunden Population erprobt. Hier zeigten sich viele Limitationen der Methode als Ganzes auf. In einer letzten Studie werden sowohl mit echten, als auch mit simulierten Daten, weitere Limitationen in der Interpretation von Modellparametern aufgedeckt.

Keywords: Gesichterwahrnehmung, Neuronale Verarbeitung von Gesichtern, funktionelle Magnetresonanztomographie, effektive Konnektivität, Dynamic Causal Modeling

Abbreviations

BMA Bayesian Model Averaging

BMS Bayesian Model Selection

BOLD Blood Oxygen Level Dependent

DCM Dynamic Causal Modeling

spDCM spectral Dynamic Causal Modeling

stDCM stochastic Dynamic Causal Modeling

DWI Diffusion Weighted Imaging

FFA Fusiform Face Area

HLM Hierarchical Linear Modeling

IFG Inferior Frontal Gyrus

MD Major Depression

MRI Magnetic Resonance Imaging

fMRI functional Magnetic Resonance Imaging

OFA Occipital Face Area

OFC Orbitofrontal Cortex

mPFC medial Prefrontal Cortex

STS (posterior) Superior Temporal Sulcus

Contents

Chapter 1

Introduction

'An agent does not have a model of its world - it is a model. In other words, the form, structure, and states of our embodied brains do not contain a model of the sensorium - they are that model. Every aspect of our brain and body can be predicted from our environment.' – *Karl Friston*

1.1 What is a model

In our everyday life, we are surrounded by models. Models are used for weather forecasting [Saxena et al., 2013], and models are used for personalized advertising [Wirtz et al., 2017]. Models are used to describe much of the primary questions of physics [Redhead, 1980], as well as our understanding of the biological evolution up to social interactions in the complex ecosystems we live in [Darwin, 1859, Hartmann et al., 2008]. Without models, we as a global society, and in particular we as scientists, would not stand at this point. That is, we as society would lack much of the outstanding knowledge we obtained access to during the last centuries. On the other hand we as scientists would lack essential toolboxes to generate such knowledge. Without models to interpret large and unstructured amounts of data, our understanding from the world would rely upon anecdotes.

The human brain is presumed to create models of the outside world [Friston and Kiebel, 2009]. Each sensory input contributes to the shape of our internal model, therefore, learning is an continuous updating of our internal models about our environment [Clark, 2013]. To raise the stakes even further, Karl Friston – as pointed out in the opening quotation – proposes not only the brain (i.e., we ourselves) containing models, but rather being large models by themselves (i.e., ourselves) [Friston, 2013].

However, in our world we usually talk about less complex models than the human brain. Models which we use as tools, e.g., in sciences. Such models can be deployed with different aims. First, models can be used to *describe* the shape of data, such as the distribution of some characteristic of a group [Fisher and Marshall, 2009]. Second, models can be used to *infer* population properties or mechanisms, based on a sample [Allua and Thompson, 2009]. Third, models can be used to *predict* current or future states of a characteristics [Larose, 2015].

Heinze and colleagues [Heinze et al., 2018] defined some minimally required properties of a model. First, it should be *valid*, such as it provides predictions with acceptable accuracy. To be valid, a model must also be *reliable*. If a model is not reliable, such as it encompasses comparable model parameters when fitted to a similar subset of data, its validity must be questioned. Second, it should be practically *useful*, hence promoting conclusions, and therefore it should be interpretable. Third, it should be *simple*. An overly complex model is deemed to be either not interpretable, or is subject to high variance, missing the fundamental ability to generalize to novel data.

The studies presented in this dissertation will make use of models encompassing different of the proposed aspects. The models of study A are proposed to be *useful*, in a way that they *describe* differences in brain connectivity between different groups of participants for which a

possible underlying mechanism is *inferred* [Kessler et al., 2020] (Appendix A). In study B, we question the *usefulness* of a very popular model by testing its robustness and *reliability*, in terms of *replicability* across different data sets [Kessler et al., 2021b] (Appendix B). All models are such *simple*, that they can be used to illustrate the neural mechanisms they are aiming at describing in an understandable fashion. However, with study C, we question the *validity* of these kinds of models in general [Kessler and Jansen, 2022] (Appendix C). In particular, the overall interpretability of such models are questioned, and commonly conducted pitfalls are illustrated, largely constraining any interpretability and therefore the models' *usefulness*.

1.2 Models to describe brain function

Researchers in cognitive neurosciences use models to render research results understandable, teachable, and therefore accessible. One well-known example is the dual-stream model of visual processing in the human brain [Van Essen and Deyoe, 1995, Ungerleider and Haxby, 1994]. The dual-stream model separates visual processing pathways along the visual hierarchy into a ventral stream, largely responsible for object identification, and a dorsal stream, largely responsible for object localization [Ungerleider and Haxby, 1994]. Both streams originate in early visual cortex, and increase in complexity of computations towards higher visual areas, i.e., towards downstream levels of the respective streams [Milner, 2017]. This exemplary but famous model is rather vague in its details, but serves as working model in research and teaching. However, it describes interactions between many, partly distant brain regions during a cognitive processing task such as visual perception. Other models of the brain describe more fine-grained operations within the brain. For example, the model of Rao and Ballard illustrates microscopic computational interactions between regions of the early visual cortex [Rao and Ballard, 1999]. On the other hand, the model of Bastos et al. illustrates even more granular computations within a single cortical column of early visual processing [Bastos et al., 2012]. Whereas these fine-grained models are suitable to test hypotheses or describe interaction in insulated regions, they neglect the complex interplay of these regions with other cortical regions. On the contrary, large-scale models describing the interactions between regions – such as the dual-stream model – can integrate rather coarse patterns of brain activation. However those lack some computational granularity and neglect the many, microscopic processes occurring within each and every single region. Both kinds of models are important to understand particular phenomena. However, integrating both kinds of models has not yet become computationally traceable. Therefore, we must often decide for one of the two kinds of models, depending on the research question and the spatial extend on which we want to test hypotheses. The models applied throughout this thesis are larger-scale

models, aiming at describing interactions between regions on a macroscopic scale.

1.3 Models of face processing

One famous model in the domain of face perception is the so-called *Haxby model* [Haxby et al., 2000]. The Haxby model aims at describing interactions between regions involved in human face perception. Haxby distinguished between *core system* and *extended system* of face perception [Haxby et al., 2000]. Whereas areas of the core system rather process basic information about faces, areas of the extended system get involved to process specific aspects such as emotional expression or biographical information [Gobbini and Haxby, 2007]. Within the core system, the Occipital Face Area (OFA) processes single features of the face, the Fusiform Face Area (FFA) processes the face rather holistically, and the (posterior) Superior Temporal Sulcus (STS) processes dynamic aspects such as facial expressions [Haxby et al., 2000]. Regions of the extended system encompass limbic regions such as the amygdala, crucial in the processing of emotional content [Adolphs, 2002], alongside with often heteromodal regions of the prefrontal cortex such as the Orbitofrontal Cortex (OFC), or the Inferior Frontal Gyrus (IFG) [Duchaine and Yovel, 2015]. The Haxby model emerged from early research results mainly based on functional Magnetic Resonance Imaging (fMRI) activation studies. It is an often-adopted working model in many studies of the face perception network (e.g., [Frässle et al., 2016b, Fairhall and Ishai, 2007, Zhang et al., 2009]), and throughout the years underwent refinements and revisions [Duchaine and Yovel, 2015]. In the dissertation at hand, we tested a computational implementation of the Haxby model [Fairhall and Ishai, 2007] – describing the neural interactions between the regions of the core system – for its replicability and stability (appendix B). The Haxby model can not only be used to visualize the sane processing patterns, aimed at understanding face perception in fundamental research. It can further be used to understand abnormal processing patterns, associated with several mental and cognitive disorders, such as prosopagnosia, autism, or depression [Rossion, 2014, Avidan and Behrmann, 2014, Mayberg, 1997, Mayberg et al., 1997, Bi and Fang, 2017, Koshino et al., 2008].

A cognitive disorder associated with impaired face perception (and recognition) abilities is prosopagnosia. Prosopagnosia is referred to as face blindness, and often results from a lesion of relevant areas of the face processing network, such as the OFA [Rossion, 2014]. Prosopagnosia has also been put into the framework of the Haxby model [Avidan and Behrmann, 2014]. Similarly, mental disorders such as autism or Major Depression (MD) can be understood within the framework of computational models in the face perception system. For instance, the Mayberg model [Mayberg, 1997] has frequently been used to explain altered patterns of emotion processing

in MD, or to evaluate treatment effects by antidepressant drugs [Mayberg et al., 1997]. In this thesis, we used this model to delineate altered connectivity patterns during emotional face processing in healthy participants with childhood trauma, a critical risk factor for MD and other mental disorders (appendix A).

1.4 Brain connectivity & Dynamic Causal Modeling

Early fMRI studies were bound to research questions which could be explained by brain activation (i.e., Blood Oxygen Level Dependent (BOLD) signaling) of particular brain regions. This type of investigation, namely the investigation of *functional segregation*, considers each neural area as insulated unit, but neglects its interactions with other areas [Tononi et al., 1994]. In contrast, the investigation of *functional integration* tries to tackle the structural, functional, or so-called effective connectivity between brain regions, by exploring the statistical co-dependencies between the single units or brain areas [Marrelec et al., 2008].

Structural connectivity describes the physical connections, e.g., axon bundles between brain regions. Using Magnetic Resonance Imaging (MRI), structural connectivity can be approximated using Diffusion Weighted Imaging (DWI) [Soares et al., 2013]. Contrarily, functional connectivity assesses the temporal commonalities between the activations of different brain regions [Rogers et al., 2007]. It can be assessed using fMRI, and can be described such as – in its most naïve form – it is nothing more as a correlation between the time series of two or more regions [Rogers et al., 2007]. Whereas structural connectivity is rather time invariant across a modest time span, functional connectivity is very task dependent, and therefore changeable within seconds as when measured with fMRI [Hindriks et al., 2016]. However, the low-pass filter property of the BOLD response hinders a very fine-grained analysis of the causal dependencies on the connectivity between brain regions, such as inference about the direction of information transfer between regions.

Effective connectivity however aims at finessing that drawback by modeling the neuronal interactions on a very fine-grained time scale [Horwitz et al., 2005]. The effective connectivity method used in the dissertation at hand is Dynamic Causal Modeling (DCM) [Friston et al., 2003]. The interactions on neuronal level are vastly described by the so-called neural state equation. Using a hemodynamic forward model, the neuronal states generated by the neural state equation are translated into a predicted hemodynamic signal, which is then compared to the observed hemodynamic signal [Buxton et al., 2004]. Neural model parameters are then estimated by iteratively adjusting the neural and hemodynamic model parameters, and by comparing the time series generated by the model with the actually observed time series [Friston et al., 2003]. This

process is usually referred to as model inversion. DCM aims at allowing for comparison between hypotheses, such as identifying connections between a set of modeled brain regions which are subject of task-dependent alterations.

More precisely: The foundation of DCM is the neural model, i.e., the neural state equation:

$$\dot{z} = Az + (\sum_{j=1}^{k} u^{(j)} B^{(j)})z + Cu$$

The neural state equation models the rate of change of activity $\dot{z}$ as the interactions between regions, depicted in A, B, and C matrices. The C matrix depicts the direct driving input into brain regions of the model. Basically, if a model solely comprised a C matrix, the informative value of the model parameters closely corresponded to activation studies targeting functional segregation. However, A and B matrices spark opportunities to explore functional integration. The A matrix thereby describes the task independent magnitude of the rate of change of activity between regions. The B matrix, most critically, describes the change in this rate which is induced my experimental perturbation. The sums of A matrix and B matrix (or rather B matrices) gets multiplied by the current activity z of the regions to obtain the rate of change. All parameters of the neural state equation get estimated during an iterative model inversion process, using observed hemodynamic data and the forward model.

DCM exists in different flavors. The vanilla implementation was mostly applicable to tackle research questions about the influence of experimental manipulations on particular connections [Friston et al., 2003]. However, with the uprising resting-state *functional connectivity* research, aiming at finding, e.g., neural biomarkers for particular disorders such as neurodegenerative disorders (e.g., [Hohenfeld et al., 2018]), autism (e.g., [Dvornek et al., 2017]), or psychiatric disorders (e.g., [Yamada et al., 2017]), two different flavors of DCM where developed for resting-state data. One of which is stochastic Dynamic Causal Modeling (stDCM) [Daunizeau et al., 2012], modeling stochastic noise. The other method is spectral Dynamic Causal Modeling (spDCM) [Razi et al., 2017], which shifts the model inversion process from the time domain to the spectral domain. Model inversion in **spCPM!** (**spCPM!**) requires less computing power and allows the inversion of larger models. However, in all articles included in the dissertation at hand, the vanilla task-based version for DCM – as described by Friston et al. [Friston et al., 2003] – was applied.

1.5 Motivation and objective of the thesis

In this dissertation, thoughts regarding the potentials as well as the limitations of effective connectivity models in basic research and clinical applications will be outlined. For this aim, I

will first present a study we have conducted using effective connectivity models in the domain of human face perception [Kessler et al., 2020] (Section 2.1). This study demonstrates a use case to reveal potential biomarkers for MD in healthy participants at risk for MD [Kessler et al., 2020] which builds upon an established model in clinical neurosciences [Mayberg, 1997].

Next, replicability of such models will be questioned [Kessler et al., 2021b] (Section 2.2). Therefore, I will present a conceptual replication of one of the earliest models of the core face perception system using contemporary state-of-the-art methods. I will demonstrate different aspects which challenge the replicability of such models, and discuss problems in the interpretation of model parameters.

Finally, I will demonstrate within a particular set of scenarios – using both real data and simulations – that a considerable amount of model parameters are predetermined in the first place [Kessler and Jansen, 2022] (Section 2.3). The results challenge the usability of effective connectivity models in a number of use cases, and therefore should alert the reader to take care in the interpretation of these models and the conclusions drawn by studies using such models.

Chapter **2**

Summary of the published results

'Without data you are just another person with an opinion.'

– *W. Edwards Deming*

2.1 Fronto-limbic dysconnectivity in healthy participants at risk for depression

In the first study included in this dissertation [Kessler et al., 2020] (appendix A), we aimed at applying effective connectivity models to build upon the so-called *Mayberg model* of fronto-limbic dysconnectivity in MD [Mayberg, 1997]. Instead of comparing healthy participants to patients with MD, we rather focused on healthy participants with particular *risks* for MD. By investigating healthy participants, it was possible to delineate connectivity parameters as potential neural biomarkers for the disorder of interest, such as MD. Such biomarkers might eventually be predictive for the occurrence of the disorder. 342 healthy participants were retrieved from a comparably large, in-house built database [Kircher et al., 2019], aiming at detailed genotyping and phenotyping of participants with current or past episodes of affective disorders as well as healthy controls. The healthy participants chosen from this database were classified either as having a family history of affective disorders (i.e., genetic or familial risk), childhood trauma experiences (i.e., environmental risk), both or none of the chosen risks. All participants underwent – among other examinations – an fMRI experiment. In the experimental paradigm participants were shown a series of faces expressing negative emotions such as fear or anger. For each of the participants, several effective connectivity models were constructed. The nodes of these models comprised the amygdala and the medial Prefrontal Cortex (mPFC). The amygdala as limbic brain structure is involved in recognition of emotions, and the mPFC as prefrontal cortical region is supposedly responsible for the control of the limbic reactivity to emotions. The constructed effective connectivitay models were fit to the hemodynamic data retrieved from the experimental paradigm.

The rationale was as follows: In healthy participants, the mPFC is supposed to exert a inhibition upon the amygdala, in order to regulate its reactivity to emotional stimuli [Delgado et al., 2008, Urry et al., 2006, Johnstone et al., 2007]. In the context of an effective connectivity model, this inhibition should manifest itself in a negative coupling parameter from mPFC to amygdala. From a mechanistic perspective, such a negative coupling parameter in turn indicates that with rising activity in mPFC, the rate of change in activity in the amygdala is decreased likewise. We hypothesized that both genetic and environmental risks reduce the inhibition of the mPFC onto the amygdala, leading to less negative coupling parameters.

The results nicely illustrated a reduction of amygdala inhibition by mPFC in healthy controls with environmental risk factors, i.e., childhood maltreatment (see [Kessler et al., 2020], appendix A, Fig. 3). This was expressed by a less negative coupling parameter from mPFC to

the amygdala. The diminished inhibition constitutes a mechanistic explanation of the amygdala over-reactivity in healthy controls with childhood maltreatment experiences reported in the literature [Dannlowski et al., 2012]. In particular, with decreased inhibition, the amygdala is active more sustainably as response to emotional facial expressions, which might in turn lead to adverse behavioral responses by the individual [Cheng et al., 2006]. However, contrary to our expectations, no reduced amygdala inhibition has been found for participants with so-called genetic risk factors, i.e., a family history of affective disorders.

The Mayberg model in its original form described the deficits in amygdala inhibition in patients suffering from MD [Mayberg, 1997]. Our model nicely complemented the Mayberg model by not only describing the current is-state of already diagnosed patients, but providing a tool to describe similar deficits in healthy participants at risk for MD and affective disorders. Whereas the ultimate aim however is to predict the future state – i.e., the probability of the individual at-risk participant to actually develop MD – longitudinal studies need to be conducted to unfold the potential of such connectivity parameters as neural biomarkers.

2.2 Conceptual replication of the Haxby model

In a next study [Kessler et al., 2021b] (appendix B), we were interested to replicate a DCM implementation of the Haxby model, which was published by Fairhall & Ishai nearly 15 years back [Fairhall and Ishai, 2007]. The authors were the first to apply both fMRI and DCM to the Haxby model. The main body of their study tackled the effective connectivity of the core system of face perception, encompassing the OFA, FFA, and STS, during face processing experiments. The DCM software package underwent a lot of methodological enhancements during the years since its first publication, rendering the results more accurate and reliable according to the developers (e.g., [The FIL Methods Group, 2020, The FIL Methods Group, 2014]). However, the results of many old studies, such as the study of Fairhall & Ishai – which can be considered as a milestone study in face perception and connectivity – remained unquestioned until today, even though the methodology has grown further. Their model is still deployed as working model for many related studies (e.g., [Elbich et al., 2019, Frässle et al., 2016b, Frässle et al., 2016a, He et al., 2015, Nagy et al., 2012, Sato et al., 2017]).

For these reasons, we aimed at replicating their model, with a contemporary, state-of-the-art implementation of DCM. To increase generalizability of the results, we analyzed four different data sets, partly acquired in our laboratory and partly freely available in the internet. Some of the data sets contained multiple measurements for each participant. In addition to replicating the original results with contemporary software and across multiple data sets, we also discovered

severe conceptual issues and miss-interpretations in the original study. Therefore, we changed major aspects in the modeling procedure and in the interpretation of results. All in one, we conducted a comprehensive conceptual replication approach, included several additions, and further straightened some conceptual flaws of the original study.

Our results demonstrated, that the core system of face perception as proposed by Haxby [Haxby et al., 2000] is more densely interconnected as described by the study of Fairhall and Ishai [Fairhall and Ishai, 2007]. Whereas in their study, primarily feedforward connections leading from OFA to FFA and from OFA to STS have been highlighted as playing a role in face perception, we delineated the similar important backward connections as well as collateral connections [Kessler et al., 2021b] (appendix B) in face perception and in emotion perception. Furthermore, we showed an astonishing stability across our tested data sets, which were heterogeneous both in their experimental set-up and their preprocessing pipelines. Furthermore, we eliminated some conceptual limitations of the original study, and discussed differences in results deriving from those. A meta analytic approach further revealed connectivity parameters which were similar across all our included data sets. This resulted in a model of which we think it is a more suitable working model than the one proposed by the authors of the original study, with a higher capability for generalization. However, it also underlined the importance for replications of these kind of models, as methodological developments advance continuously and old models are rarely updated.

2.3 Limitations in interpretabiliy of effective connectivity models

The replication of study B (Section 2.2) revealed major flaws in interpretability of many model parameters by researchers dealing with effective connectivity models such as DCM. With study C we aimed at tackling further issues of interpretability, by investigating how the shape of a model parameter is predestined by the setup of the model itself [Kessler and Jansen, 2022] (appendix C). This is crucial, because if the shape of a model parameter – i.e., if it is positive or negative – is predetermined by the experimental setting rather than by the data, interpretation upon it must be drawn with caution. We argue, that many effective connectivity models with a hierarchical structure, applied to stereotypical experimental scenarios, can't deliver additional value to the study, because the shape of many of the models' parameters are predetermined by the experimental setup and not a meaningful outcome of the experiment.

In this study, we used both real experimental data and simulated data of a stereotypical experiment. In this experiment, each modeled brain region was activated by the experimental condition, meaning the BOLD signal increased after stimulus onset of the experimental condition. Furthermore, the BOLD signal was lower in the control condition. A comparable intuition

might apply for most fMRI experiments. By modeling the regional dynamics using DCM, we demonstrated that parameters of so-called *forward connections* turn almost always positive. Contrary, parameters of so-called *backward connections* turn almost always negative. This can be easily shown by varying the region in which the experimental input enters the model (i.e., brain network) in the first place, and observing the shape of the interregional coupling parameters as response to this change.

When modeling such an experiment, the interpretation of the parameter shapes is rather straightforward. For instance, the positivity of the forward connection is necessary to distribute the activity (i.e., positive BOLD signal) across all modeled brain regions. However, the negativity of the backward connections seems to have its origins in a supportive role for the negative self-connections of each region. It therefore aids the system to decrease its activation, i.e., rendering it stable.

In addition, we demonstrated that the very same pattern is captured either by the A matrix or by the B matrix of a DCM. If a B matrix is present – i.e., modeled – it will capture the very pattern, such as the positivity of the forward connections and the negativity of the backward connections are represented within the B matrix. If it is not present, the pattern will shift to the A matrix and will be less pronounced. The reason for the differences in the absolute parameter magnitudes can be traced back to the differences in prior variances of the respective matrix parameters, with the B matrix parameters having a far higher prior variance than the A matrix parameters, allowing for a higher disparity from their prior distribution after being confronted with experimental data. However, the interpretation of A and B matrices is different from a mechanistic point of view, as the A matrix represent the context-independent couplings whereas the B matrix represents the couplings induced by experimental perturbations [Friston et al., 2003]. Therefore, misuse of A and B matrices, as it has been done in the literature e.g., by leaving out a B matrix where it was appropriate to include it or mixing up A and B matrices (e.g., [Straube et al., 2018], [Fairhall and Ishai, 2007]), can lead to misinterpretation of the acquired results.

As we worked also with simulated data in this study, we were able to test the behavior of the shape of a model, when we exactly know the ground truth model structure, i.e., the structure and shape of the model which generated the data. We were able to demonstrate, that even if in the ground truth model, no B matrix was present (equivalent to an experiment with no modulation by experimental perturbation), allowing effects within the B matrix of the tested model will pull the positivity/negativity pattern towards itself. As a consequence, the resulting model provides answers to question for which it is not supposed to have answers to from a theoretical point of view.

This study largely limits the interpretability of many model parameters, as we were able to predict the shape (i.e., positive or negative) of a parameter depending on its position in the model hierarchy. We argue, that throughout literature, many studies over-stated the meaning of the value of particular model parameters, which were falling into this very pattern, and their meaning is therefore of technical origin rather than reflecting some cognitive function.

Chapter 3

Discussion

'All generalizations are false, including this one.'

– *Mark Twain*

3.1 The failed transfer of neural network models into clinical application

Having elaborated possible applications of models and in particular connectivity models in cognitive neurosciences, the question arises why are those kind of models not implemented in practical workflows in academia or clinical application. For instance, why have neural connectivity parameters – often extolled as clinical biomarkers (e.g., [Heinzle and Stephan, 2018, Hohenfeld et al., 2018, Damoiseaux, 2012]) – not made their ways into clinical routine? One out of many reasons might be the fact that many models have a weak predictive power when it comes to forecasting the outcome for a single patient [Brennan et al., 2019, Plitt et al., 2015]. Moreover, many studies including our own [Kessler et al., 2020], provide statistical differences of connectivity parameters between two or more groups. For instance in Figure 3 of the first study included in this dissertation [Kessler et al., 2020] (appendix A) this statistical difference is illustrated on a group level. When however focusing on the level of an individual participant [Kessler et al., 2020] (e.g., appendix A, supplementary Fig. S3), the picture is less concise. The distributions of parameter estimates highly overlap between groups. Therefore, the single expression of a variable of interest is often not sufficient for a precise and therefore useful prediction. In this case, the prediction of the risk group would be possible, i.e., better than chance level, but weak. The prediction of individual disease onset however would be difficult to be evaluated with cross-sectional data.

Furthermore, prediction might work better when more parameters of interest are included, i.e., the number of non-redundant variables is increased. To build complex models which provide precise predictions, a vast amount of training data is needed – so-called big data. Scaling training data into the millions might be manageable for image classification problems such done in facial recognition models [Meng et al., 2021, Deng et al., 2019, Parkhi et al., 2015, Serengil and Ozpinar, 2020]. Whereas the capturing or even simple crawling of face images is easily scaleable, neural data is far more expensive and time consuming to collect. For instance, specialized acquisition protocols and equipment (i.e., MRI machines) and trained personnel is required. The temporal and financial efforts to acquire enough data to build models with nearly as precise predictions as in other domains still constitute an insurmountable obstacle to apply more elaborate models of machine learning to this area.

Next, out of the abundance of studies published about arbitrary group differences in connectivity parameters, those results might have low replicability and reliability. For instance, especially in the domain of imaging in cognitive neurosciences, a high portion of study results are not replicable in the first place [Poldrack et al., 2017, Hong et al., 2019]. Whereas this must be

an alarm call by itself, few lessons with impact have been adopted by the concerned research community [Shrout and Rodgers, 2018, Miłkowski et al., 2018, Poldrack and Gorgolewski, 2014]. Even if group results are replicable, their reliability might still be poor. Furthermore, in the area of functional and effective connectivity, good reliability of connectivity parameters is frequently propagandized as study result, whereas a look into the actual numbers is less promising, and fails to pass minimum stability requirements we would anticipate for any application (e.g., [Noble et al., 2021, Frässle et al., 2015, Almgren et al., 2018]).

Finally, having identified sample size issues in neuroimaging studies, poor replicability, and poor reliability as some of the driving problems in the translation from academic research results to clinical application, one fundamental flaw is limitation in interpretability of some model parameters. Potential biomarkers such as introduced in study A [Kessler et al., 2020] (appendix A) appear to have a sound interpretation within their model framework. Similarly, they can easily be embedded in theories of emotion regulation (e.g., [Mayberg, 1997]). However, having identified group differences in the parameter, neither replicability nor reliability has been shown within the study. For this, other study designs are necessary.

3.2 Limitations of the presented studies

When it comes to model interpretability, I demonstrated in study C [Kessler and Jansen, 2022] (appendix C), that model parameters might be predetermined to a large degree in the first place. The claims of study C can not be transferred to the results of study A directly, as the model is not strictly hierarchical. In study A we combined 12 different models for the resulting presented average model. However, some of the single models were hierarchical, at least those comprising only one input region. Furthermore, both regions modeled in study A were activated by the experimental condition (i.e., faces), compared to the control condition (i.e., shapes). The activation therefore needed to be propagated throughout the system following the intuition and schema carved out in study C.

The results of study B however can be more affected. The prerequisites were similar between study B and study C regarding the activation of the regions, that means each region got activated by the experimental condition (i.e., faces), and less active during control condition (i.e., houses). The models used however had only one region where the experimental input entered the system, and therefore the activation needed to be propagated throughout the system originating at this very region in a feed forward fashion. Depending on the exact model structure of all possible models in study B, this were many possible paths in total. However, the parameter estimates of the average models displayed in Figure 5 and Figure 6 of study B [Kessler et al., 2021b]

(appendix B) were dominated by the parameter estimates of the most likely model, as determined by Bayesian Model Selection (BMS) and calculated by Bayesian Model Averaging (BMA) and Hierarchical Linear Modeling (HLM).

Moreover, in study B we deliberately interpreted the resulting model in the context of predictive coding [Kessler et al., 2021b] (study B). Predictive coding is a common framework in neurosciences, aiming at explaining the neural computations on microscopic (e.g., [Bastos et al., 2012, Rao and Ballard, 1999]) or macroscopic scale (e.g., [Friston and Kiebel, 2009, Clark, 2013]). The coarse intuition is the following: in a hierarchical brain, the higher level brain region transmits predictions about the world, i.e., the sensory input, to the lower level brain region. The lower level brain region, in turn, computes a disparity between prediction and the actual sensory input (or rather information transmitted by the next lower brain region), and propagates this calculated prediction error to the higher level brain region [Clark, 2013]. In study B, we integrated the neural network parameters, i.e., positive forward connections and negative backwards connections, to the predictive coding theory, as has been done by other studies before us (e.g., [Chen et al., 2009, den Ouden et al., 2008]). Whereas the interpretations of the models' parameter estimates make sense in the predictive coding framework, the parameters are also predetermined by the model structure, according to the pattern outlined in study C. Therefore, the interpretation in any framework should rather be avoided when there can be shown, that the outcome is rather driven by the experiment itself.

The same would count if we would interpret the results of a study exploring the early visual hierarchy. Hypothetically, one could use the data and some of the models of study C to illustrate a model explaining the visual hierarchy in the framework of predictive coding. From a theoretical standpoint, this makes perfectly sense (e.g., [Rao and Ballard, 1999, Bastos et al., 2012]). However, by using the data from study C – which would be highly suitable to delineate the interregional relationships in early visual cortex using fMRI data and DCM as top level analysis – we would again pitfall into conclusions which we could not draw using the fMRI and DCM methodologies.

3.3 Putting this dissertation into the bigger picture

Finally, I want to reflect the contribution of the dissertation at hand and the included studies to the research landscape.

Study A was one of the first studies using relatively large cohort of participants to analyze network dynamics using DCM, especially in the context of psychiatric disorders, or as in our case, the sole presence of risks for psychiatric disorders. We deployed a compact neural network

model which was – as a theoretical framework – established in the field of MD, and transferred it to a new population of healthy participants with risk for MD. Despite its compactness, the model allowed for a sound interpretation on a neurocognitive level of its parameters. These result in turn allowed to explain the results of other research findings, such as findings reporting amygdala over-reactivity of participants at risk for depression [Dannlowski et al., 2012]. The results emphasize, that this over-reactivity results from a reduced inhibition by mPFC. Therefore, the study contributes to the explanation of the phenomenological findings. On the other hand, it further emphasizes the potential of deploying amygdala activation or its regulation as clinical biomarker, despite the general critics of biomarkers outlined within the dissertation at hand.

Study B however tackled a profound conceptual replication of a long established and widely used DCM model in the domain of face perception [Fairhall and Ishai, 2007]. Our study not only exterminated interpretational flaws of the original study, it further updated its' results in the light of new software updates, and highly increased its capability for generalization by analyzing several different datasets. The replication and revision of the original model within study B leads to a modified and revised version of the Haxby model, casted in a network model with quantified connections, i.e., a DCM model. We argue that this model is more suitable as working model for further studies in the domain of face perception and neural connectivity in the early face perception network.

Study C questioned the accuracy of interpretations upon a particular type of DCM models. Both by analyzing real fMRI data with DCM and by simulating DCM models we have illustrated, that a large proportion of connectivity parameters are predetermined by the modeling procedure rather than determined by the shape of the underlying neurocognitive processes which are of interest. I argue, that even if the investigated scenario was simple and straightforward, the problem is generalizable to many other studies. Therefore the problem is much wider than the scenario investigated in the manuscript. I am convinced, that among other, severe limitations of the method DCM, it is not suitable to answer most of the research questions it is applied for. I argue that the research community should overcome the fetish of using DCM as answer to all questions and emphasize, that they should open themselves for the much more diverse landscape of research methods available.

References

[Adolphs, 2002] Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opinion in neurobiology*, 12(2):169–177. Publisher: Elsevier.

[Allua and Thompson, 2009] Allua, S. and Thompson, C. B. (2009). Inferential statistics. *Air Medical Journal*, 28(4):168–171. Publisher: Elsevier.

[Almgren et al., 2018] Almgren, H., Van de Steen, F., Kühn, S., Razi, A., Friston, K., and Marinazzo, D. (2018). Variability and reliability of effective connectivity within the core default mode network: A multi-site longitudinal spectral DCM study. *Neuroimage*, 183:757–768. Publisher: Elsevier.

[Avidan and Behrmann, 2014] Avidan, G. and Behrmann, M. (2014). Impairment of the face processing network in congenital prosopagnosia. *Frontiers in bioscience (Elite edition)*, 6:236–257.

[Bastos et al., 2012] Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, 76(4):695–711.

[Bi and Fang, 2017] Bi, T. and Fang, F. (2017). Impaired Face Perception in Individuals with Autism Spectrum Disorder: Insights on Diagnosis and Treatment. *Neuroscience Bulletin*, 33(6):757–759.

[Brennan et al., 2019] Brennan, B. P., Wang, D., Li, M., Perriello, C., Ren, J., Elias, J. A., Van Kirk, N. P., Krompinger, J. W., Pope Jr, H. G., Haber, S. N., and others (2019). Use of an individual-level approach to identify cortical connectivity biomarkers in obsessive-compulsive disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(1):27–38. Publisher: Elsevier.

[Buxton et al., 2004] Buxton, R. B., Uludağ, K., Dubowitz, D. J., and Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, 23:S220–S233.

[Chen et al., 2009] Chen, C., Henson, R., Stephan, K., Kilner, J., and Friston, K. (2009). Forward and backward connections in the brain: A DCM study of functional asymmetries. *NeuroImage*, 45(2):453–462.

[Cheng et al., 2006] Cheng, D. T., Knight, D. C., Smith, C. N., and Helmstetter, F. J. (2006). Human amygdala activity during the expression of fear responses. *Behavioral neuroscience*, 120(6):1187. Publisher: American Psychological Association.

[Clark, 2013] Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204. Publisher: Cambridge University Press.

[Damoiseaux, 2012] Damoiseaux, J. S. (2012). Resting-state fMRI as a biomarker for alzheimer's disease? *Alzheimer's research & therapy*, 4(2):1–2. Publisher: Springer.

[Dannlowski et al., 2012] Dannlowski, U., Stuhrmann, A., Beutelmann, V., Zwanzger, P., Lenzen, T., Grotegerd, D., Domschke, K., Hohoff, C., Ohrmann, P., Bauer, J., and others (2012). Limbic scars: long-term consequences of childhood maltreatment revealed by functional and structural magnetic resonance imaging. *Biological psychiatry*, 71(4):286–293. Publisher: Elsevier.

[Darwin, 1859] Darwin, C. (1859). On the origin of species, 1859.

[Daunizeau et al., 2012] Daunizeau, J., Stephan, K. E., and Friston, K. J. (2012). Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? *Neuroimage*, 62(1):464–481. Publisher: Elsevier.

[Delgado et al., 2008] Delgado, M. R., Nearing, K. I., LeDoux, J. E., and Phelps, E. A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron*, 59(5):829–838. Publisher: Elsevier.

[den Ouden et al., 2008] den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., and Stephan, K. E. (2008). A dual role for prediction error in associative learning. *Cerebral Cortex*, 19(5):1175–1185.

[Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *arXiv:1801.07698 [cs]*. arXiv: 1801.07698.

[Duchaine and Yovel, 2015] Duchaine, B. and Yovel, G. (2015). A Revised Neural Framework for Face Processing. *Annual Review of Vision Science*, 1(1):393–416.

[Dvornek et al., 2017] Dvornek, N. C., Ventola, P., Pelphrey, K. A., and Duncan, J. S. (2017). Identifying autism from resting-state fMRI using long short-term memory networks. In *International workshop on machine learning in medical imaging*, pages 362–370. tex.organization: Springer.

[Elbich et al., 2019] Elbich, D. B., Molenaar, P. C., and Scherf, K. S. (2019). Evaluating the organizational structure and specificity of network topology within the face processing system. *Human Brain Mapping*, 40(9):2581–2595.

[Fairhall and Ishai, 2007] Fairhall, S. L. and Ishai, A. (2007). Effective Connectivity within the Distributed Cortical Network for Face Perception. *Cerebral Cortex*, 17(10):2400–2406.

[Fisher and Marshall, 2009] Fisher, M. J. and Marshall, A. P. (2009). Understanding descriptive statistics. *Australian critical care*, 22(2):93–97. Publisher: Elsevier.

[Friston, 2013] Friston, K. (2013). Active inference and free energy. *Behavioral and brain sciences*, 36(3):212. Publisher: Cambridge University Press.

[Friston et al., 2003] Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302.

[Friston and Kiebel, 2009] Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221. Publisher: The Royal Society London.

[Frässle et al., 2016a] Frässle, S., Krach, S., Paulus, F. M., and Jansen, A. (2016a). Handedness is related to neural mechanisms underlying hemispheric lateralization of face processing. *Scientific Reports*, 6(February):1–17. Publisher: Nature Publishing Group.

[Frässle et al., 2016b] Frässle, S., Paulus, F. M., Krach, S., Schweinberger, S. R., Stephan, K. E., and Jansen, A. (2016b). Mechanisms of hemispheric lateralization: Asymmetric interhemispheric recruitment in the face perception network. *NeuroImage*, 124:977–988. Publisher: Elsevier Inc.

[Frässle et al., 2015] Frässle, S., Stephan, K. E., Friston, K. J., Steup, M., Krach, S., Paulus, F. M., and Jansen, A. (2015). Test-retest reliability of dynamic causal modeling for fMRI. *Neuroimage*, 117:56–66. Publisher: Elsevier.

[Gobbini and Haxby, 2007] Gobbini, M. I. and Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, 45(1):32–41. Publisher: Elsevier.

[Hartmann et al., 2008] Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., and Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. *Marketing letters*, 19(3):287–304. Publisher: Springer.

[Haxby et al., 2000] Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233. Publisher: Elsevier.

[He et al., 2015] He, W., Garrido, M. I., Sowman, P. F., Brock, J., and Johnson, B. W. (2015). Development of effective connectivity in the core network for face perception. *Human Brain Mapping*, 36(6):2161–2173.

[Heinze et al., 2018] Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, 60(3):431–449. Publisher: Wiley Online Library.

[Heinzle and Stephan, 2018] Heinzle, J. and Stephan, K. E. (2018). Dynamic causal modeling and its application to psychiatric disorders. In *Computational psychiatry*, pages 117–144. Elsevier.

[Hildesheim et al., 2020] Hildesheim, F. E., Debus, I., Kessler, R., Thome, I., Zimmermann, K. M., Steinsträter, O., Sommer, J., Kamp-Becker, I., Stark, R., and Jansen, A. (2020). The trajectory of hemispheric lateralization in the core system of face processing: a cross-sectional functional magnetic resonance imaging pilot study. *Frontiers in Psychology*, 11. Publisher: Frontiers Media SA.

[Hindriks et al., 2016] Hindriks, R., Adhikari, M. H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N. K., and Deco, G. (2016). Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage*, 127:242–256. Publisher: Elsevier.

[Hohenfeld et al., 2018] Hohenfeld, C., Werner, C. J., and Reetz, K. (2018). Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker? *NeuroImage: Clinical*, 18:849–870. Publisher: Elsevier.

[Hong et al., 2019] Hong, Y.-W., Yoo, Y., Han, J., Wager, T. D., and Woo, C.-W. (2019). False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. *NeuroImage*, 195:384–395. Publisher: Elsevier.

[Horwitz et al., 2005] Horwitz, B., Warner, B., Fitzer, J., Tagamets, M.-A., Husain, F. T., and Long, T. W. (2005). Investigating the neural basis for functional and effective connectivity. Application to fMRI. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1093–1108. Publisher: The Royal Society London.

[Johnstone et al., 2007] Johnstone, T., Van Reekum, C. M., Urry, H. L., Kalin, N. H., and Davidson, R. J. (2007). Failure to regulate: counterproductive recruitment of top-down prefrontal-subcortical circuitry in major depression. *Journal of Neuroscience*, 27(33):8877–8884. Publisher: Soc Neuroscience.

[Kessler et al., 2016] Kessler, R., Bach, M., and Heinrich, S. P. (2016). Two-tactor vibrotactile navigation information for the blind: Directional resolution and intuitive interpretation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(3):279–286. Publisher: IEEE.

[Kessler et al., 2021a] Kessler, R., Henniger, O., and Busch, C. (2021a). Fingerprints, forever young? In *2020 25th international conference on pattern recognition (ICPR)*, pages 8647–8654. tex.organization: IEEE.

[Kessler and Jansen, 2022] Kessler, R. and Jansen, A. (2022). How function is bound by structure in models of effective connectivity. *preprint*.

[Kessler et al., 2021b] Kessler, R., Rusch, K. M., Wende, K. C., Schuster, V., and Jansen, A. (2021b). Revisiting the effective connectivity within the distributed cortical system for face perception. *NeuroImage: Reports*.

[Kessler et al., 2020] Kessler, R., Schmitt, S., Sauder, T., Stein, F., Yüksel, D., Grotegerd, D., Dannlowski, U., Hahn, T., Dempfle, A., Sommer, J., Steinsträter, O., Nenadic, I., Kircher, T., and Jansen, A. (2020). Long-Term Neuroanatomical Consequences of Childhood Maltreatment: Reduced Amygdala Inhibition by Medial Prefrontal Cortex. *Frontiers in Systems Neuroscience*, 14:28.

[Kircher et al., 2019] Kircher, T., Wöhr, M., Nenadic, I., Schwarting, R., Schratt, G., Alferink, J., Culmsee, C., Garn, H., Hahn, T., Müller-Myhsok, B., and others (2019). Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *European archives of psychiatry and clinical neuroscience*, 269(8):949–962. Publisher: Springer.

[Koshino et al., 2008] Koshino, H., Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., and Just, M. A. (2008). fMRI Investigation of Working Memory for Faces in Autism: Visual Coding and Underconnectivity with Frontal Areas. *Cerebral Cortex*, 18(2):289–300.

[Larose, 2015] Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

[Marrelec et al., 2008] Marrelec, G., Bellec, P., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Benali, H., and Doyon, J. (2008). Regions, systems, and the brain: hierarchical measures of functional integration in fMRI. *Medical image analysis*, 12(4):484–496. Publisher: Elsevier.

[Mayberg, 1997] Mayberg, H. S. (1997). Limbic-cortical dysregulation: A proposed model of depression. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 9(3):471–481. Place: US Publisher: American Psychiatric Assn.

[Mayberg et al., 1997] Mayberg, H. S., Brannan, S. K., Mahurin, R. K., Jerabek, P. A., Brickman, J. S., Tekell, J. L., Silva, J. A., McGinnis, S., Glass, T. G., Martin, C. C., and Fox, P. T. (1997). Cingulate function in depression: a potential predictor of treatment response. *NeuroReport*, 8(4):1057–1061.

[Meng et al., 2021] Meng, Q., Zhao, S., Huang, Z., and Zhou, F. (2021). MagFace: A Universal Representation for Face Recognition and Quality Assessment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 10.

[Milner, 2017] Milner, A. D. (2017). How do the two visual streams interact with each other? *Experimental brain research*, 235(5):1297–1308. Publisher: Springer.

[Miłkowski et al., 2018] Miłkowski, M., Hensel, W. M., and Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of computational neuroscience*, 45(3):163–172. Publisher: Springer.

[Nagy et al., 2012] Nagy, K., Greenlee, M. W., and Kovács, G. (2012). The Lateral Occipital Cortex in the Face Perception Network: An Effective Connectivity Study. *Frontiers in Psychology*, 3.

[Noble et al., 2021] Noble, S., Scheinost, D., and Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, 40:27–32. Publisher: Elsevier.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. In *Procedings of the British Machine Vision Conference 2015*, pages 41.1–41.12, Swansea. British Machine Vision Association.

[Plitt et al., 2015] Plitt, M., Barnes, K. A., and Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical*, 7:359–366. Publisher: Elsevier.

[Poldrack et al., 2017] Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115–126. Publisher: Nature Publishing Group.

[Poldrack and Gorgolewski, 2014] Poldrack, R. A. and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510–1517. Publisher: Nature Publishing Group.

[Rao and Ballard, 1999] Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group.

[Razi et al., 2017] Razi, A., Seghier, M. L., Zhou, Y., McColgan, P., Zeidman, P., Park, H.-J., Sporns, O., Rees, G., and Friston, K. J. (2017). Large-scale DCMs for resting-state fMRI. *Network Neuroscience*, 1(3):222–241.

[Redhead, 1980] Redhead, M. (1980). Models in Physics. *The British Journal for the Philosophy of Science*, 31(2):145–163.

[Rogers et al., 2007] Rogers, B. P., Morgan, V. L., Newton, A. T., and Gore, J. C. (2007). Assessing functional connectivity in the human brain by fMRI. *Magnetic resonance imaging*, 25(10):1347–1357. Publisher: Elsevier.

[Rossion, 2014] Rossion, B. (2014). Understanding face perception by means of prosopagnosia and neuroimaging. *Frontiers in bioscience*, 6(258):e307.

[Sahraei et al., 2021] Sahraei, I., Hildesheim, F. E., Thome, I., Kessler, R., Rusch, K. M., Sommer, J., Kamp-Becker, I., Stark, R., and Jansen, A. (2021). Developmental changes within the extended face processing network: A cross-sectional functional magnetic resonance imaging study. *Developmental Neurobiology*. Publisher: Wiley Online Library.

[Sato et al., 2017] Sato, W., Kochiyama, T., Uono, S., Matsuda, K., Usui, K., Usui, N., Inoue, Y., and Toichi, M. (2017). Bidirectional electric communication between the inferior occipital gyrus and the amygdala during face processing. *Human Brain Mapping*, 38(9):4511–4524.

[Saxena et al., 2013] Saxena, A., Verma, N., and Tripathi, D. K. C. (2013). A Review Study of Weather Forecasting Using Artificial Neural Network Approach. *International Journal of Engineering Research*, 2(11):7.

[Serengil and Ozpinar, 2020] Serengil, S. I. and Ozpinar, A. (2020). LightFace: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pages 23–27.

[Shrout and Rodgers, 2018] Shrout, P. E. and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69:487–510. Publisher: Annual Reviews.

[Soares et al., 2013] Soares, J., Marques, P., Alves, V., and Sousa, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in neuroscience*, 7:31. Publisher: Frontiers.

[Straube et al., 2018] Straube, B., Wroblewski, A., Jansen, A., and He, Y. (2018). The connectivity signature of co-speech gesture integration: the superior temporal sulcus modulates connectivity between areas related to visual gesture and auditory speech processing. *NeuroImage*, 181:539–549. Publisher: Elsevier.

[The FIL Methods Group, 2014] The FIL Methods Group (2014). SPM12 Release Notes.

[The FIL Methods Group, 2020] The FIL Methods Group (2020). SPM12 Update to Revision 7771.

[Thome et al., 2021] Thome, I., Hohmann, D. M., Zimmermann, K. M., Smith, M. L., Kessler, R., and Jansen, A. (2021). "I Spy with my Little Eye, Something that is a Face...": A Brain Network for Illusory Face Detection. *Cerebral Cortex*, page bhab199.

[Tononi et al., 1994] Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037. Publisher: National Acad Sciences.

[Ungerleider and Haxby, 1994] Ungerleider, L. G. and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165.

[Urry et al., 2006] Urry, H. L., Van Reekum, C. M., Johnstone, T., Kalin, N. H., Thurow, M. E., Schaefer, H. S., Jackson, C. A., Frye, C. J., Greischar, L. L., Alexander, A. L., and others (2006). Amygdala and ventromedial prefrontal cortex are inversely coupled during regulation of negative affect and predict the diurnal pattern of cortisol secretion among older adults. *Journal of Neuroscience*, 26(16):4415–4425. Publisher: Soc Neuroscience.

[Van Essen and Deyoe, 1995] Van Essen, D. C. and Deyoe, E. A. (1995). Concurrent processing in the primate visual cortex. In *The cognitive neurosciences*, pages 383–400. The MIT Press, Cambridge, MA, US.

[Wirtz et al., 2017] Wirtz, B. W., Göttel, V., and Daiser, P. (2017). SOCIAL NETWORKS: USAGE INTENSITY AND EFFECTS ON PERSONALIZED ADVERTISING. *Social Networks*, 18(2):21.

[Yamada et al., 2017] Yamada, T., Hashimoto, R.-i., Yahata, N., Ichikawa, N., Yoshihara, Y., Okamoto, Y., Kato, N., Takahashi, H., and Kawato, M. (2017). Resting-state functional connectivity-based biomarkers and functional MRI-based neurofeedback for psychiatric disorders: a challenge for developing theranostic biomarkers. *International Journal of Neuropsychopharmacology*, 20(10):769–781. Publisher: Oxford University Press US.

[Zhang et al., 2009] Zhang, H., Tian, J., Liu, J., Li, J., and Lee, K. (2009). Intrinsically organized network for face perception during the resting state. *Neuroscience letters*, 454(1):1–5. Publisher: Elsevier.

Appendix A

Study 1

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 03 June 2020
doi: 10.3389/fnsys.2020.00028

Long-Term Neuroanatomical Consequences of Childhood Maltreatment: Reduced Amygdala Inhibition by Medial Prefrontal Cortex

Roman Kessler[1,2], Simon Schmitt[1,2], Torsten Sauder[1,3], Frederike Stein[1,2], Dilara Yüksel[1,2], Dominik Grotegerd[4], Udo Dannlowski[4], Tim Hahn[4], Astrid Dempfle[5], Jens Sommer[2,6], Olaf Steinsträter[2,6], Igor Nenadic[1,2], Tilo Kircher[1,2] and Andreas Jansen[1,2,6]**

[1]Department of Psychiatry and Psychotherapy, Department of Medicine, University of Marburg, Marburg, Germany, [2]Centre for Mind, Brain and Behavior (CMBB), University of Marburg and Justus Liebig University Giessen, Marburg, Germany, [3]Department of Neurology, Bayreuth Clinic, Klinikum Bayreuth GmbH, Bayreuth, Germany, [4]Department of Psychiatry and Psychotherapy, University of Münster, Münster, Germany, [5]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany, [6]Core-Unit Brainimaging, Faculty of Medicine, University of Marburg, Marburg, Germany

OPEN ACCESS

Edited by:
Preston E. Garraghty,
Indiana University Bloomington,
United States

Reviewed by:
Amiel Rosenkranz,
Rosalind Franklin University of Medicine and Science, United States
Charles Jason Frazier,
University of Florida, United States

***Correspondence:**
Roman Kessler
kessler5@staff.uni-marburg.de
Andreas Jansen
jansena2@staff.uni-marburg.de

Received: 15 January 2020
Accepted: 30 April 2020
Published: 03 June 2020

Citation:
Kessler R, Schmitt S, Sauder T, Stein F, Yüksel D, Grotegerd D, Dannlowski U, Hahn T, Dempfle A, Sommer J, Steinsträter O, Nenadic I, Kircher T and Jansen A (2020) Long-Term Neuroanatomical Consequences of Childhood Maltreatment: Reduced Amygdala Inhibition by Medial Prefrontal Cortex. Front. Syst. Neurosci. 14:28. doi: 10.3389/fnsys.2020.00028

Similar to patients with Major depressive disorder (MDD), healthy subjects at risk for depression show hyperactivation of the amygdala as a response to negative emotional expressions. The medial prefrontal cortex is responsible for amygdala control. Analyzing a large cohort of healthy subjects, we aimed to delineate malfunction in amygdala regulation by the medial prefrontal cortex in subjects at increased risk for depression, i.e., with a family history of affective disorders or a personal history of childhood maltreatment. We included a total of 342 healthy subjects from the *FOR2107* cohort (www.for2107.de). An emotional face-matching task was used to identify the medial prefrontal cortex and right amygdala. Dynamic Causal Modeling (DCM) was conducted and neural coupling parameters were obtained for healthy controls with and without particular risk factors for depression. We assigned a *genetic risk* if subjects had a first-degree relative with an affective disorder and an *environmental risk* if subjects experienced childhood maltreatment. We then compared amygdala inhibition during emotion processing between groups. Amygdala inhibition by the medial prefrontal cortex was present in subjects without those two risk factors, as indicated by negative model parameter estimates. Having a *genetic risk* (i.e., a family history) did not result in changes in amygdala inhibition compared to *no risk* subjects. In contrast, childhood maltreatment as *environmental risk* has led to a significant reduction of amygdala inhibition by the medial prefrontal cortex. We propose a mechanistic explanation for the amygdala hyperactivity in subjects with particular risk for depression, in particular childhood maltreatment, caused by a malfunctioned amygdala

downregulation via the medial prefrontal cortex. As childhood maltreatment is a major *environmental risk* factor for depression, we emphasize the importance of this potential early biomarker.

Keywords: major depression, childhood maltreatment, fMRI, connectivity, emotion processing, dynamic causal modeling

INTRODUCTION

Major depressive disorder (MDD) is a common, chronic, costly, and debilitating disorder, affecting more than 300 million people worldwide (World Health Organization, 2017). The lifetime prevalence is in most countries in the range of 8–15% (Andrade et al., 2003; Kessler et al., 2003; Moffitt et al., 2010). MDD is caused by a complex interplay of genetic susceptibility and environmental factors, showing a heritability of ∼35% (Otte et al., 2016). Genetic risk factors are believed to decrease resilience to environmental stressors and make disorder onset more probable. Environmental risk factors include stressful life events and, in particular, childhood maltreatment (Nelson et al., 2017). Childhood maltreatment leads to an increased risk for the development of recurrent MDD and a weaker response to treatment (Nanni et al., 2011). Childhood maltreatment is also associated with persistent neurobiological alterations in brain areas involved in mood regulation (Nemeroff, 2016), strongly resembling changes reported for MDD patients (Dannlowski et al., 2012). A deeper understanding how specific risk factors for depression alter the functional neuroanatomy is important not only from a basic neuroscience perspective, but also to identify neurobiological changes that might be used as biomarkers to potentially provide preventive measures to on-risk individuals at early stages.

Functional magnetic resonance imaging (fMRI) yielded insights into the neuroanatomical correlates of MDD. One robustly replicated finding is the hyper-responsiveness of the amygdala during emotion processing (e.g., Abler et al., 2007; Dannlowski et al., 2007; Siegle et al., 2007; Suslow et al., 2010; for meta-analysis, see Fitzgerald et al., 2008; Palmer et al., 2015). Changes in activity in the amygdala and accompanying changes of activity in the medial prefrontal cortex (mPFC) have led to the formulation of the *limbic-cortical model of major depression* (Graham et al., 2013). This model, first outlined by Mayberg and colleagues (Mayberg, 1997), considers MDD as a network disorder. One key aspect is that hyper-activity in limbic areas is not adequately controlled by prefrontal regions, with an associated depressed mood (Mayberg et al., 1999). More importantly, amygdala hyperactivity is also present in subjects at *genetic* (Joormann et al., 2012) and *environmental risk* for depression, such as childhood maltreatment (Dannlowski et al., 2012). This hyperactivity is therefore not specific for MDD but may indicate a general vulnerability to mental disorders.

The *limbic-cortical model* offers a testable framework that can continuously integrate neuroimaging findings with complementary neuroanatomical, neurochemical, and electrophysiological studies in the investigation of the pathogenesis of depression. In the following, we deliberately used a simplified version of the *limbic-cortical model of Major Depression*. Our model focuses on the connection between mPFC and amygdala. This allows, on the one hand, to test whether the mPFC down-regulates the amygdala during emotion processing, and on the other hand whether this downregulation is modulated by *risk* factors.

The present study had two aims. First, we tested the *limbic-cortical model* by assessing the strength of amygdala inhibition exerted by the mPFC during an emotion processing task in a large group of healthy subjects. Second, we tested whether *genetic* (i.e., familial) and *environmental risk* factors modulate amygdala inhibition. We operationalized those risks via a family history of affective disorders and childhood maltreatment, respectively. We hypothesized that both risk factors decrease the inhibitory influence of the mPFC on the amygdala (Frodl et al., 2010; van Harmelen et al., 2010; Dannlowski et al., 2012; Joormann et al., 2012). To investigate the inhibition of mPFC to the amygdala, we applied Dynamic Causal Modeling (DCM, Friston et al., 2003) for fMRI. DCM allows for inferences about the directionality of brain connectivity and aims at inferring neural interactions from observational data. As DCM is strongly hypothesis-driven, it allows us to test hypotheses within the borders of a network model. Furthermore, previous studies have used such models to decipher disorder and medication effects on limbic-cortical circuitry (de Almeida et al., 2009; Sladky et al., 2015a; Sladky et al., 2015b).

MATERIALS AND METHODS

Subjects

Neuroimaging, clinical and neuropsychological data were obtained from the *FOR2107* cohort[1]. *FOR2107* is an ongoing multicenter study that aims to decipher the neurobiological foundations of affective disorders (Kircher et al., 2019). A detailed study description, including recruitment and assessment procedures, is given elsewhere (Vogelbacher et al., 2018; Kircher et al., 2019). Neuroimaging was performed at two centers, the University of Marburg and the University of Münster. The study was approved by the ethics committees of all participating institutions. Written informed consent was obtained from all subjects after a complete description of the study.

A first data freeze (v1.00) was conducted after 1,000 subjects (both patients and controls) were included in the study. For the selection of our final sample, we proceeded as follows: First, we decided to include only subjects measured at the University of Marburg to reduce variance related to different MR scanners (see Vogelbacher et al., 2018) for a comparison of data characteristics

[1] www.for2107.de

of both sites), leading to a sample size of 800 subjects. Second, we selected all subjects without any present or past psychiatric disorders, leading to a sample size of 352 subjects. Third, we excluded subjects with missing relevant imaging, clinical or neuropsychological data, leading to a final sample size of 342 (135 men, mean age 33.4 × 12.6 years, range 18–65 years). Subjects' characteristics (sex, age, verbal IQ, years of education, BDI, and HAMD scores) are summarized in **Supplementary Table S1**.

The subjects were classified according to their risk status as having a *genetic risk* (i.e., familial risk, $n = 63$), an *environmental risk* ($n = 44$), or *no risk* factors ($n = 247$). Twelve subjects had both a *genetic* and *environmental risk*. G*enetic risk* was assigned if at least one first degree relative was suffering from an affective disorder. We use the word "genetic risk" as a proxy for a familial risk, knowing that we are not examining concrete genotypes (see "Discussion" section). An *environmental risk* was assigned when two subscales of the Childhood Trauma Questionnaire (CTQ, Bernstein et al., 1997) exceeded a critical threshold (10 for emotional abuse, eight for physical abuse, eight for sexual abuse, 15 for emotional neglect, eight for physical neglect). We hypothesized that both risk factors independently decreased the inhibitory influence of the mPFC on the amygdala (Dannlowski et al., 2012; Joormann et al., 2012).

Experimental Design

All subjects were measured with a large neuroimaging battery assessing both brain function and structure. The study protocol is described in detail elsewhere (Kircher et al., 2019). In the present study, we analyzed the fMRI data from an emotional face-matching task (Hariri et al., 2002). It aims at activating face processing regions (e.g., fusiform face area, FFA), limbic regions (e.g., amygdala), and prefrontal regions. In the active condition, subjects viewed gray-scale images of fearful or angry faces (Ekman, 1992), in the control condition they viewed geometric shapes (circles and ellipsoids). In each trial, three items were presented. A target image was located at the top, two further images on the left and right side at the bottom, whereby one of these images was identical to the target image. The subject was instructed to indicate which of these two images was identical to the target image by pressing a corresponding button on an MRI-compatible response pad. The task was set up as block design, with six face and shape trials, respectively, per block. Blocks had a duration of 44 s (faces) and 32 s (shapes), respectively. Five shape blocks and four faces blocks were presented in an alternating order, starting with a shapes block. Blocks were separated by short inter-block-intervals. The paradigm lasted 6 min 14 s. Subjects of different subgroups performed similar with respect to hit rates and reaction times in this paradigm (**Supplementary Table S2**).

MRI Data Acquisition

MRI data were acquired at a 3T MRI scanner (Tim Trio, Siemens, Erlangen, Germany), located at the Department of Psychiatry, University of Marburg, using a 12-channel head matrix Rx-coil. A T2*-weighted echo-planar imaging (EPI) sequence sensitive to blood oxygen level-dependent (BOLD) contrast was used with the following parameters: TE = 30 ms, TR = 2,000 ms, FoV = 210 mm, matrix = 64 × 64, slice thickness = 3.8 mm, distance factor = 10%, phase encoding direction anterior >> posterior, flip angle = 90°, no parallel imaging, bandwidth 2,232 Hz/Px, ascending acquisition, axial acquisition, 33 slices, slice alignment parallel to AC-PC line tilted 20° in the dorsal direction. A quality assurance (QA) protocol was implemented to monitor scanner stability by regular phantom measurements, similar to the "Glover protocol" implemented in the FBIRN consortium (Friedman and Glover, 2006). The QA protocol is described in detail elsewhere (Vogelbacher et al., 2018).

MRI Data Analysis

Analysis of Brain Activity

fMRI data were analyzed with the software Statistical Parametric Mapping (SPM8, r2975)[2] based on MATLAB 7.9.0 R2009b using standard routines and templates. *Preprocessing*: the initial three functional images were excluded from further analysis to exclude T1 stabilization effects. Functional images were realigned onto the mean image of the series using a six parameter rigid-body transformation, spatially normalized into standard MNI space, and resampled to a resolution of 2 × 2 × 2 mm^3. Finally, the images were spatially smoothed using an 8 mm full-width-half-maximum (FWHM) isotropic Gaussian kernel. *Statistical analysis*: statistical analysis was performed using a general linear model (GLM) framework to create three-dimensional maps concerning the estimated regressor response amplitude. At the individual subject level, fMRI responses for both conditions (faces, shapes) were modeled in a block design using the canonical hemodynamic response function implemented in SPM8 convolved with a vector of onset times for the different stimulus blocks. High-pass filtering was applied with a cut-off frequency of 1/128 Hz to attenuate low-frequency components. Weighted beta-images and t-statistic images were created by contrasting the faces-condition (contrast weight 1) against the shapes-condition (contrast weight −1). At the group level, brain activation was assessed using a one-sample *t*-test for the contrast (faces > shapes).

Analysis of Brain Connectivity

Connectivity changes between the mPFC and the amygdala were assessed using Dynamic Causal Modeling [DCM, Friston et al., 2003), SPM12, r6685, DCM12, r6591]. DCM is a Bayesian framework for investigating the effective connectivity in a neural network based on neuroimaging data. In the present implementation, DCM describes the brain as a deterministic input-output system using a bilinear differential equation:

$$\frac{dz}{dt} = \left(A + \sum_{j=1}^{m} u_j B^j\right) z + Cu,$$

where z depicts the neuronal activities, u corresponds to the experimental input. A describes the endogenous (fixed or context-independent) connection strengths, B^j defines how the experimental manipulation u_j affects the connections among the network regions (modulatory connectivity), and C describes how

[2]http://www.fil.ion.ucl.ac.uk/spm/

the driving inputs directly influence the neuronal state of the network regions. The dynamics of the neuronal states in each region are translated into predictions of the measured BOLD signal by a hemodynamic forward model (Balloon-Windkessel model; (Buxton et al., 1998). Using a Variational Laplace approach with Gaussian assumptions on the prior and posterior distributions, the posterior densities of the model parameters (i.e., conditional mean and covariance) can be estimated by maximizing the negative free energy.

The starting point for a DCM analysis is the selection of a fixed set of regions, their possible connections, the driving inputs, and the modulatory inputs. Different models can be compared to identify which models best predict the data. DCM enables inferences at different levels, on the one hand, inference on model space, on the other hand, inference on parameter space of any given model. In the following, we will describe: (i) the extraction of time series in specific regions of interest (ROIs), the basis for estimating models; (ii) the model space definition; and (iii) the statistical inferences conducted with the model parameters of interest.

Time Series Extraction

fMRI time series were extracted from the mPFC and the right amygdala, analogous to the procedure described by Sladky et al. (2015b). First, we calculated the group activation pattern for the contrast (faces > shapes) using a one-sample *t*-test on the weighted beta-images of all subjects. We determined mPFC (MNI: 2, 46, −16) and right amygdala (MNI: 20, −6, −20) by selecting voxels that showed the most significant activations concerning the *t*-test in those areas. Subsequently, we identified the single subject peak voxel coordinates using a searchlight approach. For this, single subjects' activation maps were thresholded at $p < 0.99$, uncorrected, and the most strongly activated voxel was determined for each subject for the mPFC (within a search radius of 12 mm around group peak) and the right amygdala (within a search radius of 8 mm around group peak). See **Figure 1** for a graphical depiction of the localization of the regions. We selected such a liberal threshold to avoid dropping single subjects due to sub-threshold activation out of our DCM analysis. This would have created a selective sample with only "strongly"-activating subjects and generalizations would not have been possible.

At last, the first principal component of the time series in the mPFC and the right amygdala, respectively, was extracted including all voxels inside a radius of 4 mm around the subject-specific peak voxel.

Model Space Definition

Based on the *limbic-cortical model* of major depression (see "Introduction" section), we investigated the coupling between the mPFC and the right amygdala in a two-region model (**Figure 2**). We chose the right rather than bilateral amygdala because the most consistent findings regarding connectivity and risk factors focus on the right amygdala (e.g., Del-Ben et al., 2005; Anderson et al., 2007; Dalby et al., 2010; Windischberger et al., 2010; Dannlowski et al., 2012; Zhang et al., 2012; Sladky et al., 2015b). The choice of our model space was motivated by previous studies using a similar approach (de Almeida et al., 2009; Sladky et al., 2015a,b). We assumed reciprocal structural connectivity between both regions (Klingler and Gloor, 1960; Catani et al., 2002; Ghashghaei and Barbas, 2002). Therefore, the A-matrix was identical in all models. We created 12 different models, differing in their B- and C-matrices. The face blocks served as direct driving input (C-matrix) into the system, either via the mPFC, the amygdala, or both regions. These face regressors served also as modulatory input (B-matrix) on the connection from mPFC to the amygdala, on the connection from the amygdala to mPFC, on both connections or none connection.

Statistical Inference

We assessed the impact of risk status on amygdala inhibition. Our parameter of interest was, therefore, the modulatory B-matrix parameter of the fronto-amygdala connection. Bayesian Model Averaging (BMA) was conducted over the whole model space of a subject to compute a weighted average of each model parameter. The weighting was determined by the posterior probability of each model. This approach is considered as useful complementation to Bayesian Model Selection (BMS, Stephan et al., 2009) when none of the models tested outperformed all others (as was the case in the present study; see **Supplementary Table S3**).

A Bayesian estimation (*BEST*) procedure implemented in R (version 3.5.1; Kruschke, 2013) was used to calculate group differences. As input data, we used the posterior point estimates of all subjects' DCM parameters (i.e., modulatory fronto-amygdala connection) after subject-specific BMA. We used uninformative default priors. In a first step, a Bayesian MCMC process generated random draws from the posterior distribution

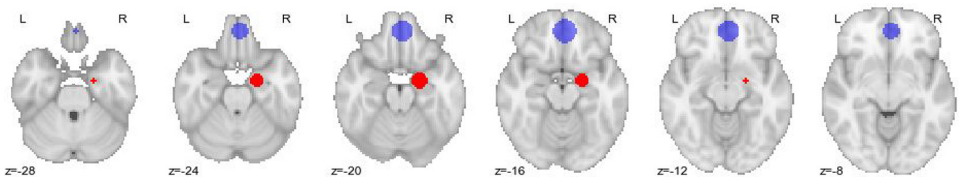


FIGURE 1 | Graphical depiction of the regions-of-interest for the Dynamic Causal Modeling (DCM) analysis. Medial prefrontal cortex (mPFC; blue; peak voxel at MNI coordinates 2, 46, −16) and right amygdala (red; peak voxel at MNI coordinates 20, −6, −20) are shown on axial slices. As the center of the sphere, we used the peak voxels of the group-level activation map. Numbers indicate the MNI z-coordinate.

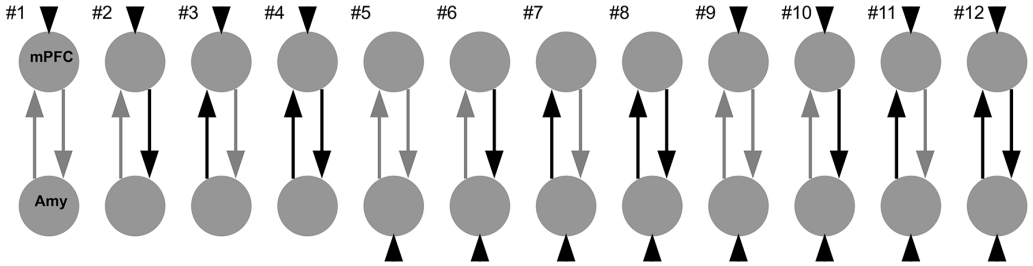


FIGURE 2 | Model space consisting of 12 different DCMs. Faces with emotional expressions served as input into the system (C-matrix, short arrows), either on the mPFC, the amygdala, or both regions. The two regions were always reciprocally connected (A-matrix, grey arrows). Faces either modulated one connection, both connections, or none of the connections (B-matrix, black arrows).

of group means and differences of means (500,000 samples each). We used the distribution of mean differences to infer the credibility of group differences. With this, posterior distributions for group mean comparisons were generated, similar to a *t*-test. But rather than *p*-values, Bayesian estimation provides probabilistic statements about values of interest (for more information, see Kruschke, 2010, 2013; Kruschke and Liddell, 2018). For example, we can state that with a probability of 95% the true value (i.e., mean connection strength) is higher for group A than for group B. Furthermore, an (e.g., 95%) highest density interval (HDI) marks a region of the credibility of parameter values. Obtaining a 95% HDI in the difference distribution that lies fully above or below zero, we can conclude a *credible difference*. Furthermore, we report effect sizes of the difference distribution between groups.

First, we computed three posterior distributions for the fronto-amygdala modulatory parameter, one for each group (*no risk*, *genetic risk*, and *environmental risk*). Subjects with *both risks* were included in both risk groups equally. We further computed the difference distributions between the respective risk groups and the *no-risk* group. We hypothesized that both risk factors independently decrease the inhibitory influence of the mPFC on the amygdala (Dannlowski et al., 2012; Joormann et al., 2012).

To account for confounding factors such as age, sex, and BDI score, we additionally conducted a multiple regression analysis (see **Supplementary Analysis**).

RESULTS

In the following, we will present subgroup-specific posterior parameter estimates after BMA and BEST. Our parameter of interest was the modulatory B-matrix parameter of the fronto-amygdala connection.

For participants without any of our examined risk factors, the coupling between mPFC and amygdala was negative, characterized by a mean parameter estimate of −0.366 (**Figure 3**, top left). Importantly, the 95% HDI interval was completely below zero, indicating a credible difference from zero. In this group, the mPFC therefore clearly exerted an inhibitory influence on amygdala activity during face processing.

For participants with a family history of affective disorders (i.e., *genetic risk*), the coupling strength was similar (mean parameter estimate −0.417, **Figure 3**, top center). The 95% HDI was completely located in the negative range, indicating that also in this group the mPFC exerted a clear inhibitory influence on amygdala activity during face processing. The differences of means between the *no risk* and the *genetic risk* group were 0.049 (**Figure 3**, top right). Since both the distribution of differences between means accumulated at zero and the 95% HDI intersected zero, there was no evidence for a different coupling strength between both groups. The effect size of the difference distribution was 0.03 (**Supplementary Figure S1**).

For participants with an *environmental risk* (i.e., childhood maltreatment) the parameter estimate of the fronto-amygdala coupling accumulated around zero (mean parameter estimate −0.035, **Figure 3**, bottom center). The difference of means between the *no risk* and the *environmental risk* group was −0.331 (**Figure 3**, bottom right). Importantly, the mean of the *no-risk* group was with a probability of 99.5% more negative than the mean of the *environmental risk* group. Similarly, the 95% HDI was completely in the negative range (**Figure 3**, bottom right). This showed that the inhibitory influence of the mPFC on amygdala activity during face processing was diminished in the *environmental risk* group compared to the *no-risk* group. The corresponding effect size was −0.46 (**Supplementary Figure S2**).

An additional multiple regression analysis confirmed those results (see **Supplementary Analysis**). In the regression, we found an overall significant amygdala inhibition in subjects at *no risk* ($p < 0.001$), and a significant reduction of this inhibition by childhood maltreatment as *environmental risk* ($p = 0.02$, see **Supplementary Analysis**). Neither effects of age, sex, or BDI have been found.

DISCUSSION

In the present study, we tested a neurobiological model for the inhibition of the amygdala response to emotional stimuli in a large sample of healthy subjects. In particular, we tested whether

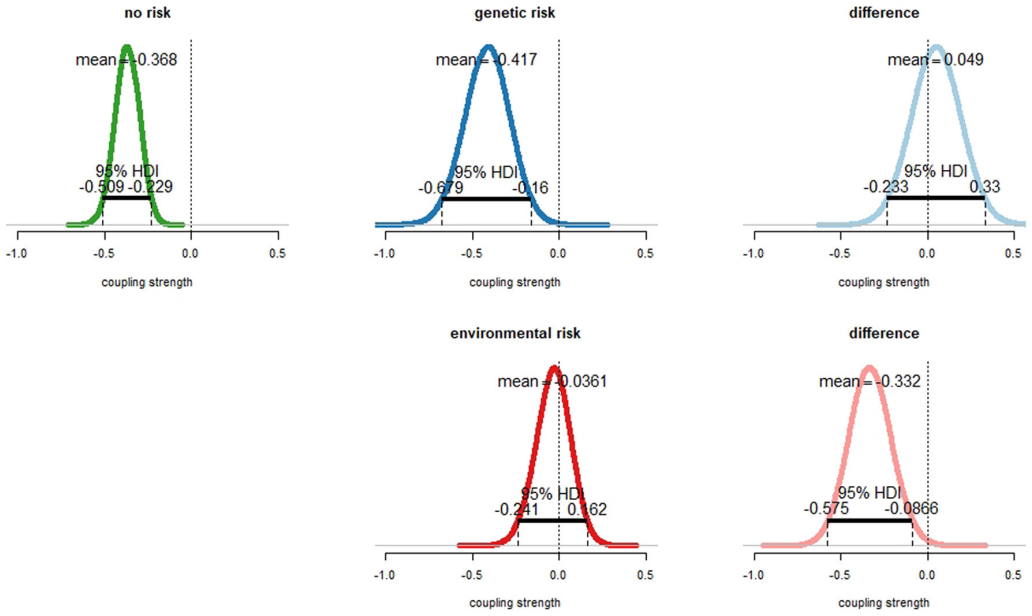


FIGURE 3 | Effect of emotional face processing on the fronto-amygdala connection in healthy participants with and without particular risks for depression. Displayed are sampling distributions for the mean for each subgroup (left and middle column) as obtained via Bayesian estimation ("BEST") and sampling distributions for the difference of group means (right column). Top row: subjects with *genetic risk* (family history) for major depressive disorder (MDD; top center) exhibited similar amygdala inhibition than those without risk (top left). 95% highest density intervals (HDIs) fell fully into the negative range. There was no credible difference between groups (top right). The 95% HDI well accumulated around zero. Bottom row: amygdala inhibition in healthy participants with *environmental risk* for depression (i.e., childhood maltreatment). Fronto-amygdala connectivity during emotional face processing was strongly diminished in healthy participants exhibiting an *environmental risk* with a probability of 99.5%, with the 95% HDI accumulating completely in the negative range.

this inhibition is modulated by *genetic* and *environmental risk* factors such as a family history of affective disorders and childhood maltreatment, respectively. Our results showed that amygdala inhibition by medial prefrontal cortex regions was strongly diminished in subjects who experienced childhood maltreatment, but not in subjects with genetic (i.e., familial) risk factors.

In the following, we will first discuss some background on the amygdala function and the necessity of amygdala inhibition. Then we will introduce the *limbic-cortical model* for depression. We will demonstrate how this network model explains amygdala hyperactivity in on-risk subjects, particularly those with past childhood maltreatment. Our results complement findings of amygdala hyperactivation in subjects with childhood maltreatment, and we propose a mechanistic model for how this hyperactivation may be caused.

The Amygdala Prefrontal Pathway in Emotion Regulation

Amygdala's activity is generally associated with the processing of emotionally salient stimuli, e.g., fearful facial expressions (Davis, 1992; Adolphs, 2002; Fitzgerald et al., 2006; Pessoa and Adolphs, 2011). The amygdala can respond to biologically relevant stimuli quickly (Méndez-Bértolo et al., 2016), allowing for a fast modulation of specialized cortical processing as well as behavioral, vegetative and endocrine reactions (LeDoux, 1998). Proper amygdala functioning was therefore of major advantage throughout vertebrate evolution. However, amygdala activity needs regulation, for instance after a stimulus has been evaluated as harmless. Such control is functionally related to the prefrontal cortex (Kim and Whalen, 2009; Agustín-Pavón et al., 2012), in particular to the orbitofrontal cortex (ORB), ventromedial prefrontal cortex (vmPFC) and anterior cingulate cortex (ACC; Mayberg, 1997; Mayberg et al., 1999; Etkin et al., 2011; Motzkin et al., 2015). Studies report overlapping functionalities of these three medial frontal regions (Etkin et al., 2011; Marusak et al., 2016). Lesions in medial prefrontal areas are associated with impaired down-regulation of fear and anxiety (Agustín-Pavón et al., 2012; Motzkin et al., 2015), implicating its role as an emotion control region. Additionally, metabolic alterations of those regulatory regions have been found for disorders such as MDD, which are accompanied by impaired emotion control abilities (Portella et al., 2011).

The amygdala has reciprocal anatomical connections to medial prefrontal regions, e.g., via the uncinate fasciculus (UF; Ebeling and von Cramon, 1992; Thiebaut de Schotten et al., 2012; Von Der Heide et al., 2013), which has been linked to inhibitory signaling from the mPFC to the amygdala (Kim and Whalen, 2009; Motzkin et al., 2015). Top-down signaling from mPFC towards the amygdala may be regarded

as *safety signaling*, with the mPFC supposedly calming down the amygdala (Harrison et al., 2017). Dysfunctions of amygdala downregulation in MDD have been associated with structural abnormalities in the UF, showing, for instance, an inverse relationship between UF volume and trait anxiety (Kim and Whalen, 2009; Baur et al., 2012) and weakened UF white matter structural integrity in MDD (de Kwaasteniet et al., 2013), particularly right-hemispheric (Dalby et al., 2010; Zhang et al., 2012). In an often-used analogy, the amygdala is regarded as a barking watchdog, while the mPFC is the dog's owner, evaluating the relevance of the barking dog and therefore differentiating between harmless and potentially hazardous events. In MDD however, the owner fails to regulate his or her watchdog as effectively as necessary, and the dog keeps alarming longer or louder as usual.

The Limbic-Cortical Model

A network model describing the interaction of mPFC and amygdala was first outlined by *Mayberg and colleagues* in the context of MDD (Mayberg, 1997). Its initial formulation proposed aberrant networking of a variety of cortical and subcortical areas. It proposes hypo-activity in the dorsal cortical and dorsal limbic areas and accompanying hyperactivity in ventral (para-) limbic areas in MDD. This activation pattern was supposed to flip with treatment (Mayberg, 1997), and medial prefrontal areas are to mediate between those major compartments (Mayberg, 1997). It's baseline activity has further been proposed as a biomarker for treatment success (Mayberg, 1997). Over the years the Mayberg model has been adapted and revised in very different fashions. For instance, the ventromedial prefrontal cortex (vmPFC) is often described as the regulatory region, inhibiting the amygdala in healthy subjects (e.g., Johnstone et al., 2007; Dutcher and Creswell, 2018) and lacking such inhibition in MDD (e.g., Johnstone et al., 2007). Other studies assigned such a regulatory function rather than the (ORB, Sladky et al., 2015b), but also (ACC, Johnstone et al., 2007; Etkin et al., 2011). In neuroimaging studies, regions such as vmPFC, ORB, and sometimes ACC are named in a very heterogeneous fashion, complicating the comparison of studies and findings. We derived both regions of interest from local peaks within the respective areas. Therefore, we named our prefrontal region, which encompassed both vmPFC and medial ORB, "mPFC" to keep it sufficiently general.

We applied the *limbic-cortical model* to data derived by healthy subjects with and without particular risk status for MDD rather than MDD patients themselves. We hypothesized that both of our examined risks may be associated with aberrant networking of this emotion regulation circuit, which then, in turn, may contribute to disorder onset. In the present study, we are not able to evaluate a causality chain due to the cross-sectional data used. However, we were able to evaluate the network model in healthy individuals without those two risk factors by showing, that there is indeed a down-regulation of the amygdala by mPFC during emotion processing, indicated by negative parameter estimates. We then examined how the network model behaves in subjects at-risk. In future studies, using longitudinal data that is currently collected in the *FOR2107* cohort, we will be able to further refine our findings by applying our models also to patient data.

The Impact of Risk Factors

MDD is most likely caused by a combination of some polygenetic predisposition and environmental factors. Showing high heritability, a family history of MDD may have a major impact on an individual, e.g., lowering resilience to adverse life events (Joormann et al., 2012). On the other hand, there are environmental factors, elevating the probability of clinical depression. One factor, leading to increased risk for depression, is childhood maltreatment (Kessler, 1997; Gilbert et al., 2009). Childhood maltreatment probably leads to psychological and biological vulnerabilities and higher sensibility to stressors (Kessler, 1997; Beck, 2008; Danese et al., 2008; Nanni et al., 2011), increasing the probability of disorder onset. Furthermore, MDD patients that experienced childhood maltreatment show lower treatment outcome (Hammen et al., 2000; Lanquillon et al., 2000; Nanni et al., 2011). On a neural system level, healthy subjects with a family history of MDD show amygdala hyperactivity in emotional tasks (Joormann et al., 2012). Similarly, healthy subjects with childhood trauma experiences show amygdala hyperactivity as a response to emotional faces, much like patients suffering from MDD (Dannlowski et al., 2012), accompanied with structural alterations in the prefrontal cortex (Frodl et al., 2010; van Harmelen et al., 2010; Dannlowski et al., 2012). Early life events, therefore, may establish long-lasting changes in emotional processing and associated unfavorable alterations in brain structure, function, and connectivity.

In our analysis, we tackled the question of amygdala inhibition by mPFC in healthy subjects at-risk. We operationalized a *genetic risk* by assigning it to a subject if a first-degree relative ever had a diagnosed affective disorder. We found no credible differences in amygdala inhibition between the *no risk* and the *genetic risk* group (**Figure 3**, bottom). This was contrary to our hypothesis as we expected a weaker inhibition in those subjects under *genetic* (i.e., familial) *risk*. Likewise, the *environmental risk* was operationalized via childhood maltreatment (see "Materials and Methods" section). We found that childhood maltreatment was associated with a strong reduction of amygdala inhibition (**Figure 3**, bottom). In the framework of our network model—an operationalization of the *limbic-cortical model*—we, therefore, provide a mechanistic explanation for the observed amygdala hyperactivity in healthy subjects with childhood trauma experiences (Dannlowski et al., 2012), namely a failure of amygdala regulation by prefrontal control regions.

Limitations

We acknowledge some limitations of our analyses. First, we used a simplified model including only two regions, covering only a small part of the brain regions associated with emotion processing. A widely distributed network of regions would form a better picture but comes with higher computational costs. Second, we identified one possible prefrontal region for our analysis, derived from our group activation data. Literature, however, reveals many different localizations of potential prefrontal control regions, with overlapping functionality but

variability in their designations (Etkin et al., 2011; Marusak et al., 2016). We refer to the Mayberg studies with our results, which can be seen as the basis for the *limbic-cortical model* of MDD (Graham et al., 2013). It provides us a suitable framework for our hypotheses. However, the prefrontal control region we used differed from the regions within the original studies. Additionally, we operationalized a *genetic risk* via a family history of affective disorders. However, this does not capture any concrete genotype. With this kind of operationalization, we may also not distinguish between a true *genetic risk* due to inheritance, and an *environmental* factor such as emotional neglect due to the indirect consequences of a parent's disorder. Therefore, our assigned *genetic risk* can be better understood as a familial risk, including both genetic and environmental factors.

Conclusion

In this article, we constructed and evaluated a model proposing that childhood maltreatment but not a family history of affective disorders are characterized by a reduced inhibition of the amygdala by mPFC. In the context of our model, we illustrate a potential mechanism for the frequently reported amygdala hyperactivation in MDD during emotion processing. More importantly, the model provides a mechanistic explanation for amygdala hyperactivation in healthy subjects with childhood trauma experiences. Model parameters such as this may constitute vulnerability markers for clinical symptoms in later life and maybe predictive for treatment success. Information of such model parameters may be used for early therapeutic intervention in at-risk individuals, to prevent disorder onset and poor treatment response in later life stages, when pathological connections are tightened and more difficult to treat.

DATA AVAILABILITY STATEMENT

All PIs take responsibility for the integrity of the respective study data and their components. All authors and co-authors had full access to all study data. Code for crucial analyses as well as statistical maps, subject-specific DCM models, and further data is available in a public repository of the first author (https://github.com/kesslerr/limbiccortical).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkomission FB 20 Medizin, Baldingerstraße, 35032 Marburg & Ethik-Kommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität Münster, Gartenstraße 210-214, 48147 Münster. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RK: conceptualization of analyses, conduction of analyses, interpretation of the data, drafting, and revision of the manuscript. SS and TS: data collection, revision of the manuscript, and interpretation of the data. FS and DY: data collection. DG: data collection and provided data infrastructure. UD: design of fMRI protocol and financially enabled the study. TH: financially enabled the study and interpretation of the data, and revision of the manuscript. AD: financially enabled the study. JS and OS: provided data infrastructure. IN: financially enabled the study, and interpretation of the data. TK: design of fMRI protocol, financially enabled the study, and revision of the manuscript. AJ: conceptualization of analyses, conduction of analyses, interpretation of the data, provided data infrastructure, design of fMRI protocol, drafting and revision of the manuscript, and financially enabled the study.

FUNDING

This work is part of the German multicenter consortium "Neurobiology of Affective Disorders. A translational perspective on brain structure and function," funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; Forschungsgruppe/Research Unit FOR2107). Principal investigators (PIs) with respective areas of responsibility in the FOR2107 consortium are: Work Package WP1, FOR2107/MACS cohort and brainimaging: TK (speaker FOR2107; DFG grant numbers KI 588/14-1, KI 588/14-2), UD (co-speaker FOR2107; DA 1151/5-1, DA 1151/5-2), Axel Krug (KR 3822/5-1, KR 3822/7-2), IN (NE 2254/1-2), Carsten Konrad (KO 4291/3-1). WP2, animal phenotyping: Markus Wöhr (WO 1732/4-1, WO 1732/4-2), Rainer Schwarting (SCHW 559/14-1, SCHW 559/14-2). WP3, miRNA: Gerhard Schratt (SCHR 1136/3-1, 1136/3-2). WP4, immunology, mitochondriae: Judith Alferink (AL 1145/5-2), Carsten Culmsee (CU 43/9-1, CU 43/9-2), Holger Garn (GA 545/5-1, GA 545/7-2). WP5, genetics: Marcella Rietschel (RI 908/11-1, RI 908/11-2), Markus Nöthen (NO 246/10-1, NO 246/10-2), Stephanie Witt (WI 3439/3-1, WI 3439/3-2). WP6, multi-method data analytics: AJ (JA 1890/7-1, JA 1890/7-2), Tim Hahn (HA 7070/2-2), Bertram Müller-Myhsok (MU1315/8-2), AD (DE 1614/3-1, DE 1614/3-2). CP1, biobank: Petra Pfefferle (PF 784/1-1, PF 784/1-2), Harald Renz (RE 737/20-1, 737/20-2). CP2, administration: TK (KI 588/15-1, KI 588/17-1), UD (DA 1151/6-1), Carsten Konrad (KO 4291/4-1).

ACKNOWLEDGMENTS

WP1: Henrike Bröhl, Katharina Brosch, Bruno Dietsche, Rozbeh Elahi, Jennifer Engelen, Sabine Fischer, Jessica Heinen, Svenja Klingel, Felicitas Meier, Tina Meller, Torsten Sauder, Simon Schmitt, Frederike Stein, Annette Tittmar, Dilara Yüksel (Dept. of Psychiatry, Marburg University); Mechthild Wallnig, Rita Werner (Core-Facility Brainimaging, Marburg University); Carmen Schade-Brittinger, Maik Hahmann (Coordinating Centre for Clinical Trials, Marburg). Michael Putzke (Psychiatric Hospital, Friedberg); Rolf Speier, Lutz Lenhard (Psychiatric Hospital, Haina); Birgit Köhnlein (Psychiatric Practice, Marburg); Peter Wulf, Jürgen Kleebach, Achim Becker (Psychiatric Hospital Hephata, Schwalmstadt-Treysa); Ruth Bär (Care facility Bischoff, Neukirchen); Matthias Müller; Michael Franz, Siegfried Scharmann, Anja Haag, Kristina Spenner, Ulrich

Ohlenschläger (Psychiatric Hospital Vitos, Marburg); Matthias Müller, Michael Franz, Bernd Kundermann (Psychiatric Hospital Vitos, Gießen); Christian Bürger, Katharina Dohm, Fanni Dzvonyar, Verena Enneking, Stella Fingas, Katharina Förster, Janik Goltermann, Dominik Grotegerd, Hannah Lemke, Susanne Meinert, Nils Opel, Ronny Redlich, Jonathan Repple, Kordula Vorspohl, Bettina Walden, Dario Zaremba (Dept. of Psychiatry, University of Münster); Harald Kugel, Jochen Bauer, Walter Heindel, Birgit Vahrenkamp (Dept. of Clinical Radiology, University of Münster); Gereon Heuft, Gudrun Schneider (Dept. of Psychosomatics and Psychotherapy, University of Münster); Thomas Reker (LWL-Hospital Münster); Gisela Bartling (IPP Münster); Ulrike Buhlmann (Dept. of Clinical Psychology, University of Münster); WP2: Marco Bartz, Miriam Becker, Christine Blöcher, Annuska Berz, Moria Braun, Ingmar Conell, Debora dalla Vecchia, Darius Dietrich, Ezgi Esen, Sophia Estel, Jens Hensen, Ruhkshona Kayumova; Theresa Kisko, Rebekka Obermeier, Anika Pütz, Nivethini Sangarapillai, Özge Sungur, Clara Raithel, Tobias Redecker, Vanessa Sandermann, Finnja Schramm, Linda Tempel, Natalie Vermehren, Jakob Vörckel, Stephan Weingarten, Maria Willadsen, Cüneyt Yildiz (Faculty of Psychology, Marburg University); WP4: Jana Freff, Silke Jörgens, Kathrin Schwarte (Dept. of Psychiatry, University of Münster); Susanne Michels, Goutham Ganjam, Katharina Elsässer (Faculty of Pharmacy, Marburg University); Felix Ruben Picard, Nicole Löwer, Thomas Ruppersberg (Institute of Laboratory Medicine and Pathobiochemistry, Marburg University); WP5: Helene Dukal, Christine Hohmeyer, Lennard Stütz, Viola Schwerdt, Fabian Streit, Josef Frank, Lea Sirignano (Dept. of Genetic Epidemiology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University); Stefanie Heilmann-Heimbach, Stefan Herms, Per Hoffmann (Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn); Andreas J. Forstner (Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn; Centre for Human Genetics, Marburg University); WP6: Anastasia Benedyk, Miriam Bopp, Roman Keßler, Maximilian Lückel, Verena Schuster, Christoph Vogelbacher (Dept. of Psychiatry, Marburg University); Jens Sommer, Olaf Steinsträter (Core-Facility Brainimaging, Marburg University); Thomas W.D. Möbius (Institute of Medical Informatics and Statistics, Kiel University); CP1: Julian Glandorf, Fabian Kormann, Arif Alkan, Fatana Wedi, Lea Henning, Alena Renker, Karina Schneider, Elisabeth Folwarczny, Dana Stenzel, Kai Wenk, Felix Picard, Alexandra Fischer, Sandra Blumenau, Beate Kleb, Doris Finholdt, Elisabeth Kinder, Tamara Wüst, Elvira Przypadlo, Corinna Brehm (Comprehensive Biomaterial Bank Marburg, Marburg University). The FOR2107 cohort project (WP1) was approved by the Ethics Committees of the Medical Faculties, University of Marburg (AZ: 07/14) and University of Münster (AZ: 2014-422-bS).

SUPPLEMENTARY MATERIAL

The corresponding single-subject parameter estimates are displayed in **Supplementary Figure S3**. The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnsys.2020.00028/full#supplementary-material.

REFERENCES

Abler, B., Erk, S., Herwig, U., and Walter, H. (2007). Anticipation of aversive stimuli activates extended amygdala in unipolar depression. *J. Psychiatr. Res.* 41, 511–522. doi: 10.1016/j.jpsychires.2006.07.020

Adolphs, R. (2002). Neural systems for recognizing emotion. *Curr. Opin. Neurobiol.* 12, 169–177. doi: 10.1016/s0959-4388(02)00301-x

Agustín-Pavón, C., Braesicke, K., Shiba, Y., Santangelo, A. M., Mikheenko, Y., Cockroft, G., et al. (2012). Lesions of ventrolateral prefrontal or anterior orbitofrontal cortex in primates heighten negative emotion. *Biol. Psychiatry* 72, 266–272. doi: 10.1016/j.biopsych.2012.03.007

Anderson, I. M., Del-Ben, C. M., McKie, S., Richardson, P., Williams, S. R., Elliott, R., et al. (2007). Citalopram modulation of neuronal responses to aversive face emotions: a functional MRI study. *Neuroreport* 18, 1351–1355. doi: 10.1097/wnr.0b013e3282742115

Andrade, L., Caraveo-Anduaga, J. J., Berglund, P., Bijl, R. V., De Graaf, R., Vollebergh, W., et al. (2003). The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *Int. J. Methods Psychiatr. Res.* 12, 3–21. doi: 10.1002/mpr.138

Baur, V., Hänggi, J., and Jäncke, L. (2012). Volumetric associations between uncinate fasciculus, amygdala and trait anxiety. *BMC Neurosci.* 13:4. doi: 10.1186/1471-2202-13-4

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *Am. J. Psychiatry* 165, 969–977. doi: 10.1176/appi.ajp.2008.08050721

Bernstein, D. P., Ahluvalia, T., Pogge, D., and Handelsman, L. (1997). Validity of the childhood trauma questionnaire in an adolescent psychiatric population. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 340–348. doi: 10.1097/00004583-199703000-00012

Buxton, R. B., Wong, E. C., and Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the ballon model. *Magn. Reson. Med.* 39, 855–864. doi: 10.1002/mrm.1910390602

Catani, M., Howard, R. J., Pajevic, S., and Jones, D. K. (2002). Virtual *in vivo* interactive dissection of white matter fasciculi in the human brain. *NeuroImage* 17, 77–94. doi: 10.1006/nimg.2002.1136

Dalby, R. B., Frandsen, J., Chakravarty, M. M., Ahdidan, J., Sørensen, L., Rosenberg, R., et al. (2010). Depression severity is correlated to the integrity of white matter fiber tracts in late-onset major depression. *Psychiatry Res.* 184, 38–48. doi: 10.1016/j.pscychresns.2010.06.008

Danese, A., Moffitt, T., Pariante, C., Antony, A., Poulton, R., and Caspi, A. (2008). Elevated inflammation levels in depressed adults with a history of childhood maltreatment. *Arch. Gen. Psychiatry* 65, 409–415. doi: 10.1001/archpsyc.65.4.409

Dannlowski, U., Ohrmann, P., Bauer, J., Kugel, H., Arolt, V., Heindel, W., et al. (2007). Amygdala reactivity to masked negative faces is associated with automatic judgmental bias in major depression: a 3 T fMRI study. *J. Psychiatry Neurosci.* 32, 423–429. doi: 10.30965/9783657764082_082

Dannlowski, U., Stuhrmann, A., Beutelmann, V., Zwanzger, P., Lenzen, T., Grotegerd, D., et al. (2012). Limbic scars: long-term consequences of childhood maltreatment revealed by functional and structural magnetic resonance imaging. *Biol. Psychiatry* 71, 286–293. doi: 10.1016/j.biopsych.2011.10.021

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annu. Rev. Neurosci.* 15, 353–375. doi: 10.1146/annurev.ne.15.030192.002033

de Almeida, J. R. C., Versace, A., Mechelli, A., Hassel, S., Quevedo, K., Kupfer, D. J., et al. (2009). Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biol. Psychiatry* 66, 451–459. doi: 10.1016/j.biopsych.2009.03.024

de Kwaasteniet, B., Ruhe, E., Caan, M., Rive, M., Olabarriaga, S., Groefsema, M., et al. (2013). Relation between structural and functional connectivity in major depressive disorder. *Biol. Psychiatry* 74, 40–47. doi: 10.1016/j.biopsych.2012.12.024

Del-Ben, C. M., Deakin, J. F. W., Mckie, S., Delvai, N. A., Williams, S. R., Elliott, R., et al. (2005). The effect of citalopram pretreatment on neuronal responses to neuropsychological tasks in normal volunteers: an fMRI study. *Neuropsychopharmacology* 30, 1724–1734. doi: 10.1038/sj.npp.1300728

Dutcher, J. M., and Creswell, J. D. (2018). Behavioral interventions in health neuroscience. *Ann. N Y Acad. Sci.* 1428, 51–70. doi: 10.1111/nyas.13913

Ebeling, U., and von Cramon, D. (1992). Topography of the uncinate fascicle and adjacent temporal fiber tracts. *Acta Neurochir.* 115, 143–148. doi: 10.1007/bf01406373

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200.

Etkin, A., Egner, T., and Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends Cogn. Sci.* 15, 85–93. doi: 10.1016/j.tics.2010.11.004

Fitzgerald, D. A., Angstadt, M., Jelsone, L. M., Nathan, P. J., and Phan, K. L. (2006). Beyond threat: amygdala reactivity across multiple expressions of facial affect. *NeuroImage* 30, 1441–1448. doi: 10.1016/j.neuroimage.2005.11.003

Fitzgerald, P. B., Laird, A. R., Maller, J., and Daskalakis, Z. J. (2008). A meta-analytic study of changes in brain activation in depression. *Hum. Brain Mapp.* 29, 683–695. doi: 10.1002/hbm.20426

Friedman, L., and Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *J. Magn. Reson. Imaging* 23, 827–839. doi: 10.1002/jmri.20583

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage* 19, 1273–1302. doi: 10.1016/s1053-8119(03)00202-7

Frodl, T., Reinhold, E., Koutsouleris, N., Reiser, M., and Meisenzahl, E. M. (2010). Interaction of childhood stress with hippocampus and prefrontal cortex volume reduction in major depression. *J. Psychiatr. Res.* 44, 799–807. doi: 10.1016/j.jpsychires.2010.01.006

Ghashghaei, H. T., and Barbas, H. (2002). Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience* 115, 1261–1279. doi: 10.1016/s0306-4522(02)00446-3

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., and Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *Lancet* 373, 68–81. doi: 10.1016/s0140-6736(08)61706-7

Graham, J., Salimi-Khorshidi, G., Hagan, C., Walsh, N., Goodyer, I., Lennox, B., et al. (2013). Meta-analytic evidence for neuroimaging models of depression: state or trait? *J. Affect. Disord.* 151, 423–431. doi: 10.1016/j.jad.2013.07.002

Hammen, C., Henry, R., and Daley, S. E. (2000). Depression and sensitization to stressors among young women as a function of childhood adversity. *J. Consult. Clin. Psychol.* 68, 782–787. doi: 10.1037/0022-006x.68.5.782

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., and Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *NeuroImage* 17, 317–323. doi: 10.1006/nimg.2002.1179

Harrison, B. J., Fullana, M. A., Via, E., Soriano-Mas, C., Vervliet, B., Martínez-Zalacaín, I., et al. (2017). Human ventromedial prefrontal cortex and the positive affective processing of safety signals. *NeuroImage* 152, 12–18. doi: 10.1016/j.neuroimage.2017.02.080

Johnstone, T., van Reekum, C. M., Urry, H. L., Kalin, N. H., and Davidson, R. J. (2007). Failure to regulate: counterproductive recruitment of top-down prefrontal-subcortical circuitry in major depression. *J. Neurosci.* 27, 8877–8884. doi: 10.1523/jneurosci.2063-07.2007

Joormann, J., Cooney, R. E., Henry, M. L., and Gotlib, I. H. (2012). Neural correlates of automatic mood regulation in girls at high risk for depression. *J. Abnorm. Psychol.* 121, 61–72. doi: 10.1037/a0025294

Kessler, R. C. (1997). The effects of stressful life events on depression. *Ann. Rev. Psychol.* 48, 191–214. doi: 10.1146/annurev.psych.48.1.191

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major. *JAMA* 289, 3095–3105. doi: 10.1001/jama.289.23.3095

Kim, M. J., and Whalen, P. J. (2009). The structural integrity of an amygdala-prefrontal pathway predicts trait anxiety. *J. Neurosci.* 29, 11614–11618. doi: 10.1523/JNEUROSCI.2335-09.2009

Kircher, T., Wöhr, M., Nenadic, I., Schwarting, R., Schratt, G., and Alferink, J. (2019). Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *Eur. Arch. Psychiatry Clin. Neurosci.* 269, 949–962. doi: 10.1007/s00406-018-0943-x

Klingler, J., and Gloor, P. (1960). The connections of the amygdala and of the anterior temporal cortex in the human brain. *J. Comp. Neurol.* 115, 333–369. doi: 10.1002/cne.901150305

Kruschke, J. K. (2010). What to believe: bayesian methods for data analysis. *Trends Cogn. Sci.* 14, 293–300. doi: 10.1016/j.tics.2010.05.001

Kruschke, J. K. (2013). Bayesian estimation supersedes the T test. *J. Exp. Psychol. Gen.* 142, 573–603. doi: 10.1037/a0029146

Kruschke, J. K., and Liddell, T. M. (2018). The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. Bull. Rev.* 25, 178–206. doi: 10.3758/s13423-016-1221-4

Lanquillon, S., Krieg, J.-C., Bening-Abu-Shach, U., and Vedder, H. (2000). Cytokine production and treatment response in major depressive disorder. *Neuropsychopharmacology* 22, 370–379. doi: 10.1016/s0893-133x(99)00134-7

LeDoux, J. (1998). Fear and the brain: where have we been and where are we going? *Biol. Psychiatry* 44, 1229–1238. doi: 10.1016/s0006-3223(98)00282-0

Marusak, H. A., Thomason, M. E., Peters, C., Zundel, C., Elrahal, F., and Rabinak, C. A. (2016). You say 'prefrontal cortex' and I say 'anterior cingulate': meta-analysis of spatial overlap in amygdala-to-prefrontal connectivity and internalizing symptomology. *Transl. Psychiatry* 6:e944. doi: 10.1038/tp.2016.218

Mayberg, H. S. (1997). Limbic-cortical dysregulation: a proposed model of depression. *J. Neuropsychiatry Clin. Neurosci.* 9, 471–481. doi: 10.1176/jnp.9.3.471

Mayberg, H. S., Liotti, M., Brannan, S. K., McGinnis, S., Mahurin, R. K., Jerabek, P. A., et al. (1999). Reciprocal limbic-cortical function and negative mood: converging PET findings in depression and normal sadness. *Am. J. Psychiatry* 156, 675–682. doi: 10.1176/ajp.156.5.675

Méndez-Bértolo, C., Moratti, S., Toledano, R., Lopez-Sosa, F., Martínez-Alvarez, R., Mah, Y. H., et al. (2016). A fast pathway for fear in human amygdala. *Nat. Neurosci.* 19, 1041–1049. doi: 10.1038/nn.4324

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., et al. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychol. Med.* 40, 899–909. doi: 10.1017/S0033291709991036

Motzkin, J. C., Philippi, C. L., Wolf, R. C., Baskaya, M. K., and Koenigs, M. (2015). Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. *Biol. Psychiatry* 77, 276–284. doi: 10.1016/j.biopsych.2014.02.014

Nanni, V., Uher, R., and Danese, A. (2011). Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. *Am. J. Psychiatry* 169, 141–151. doi: 10.1176/appi.ajp.2011.11020335

Nelson, J., Klumparendt, A., Doebler, P., and Ehring, T. (2017). Childhood maltreatment and characteristics of adult depression: meta-analysis. *Br. J. Psychiatry* 210, 96–104. doi: 10.1192/bjp.bp.115.180752

Nemeroff, C. B. (2016). Paradise lost: the neurobiological and clinical consequences of child abuse and neglect. *Neuron* 89, 892–909. doi: 10.1016/j.neuron.2016.01.019

Otte, C., Gold, S. M., Penninx, B. W., Pariante, C. M., Etkin, A., Fava, M., et al. (2016). Major depressive disorder. *Nat. Rev. Dis. Primers* 2:16065. doi: 10.1038/nrdp.2016.65

Palmer, S. M., Crewther, S. G., and Carey, L. M. (2015). A meta-analysis of changes in brain activity in clinical depression. *Front. Hum. Neurosci.* 8:1045. doi: 10.3389/fnhum.2014.01045

Pessoa, L., and Adolphs, R. (2011). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* 11, 773–783. doi: 10.1038/nrn2920

Portella, M. J., de Diego-Adeliño, J., Gómez-Ansón, B., Morgan-Ferrando, R., Vives, Y., Puigdemont, D., et al. (2011). Ventromedial prefrontal spectroscopic abnormalities over the course of depression: a comparison among first episode, remitted recurrent and chronic patients. *J. Psychiatr. Res.* 45, 427–434. doi: 10.1016/j.jpsychires.2010.08.010

Siegle, G. J., Thompson, W., Carter, C. S., Steinhauer, S. R., and Thase, M. E. (2007). Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: related and independent

features. *Biol. Psychiatry* 61, 198–209. doi: 10.1016/j.biopsych.2006.05.048

Sladky, R., Höflich, A., Küblböck, M., Kraus, C., Baldinger, P., Moser, E., et al. (2015a). Disrupted effective connectivity between the amygdala and orbitofrontal cortex in social anxiety disorder during emotion discrimination revealed by dynamic causal modeling for fMRI. *Cereb. Cortex* 25, 895–903. doi: 10.1093/cercor/bht279

Sladky, R., Spies, M., Hoffmann, A., Kranz, G., Hummer, A., Gryglewski, G., et al. (2015b). (S)-citalopram influences amygdala modulation in healthy subjects: a randomized placebo-controlled double-blind fMRI study using dynamic causal modeling. *NeuroImage* 108, 243–250. doi: 10.1016/j.neuroimage.2014.12.044

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage* 46, 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025

Suslow, T., Konrad, C., Kugel, H., Rumstadt, D., Zwitserlood, P., Schöning, S., et al. (2010). Automatic mood-congruent amygdala responses to masked facial expressions in major depression. *Biol. Psychiatry* 67, 155–160. doi: 10.1016/j.biopsych.2009.07.023

Thiebaut de Schotten, M., Dell'Acqua, F., Valabregue, R., and Catani, M. (2012). Monkey to human comparative anatomy of the frontal lobe association tracts. *Cortex* 48, 82–96. doi: 10.1016/j.cortex.2011.10.001

van Harmelen, A. L., van Tol, M. J., van der Wee, N. J. A., Veltman, D. J., Aleman, A., Spinhoven, P., et al. (2010). Reduced medial prefrontal cortex volume in adults reporting childhood emotional maltreatment. *Biol. Psychiatry* 68, 832–838. doi: 10.1016/j.biopsych.2010.06.011

Vogelbacher, C., Möbius, T. W. D., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., et al. (2018). The marburg-münster affective disorders cohort study (MACS): a quality assurance protocol for MR neuroimaging data. *NeuroImage* 172, 450–460. doi: 10.1016/j.neuroimage.2018.01.079

Von Der Heide, R. J., Skipper, L. M., Klobusicky, E., and Olson, I. R. (2013). Dissecting the uncinate fasciculus: disorders, controversies and a hypothesis. *Brain* 136, 1692–1707. doi: 10.1093/brain/awt094

Windischberger, C., Lanzenberger, R., Holik, A., Spindelegger, C., Stein, P., Moser, U., et al. (2010). Area-specific modulation of neural activation comparing escitalopram and citalopram revealed by pharmaco-fMRI: a randomized cross-over study. *NeuroImage* 49, 1161–1170. doi: 10.1016/j.neuroimage.2009.10.013

World Health Organization. (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Geneva: World Health Organization.

Zhang, A., Leow, A., Ajilore, O., Lamar, M., Yang, S., Joseph, J., et al. (2012). Quantitative tract-specific measures of uncinate and cingulum in major depression using diffusion tensor imaging. *Neuropsychopharmacology* 37, 959–967. doi: 10.1038/npp.2011.279

Conflict of Interest: TK received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, neuraxpharm. Markus Wöhr is scientific advisor of Avisoft Bioacoustics.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers

Supplementary Material

1 Supplementary Methods

1.1 Subjects Characteristics

Table S1: Subjects' characteristics. N: Total number of subjects in the subgroup. HAMD is the Hamilton depression score (Hamilton, 1960). BDI is Becks Depression Inventory score (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961).

group	no risk	genetic risk	environmental risk	both risks
N	247	51	32	12
sex (m/f)	99/148	20/31	13/19	3/9
age	33 ± 13	32 ± 13	40 ± 12	38 ± 13
verbal IQ	115 ± 14	114 ± 15	117 ± 16	113 ± 11
years education	14 ± 2	14 ± 3	14 ± 3	14 ± 3
BDI	4.2 ± 3.9	4.9 ± 5.6	5.4 ± 5.5	10 ± 7.7
HAMD	1.0 ± 1.5	1.6 ± 1.9	2.2 ± 2.5	4.0 ± 6.1

1.2 Task Performance

Table S2: Hit rates and reaction times (RT) for the different subgroups.

	mean hit rate faces	sd hit rate faces	mean hit rate shapes	sd hit rate shapes	mean RT faces	sd RT faces	mean RT shapes	sd RT shapes
no risk	0.91	0.25	0.89	0.25	1223.05	377.33	1041.71	295.47
genetic risk	0.90	0.26	0.88	0.26	1171.51	399.16	1005.99	306.22
environmental risk	0.91	0.25	0.90	0.24	1206.05	387.28	1030.48	304.58
both risks	0.91	0.25	0.90	0.24	1206.05	387.28	1030.48	304.58

1.3 Bayesian Model Selection

Table S3: Model exceedance probabilities and posterior probabilities. Bayesian Model Selection was conducted in each subgroup separately. Model exceedance probabilities describe the probability that one model is more like than all competing models generating the data. Posterior model probabilities determine the relative probability of a model and further its contribution of a model to the respective subgroups average model (BMA). For a graphical description of the twelve models see Figure 2. The subgroup-specific model probabilities as displayed here were not of particular importance for the group comparisons using BEST. Instead, we used subject-specific model probabilities to calculate each subject's individual average model.

posterior probabilities	**model**											
	1	2	3	4	5	6	7	8	9	10	11	12
no risk	0.01	0.01	0.01	0.30	0.02	0.02	0.04	0.36	0.00	0.04	0.15	0.04
genetic risk	0.02	0.03	0.02	0.16	0.02	0.04	0.03	0.30	0.02	0.10	0.18	0.09
environmental risk	0.04	0.11	0.08	0.18	0.03	0.05	0.04	0.29	0.03	0.07	0.04	0.04
both risks	0.04	0.04	0.05	0.22	0.05	0.06	0.09	0.20	0.05	0.09	0.05	0.07

exceedance probabilities	**model**											
	1	2	3	4	5	6	7	8	9	10	11	12
no risk	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.81	0.00	0.00	0.00	0.00
genetic risk	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.83	0.00	0.01	0.09	0.00
environmental risk	0.00	0.04	0.02	0.19	0.00	0.01	0.00	0.73	0.00	0.00	0.00	0.00
both risks	0.01	0.01	0.01	0.44	0.01	0.02	0.06	0.34	0.01	0.05	0.01	0.03

2 Supplementary Results

2.1 Bayesian Estimation

For both comparisons (either no risk vs. genetic risk or no risk vs. environmental risk) we used Bayesian Estimation ("BEST", (Kruschke, 2013)) to calculate e.g. differences of means. Furthermore, effect sizes, and differences in variances etc. are displayed in the following graphics.

2.1.1 No risk vs. genetic risk

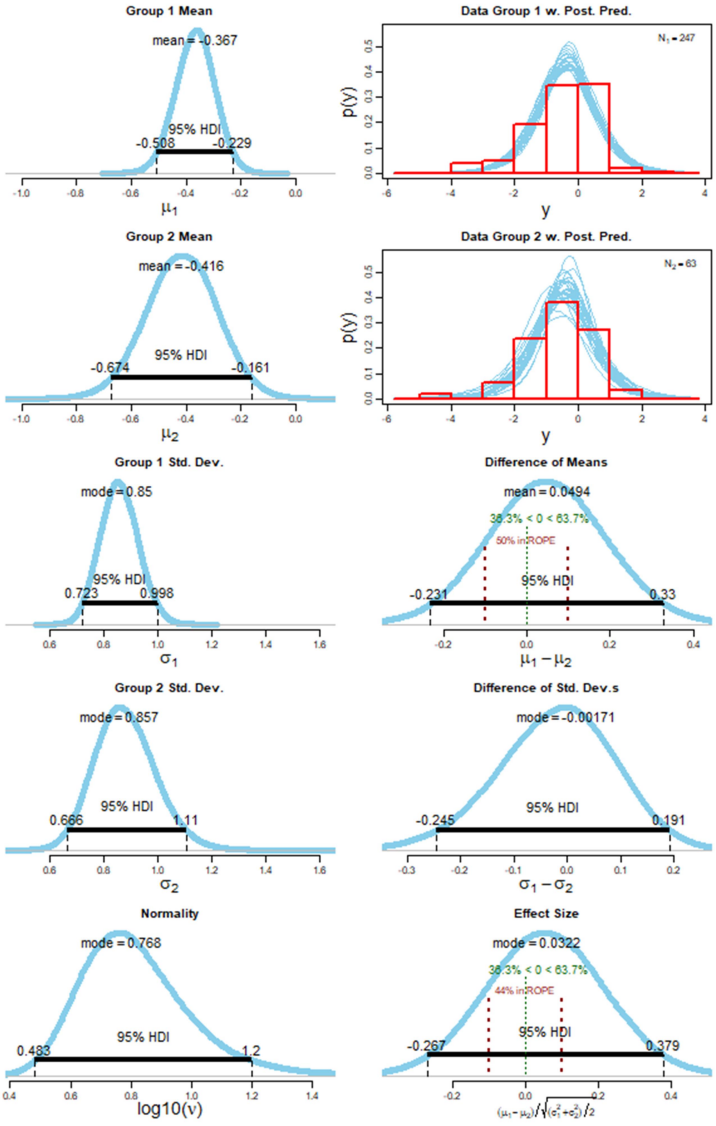


Figure S1: Detailed Bayesian Estimation ('BEST') results for the comparisons of subjects without risk (group 1) and subjects with genetic risk (group 2). Displayed are estimated group means and standard deviations, differences of means and standard deviations, degree of normality and effect sizes (Kruschke, 2013) alongside with highest density intervals (HDI).

2.1.2 No risk vs. environmental risk

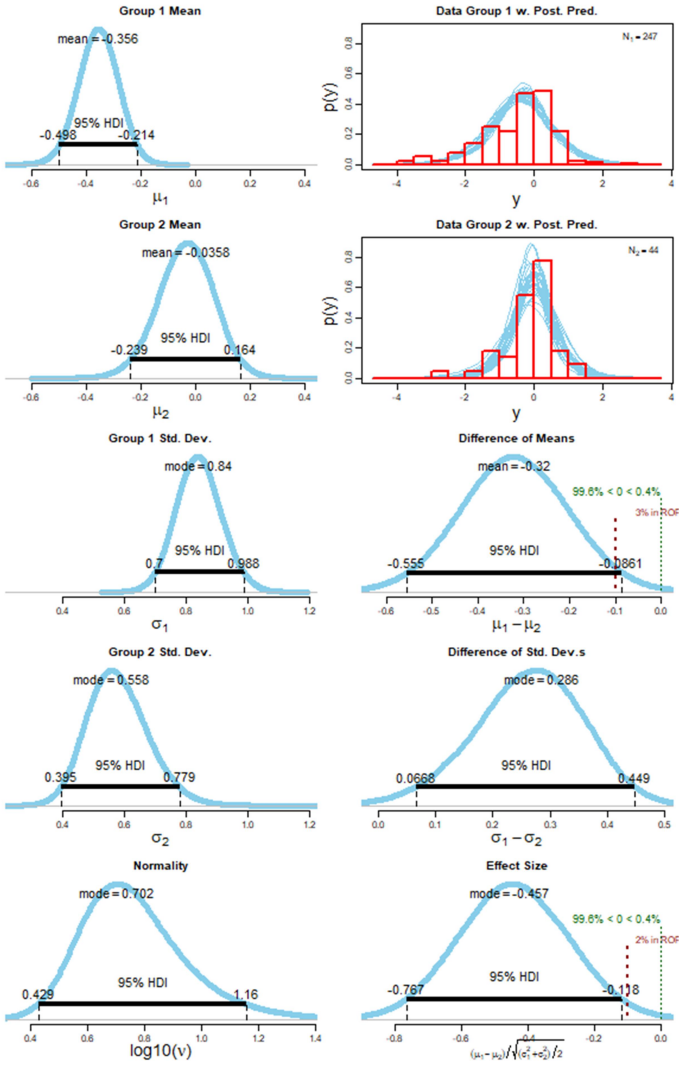


Figure S2: Detailed Bayesian Estimation ('BEST') results for the comparisons of subjects without risk (group 1) and subjects with environmental risk (group 2). Displayed are estimated group means and standard deviations, differences of means and standard deviations, degree of normality and effect sizes (Kruschke, 2013) alongside with highest density intervals (HDI).

3 Supplementary Analysis: Linear regression

Groups were not explicitly matched for sex, age, verbal IQ or years of education. Instead, we selected as many subjects as possible to increase the sensitivity of our analyses. To strengthen our results and to correct for these covariates, we additionally conducted a multiple linear regression analysis. We included covariates as age, sex, and BDI into our analysis as covariates.

3.1 Methods

Group differences between the modulatory B-matrix parameter of the fronto-amygdala connection were assessed using a linear regression model using R (version 3.5.1). We constructed the linear regression model of amygdala inhibition as a function of risk factors. As risk factors, we categorically modeled a family history (*genetic risk*) and childhood maltreatment (*environmental risk*). We further included major possible confounding variables as age (mean-centered), sex (female = 0, male = 1), and BDI in our model. The intercept of the model represents the average amygdala inhibition (if negative) of (female) subjects with none of the above risks and a BDI of zero. Other regression parameters give insight about the significance and strength of risk factors and confounding variables. If slopes are positive, the modeled factor decreases amygdala inhibition by mPFC. If negative, those factors increase amygdala inhibition. We further used a step-wise backward regression, by iteratively pruning the model parameters with highest p-value (until all $p<0.05$) to get a simpler model with only significant predictor variables.

3.2 Results

A multiple linear regression model was constructed to predict the influence of mPFC onto the amygdala during emotion processing. Predictors were *genetic risk* (family history, categorical), *environmental risk* (childhood maltreatment, categorical), age (mean-centered), sex (female = 0), and BDI. After stepwise backward regression using ordinary least squares (OLS), a significant regression equation was found ($p=0.02$) with an adjusted R-squared of 0.013. The predicted influence of mPFC onto the amygdala was -0.503 (intercept, 95% CI -0.618 & -0.387, $p<0.001$), with an increase of the parameter estimate by childhood maltreatment (binary) of 0.381 (95% CI 0.06 & 0.703, $p=0.02$). Therefore, the parameter estimate for healthy subjects at no risk was negative (i.e. amygdala inhibition). With childhood maltreatment, this inhibition was decreased significantly (i.e. reduced inhibition).

Using a full model without backward regression, we obtain similar results. Only the intercept (amygdala inhibition at no risk) and childhood maltreatment as predictor became significant. A family history, as well as covariates such as age, sex, and BDI, remained non-significant. The corresponding single-subject parameter estimates are displayed in Figure S3.

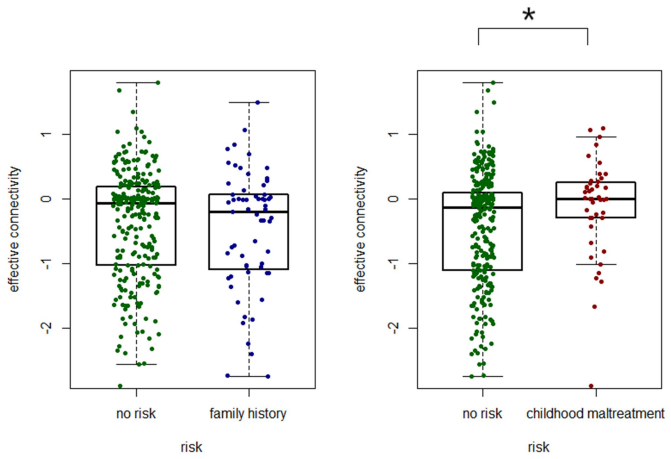


Figure S3: Single subject parameter estimates of the fronto-amygdala connection and result of multiple linear regression analysis.

4 Supplementary Analysis: Full-factorial analysis

To evaluate activation differences between groups, we constructed a 2x2 full-factorial analysis on the second level fMRI data. The first factor was *environmental risk* (childhood maltreatment), and the second factor *genetic risk* (family history).). We then analyzed "main effect of environmental risk" and "main effect of genetic risk" as F-contrasts. We masked the resulting contrast with the very same masks for right amygdala ("rAmy") and medial prefrontal cortex ("mPFC"), which we used to extract the time series for our subjects. No voxels exceeded significance in the F-contrasts at thresholds of $p<0.01$ (uncorrected for multiple comparisons). We therefore conclude no meaningful activation differences between no-risk and at-risk groups on activation level within our sample.

5 Supplementary References

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). Inventory for Measuring Depression. *Archives of General Psychiatry*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Hamilton, M. (1960). Scale for depression. *J. Neurol. Neurosurg. Psychiat.*, (23), 56–62.

Kruschke, J. K. (2013). Bayesian estimation supersedes the T test. *Journal of Experimental Psychology: General*, *142*(2), 573–588. https://doi.org/10.1037/a0029177

Appendix B

Study 2

Neuroimage: Reports 1 (2021) 100045

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports

Revisiting the effective connectivity within the distributed cortical network for face perception

Roman Kessler [a,b,c,d,*], Kristin M. Rusch [a,b,e], Kim C. Wende [a,f], Verena Schuster [a,b,g], Andreas Jansen [a,b,h,**]

[a] Laboratory for Multimodal Neuroimaging, Department of Psychiatry and Psychotherapy, University of Marburg, Germany
[b] Center for Mind, Brain and Behavior, University of Marburg and University of Giessen, Germany
[c] Norwegian University of Science and Technology (NTNU), Gjøvik, Norway
[d] University of Applied Sciences, Darmstadt, Germany
[e] Department of Neurology and Neurorehabilitation, Hospital zum Heiligen Geist, Academic Teaching Hospital of the Heinrich-Heine-University Düsseldorf, Kempen, Germany
[f] Institute of Medical Psychology and Medical Sociology, Faculty of Medicine, University of Freiburg, Germany
[g] The Neuro (Montréal Neurological Institute-Hospital), McGill University, Montréal, Canada
[h] Core-Unit Brainimaging, Faculty of Medicine, University of Marburg, Germany

ARTICLE INFO

Keywords:
Conceptual replication
Dynamic causal modeling
Emotion processing
Face perception
fMRI

ABSTRACT

The classical core system of face perception consists of the occipital face area (OFA), fusiform face area (FFA), and posterior superior temporal sulcus (STS). The functional interaction within this network, more specifically the effective connectivity, was first described by Fairhall and Ishai (2007) using functional magnetic resonance imaging and dynamic causal modeling. They proposed that the core system is hierarchically organized; information is processed in a parallel and predominantly feed-forward fashion from the OFA to downstream regions such as the FFA and STS, with no lateral connectivity, i.e., no connectivity between the two downstream regions (FFA and STS). Over a decade later, we conducted a conceptual replication of their model using four different functional magnetic resonance imaging data sets. The effective connectivity within the core system was assessed with contemporary versions of dynamic causal modeling.

The resulting model of the core system of face perception was densely interconnected. Using hierarchical linear modeling, we identified several significant forward, backward, and lateral connections in the core system of face perception across the data sets. Face perception increased the forward connectivity from the OFA to the FFA and OFA to the STS and increased the inhibitory backward connectivity from the FFA to the OFA, as well as the lateral connectivity between the FFA and STS. Emotion perception increased forward connectivity between the OFA and STS and decreased the lateral connectivity between the FFA and STS. Face familiarity did not significantly alter these connections.

Our results revise the 2007 model of the core system of face perception. We discuss the potential meaning of the resulting model parameters and propose that our revised model is a suitable working model for further studies assessing the functional interaction within the core system of face perception. Our work further emphasizes the general importance of conceptual replications.

1. Introduction

Face processing is mediated by a widely distributed neural network. This network is often divided into a core system and an extended system (Haxby model, Haxby et al., 2000). The core system is involved in the processing of basic information about faces. It consists of several bilateral brain regions in the occipitotemporal cortex; specifically, the occipital face area (OFA) in the inferior occipital gyrus, the fusiform face area (FFA) in the middle fusiform gyrus, and an area in the posterior superior temporal sulcus (STS). According to the Haxby model, the OFA

* Corresponding author. Department of Psychiatry and Psychotherapy, University of Marburg, Rudolf-Bultmann-Str. 8, 35039, Marburg Germany.
** Corresponding author. Laboratory for Multimodal Neuroimaging, Department of Psychiatry and Psychotherapy, University of Marburg, Germany.
E-mail addresses: kessler5@staff.uni-marburg.de (R. Kessler), jansena2@staff.uni-marburg.de (A. Jansen).

https://doi.org/10.1016/j.ynirp.2021.100045
Received 25 June 2021; Accepted 6 August 2021
Available online 17 August 2021

is responsible for the early processing of physical features of face stimuli and sends its output to both the FFA and STS. The FFA is associated with the representation of invariant aspects of the face (e.g., face identity), while the STS processes changeable aspects of facial expression (e.g., lip movements and the direction of eye-gaze). Beyond the core system, there are several additional regions that contribute to face perception, such as the amygdala, insula, inferior frontal gyrus, and orbitofrontal cortex (Gobbini and Haxby, 2007). This extended system tends to be task-specific and comes into play if additional information is extracted from faces, such as attractiveness or biographical information.

Only a few studies have previously investigated the assumptions made by the Haxby model with respect to the interplay between the face-sensitive regions. An understanding of the interaction between these areas, however, is crucial for unraveling how the human brain processes faces and might also provide new insights into the pathophysiology of disorders where face perception is impaired (e.g., prosopagnosia). One method to test the interactions between brain regions is dynamic causal modeling (DCM) (Friston et al., 2003). DCM is used to test hypotheses about the neural network structure. It estimates the directed coupling between brain areas (effective connectivity) and the changes in coupling caused by experimental manipulations (i.e., context). A few different neural network models (i.e., DCMs) have been developed for the face perception system. These DCMs assessed the neural dynamics within the core system of face perception, the interaction between the core system and extended system, and the effects of 'emotions' and 'fame' on the effective connectivity within those networks (e.g., Dima et al., 2011; Fairhall and Ishai, 2007; Furl, 2015; Furl et al., 2015; Herrington et al., 2011). They were typically limited to one hemisphere but have recently been expanded by bilateral DCMs, including interactions between both hemispheres (Frässle et al, 2016b, 2016c, 2016b).

The first study that used DCM to describe the interactions between face-sensitive brain regions was published almost 15 years ago. In this study, Fairhall and Ishai (2007) tested DCMs, which were built based on the Haxby model and described the interactions within the core system. Not only did they show how the OFA, FFA, and STS interact during face processing, but they also assessed how factors like emotional valence and the fame of faces influenced those interactions (see Fig. 1 for a graphical depiction of their model). Their study's main results were:

i. The OFA propagates face-specific content simultaneously to the FFA and STS in a feed-forward fashion.
ii. Backward connections to the OFA and collateral connections between the FFA and STS were not present in their proposed model.
iii. Emotional valence enhanced connectivity from the OFA to the FFA.
iv. 'Fame' enhanced connectivity from the OFA to the FFA.

The Fairhall and Ishai (2007) study has been highly influential and widely cited since it was published, and it further makes far-reaching claims on how the brain regions in the core system interact during face processing and how these interactions are modulated. Various studies investigating the connectivity within the core system of face perception have been published, building upon these results (Elbich et al., 2019; Frässle et al., 2016a, 2016b, 2016c, 2016b; He et al., 2015; Lohse et al., 2016; Nagy et al., 2012; Nguyen et al., 2014; Sato et al., 2017). However, the study's results have never been formally replicated, neither in different samples nor with different strategies of analysis. Therefore, the aim of the present study was to investigate the degree to which we can reproduce the results from the study by Fairhall and Ishai (2007).

Concerns about the reproducibility of neuroimaging findings have been steadily raised in recent years since numerous studies have shown that the results of previous experiments could not be replicated (Gorgolewski and Poldrack, 2016). One reason for this is that results obtained can be highly dependent on the tools being used as well as differences in the experimental setup, pipeline, or statistical methods (Bedenbender et al., 2011; Botvinik-Nezer et al., 2020; Weissenbacher et al., 2009). Reproducibility can be assessed with different approaches (Diener and Biswas-Diener, 2016). An exact replication can be performed by attempting to repeat the original study in the best way possible, i.e., using identical paradigms and tools for analysis. However, there is also the option of a conceptual replication, wherein the researchers are not interested in simply repeating the steps of the original study in an exact and sequential manner. Instead, they may be interested in answering the very same research question as that in the original study by using tools that are similarly suitable to find those answers. Both types of replications are important since they each give us new but complementary information. While exact replications strengthen our belief in the findings from the original research, conceptual replications can strengthen the theoretical idea behind the findings. In other words, conceptual replications offer insights into how generalizable the findings are. In the present study, we aimed to conduct a conceptual replication of the core results of the study by Fairhall and Ishai (2007). We were not interested in whether these results could be reproduced in one specific sample, with one specific face perception task, and with one specific analysis pipeline. Rather, we aimed to assess whether the findings can be replicated over several samples, different implementations of face processing tasks, and different analysis methods.

In summary, we investigated face-specific interactions in the core system, i.e., between the OFA, FFA, and STS in the right hemisphere. We expected similar results to those from the study conducted by Fairhall & Ishai (Fairhall and Ishai, 2007), namely an increase in forward connectivity from the OFA to the FFA and from the OFA to the STS. Furthermore, we investigated the influence of 'emotion' and 'fame' on the strength of the connections between brain regions of the core system.

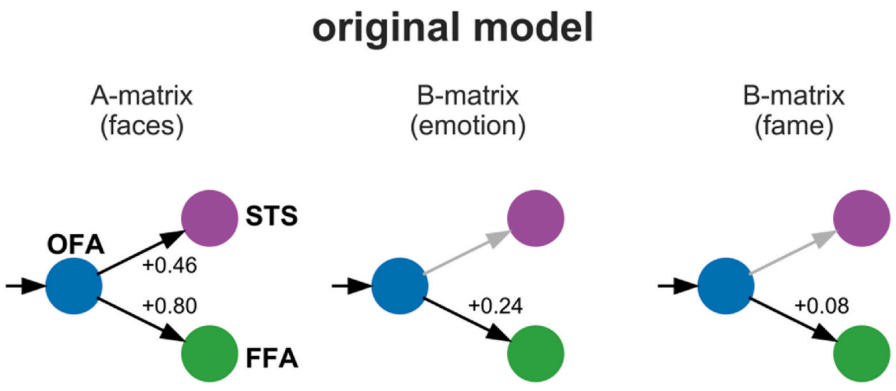


Fig. 1. Dynamic causal model of the interactions within the core system of face perception by Fairhall and Ishai (2007). Driving input (faces) enters the OFA, which propagates the information in a parallel manner toward the FFA and STS. Assumptions about the effect of faces were drawn from the A-matrix. Assumptions about the effects of emotion and fame were drawn from separate B-matrices (see Material and Methods for further information on the terminology of DCM, see Discussion for further information on the modeling strategy).

We expected 'emotion' and 'fame' to increase the connectivity from the OFA to the FFA, similar to what was observed in the original study. By analyzing four different samples, three of which were acquired in our laboratory, we aimed to increase the generalizability of our results. In all four samples, the processing of faces was investigated; emotion processing was additionally assessed in two samples. The fourth sample, which was retrieved from an open neuroimaging platform (Wakeman and Henson, 2015), allowed us to investigate the effect of 'fame'. All the studies from which the samples were obtained used distinct paradigms and participants. To combine our results with these studies, we applied a hierarchical linear modeling (HLM) approach.

2. Material and Methods

2.1. Study samples

We analyzed four samples of healthy participants (referred to in the manuscript as data sets A–D, studies A–D, paradigms A–D, samples A–D, and so forth). Three of these data sets (A, B, and C) were retrieved from ongoing (and therefore yet unpublished) studies in our lab (Laboratory for Multimodal Neuroimaging, Department of Psychiatry, University of Marburg, Germany). Studies A and B were originally planned to investigate the changes in connectivity in the face perception network associated with facial emotion processing. Study C initially assessed the impact of female hormones on brain structure and function, and, on the face-processing network. Written informed consent was provided by all the participants. The fourth data set (data set D) was obtained from the OpenNeuro project (openneuro.org), accession number ds000117, Wakeman and Henson (2015). In Table 1, we summarized detailed information on the participants' characteristics of studies A-D and the original study (Fairhall and Ishai, 2007) (henceforth referred to as 'study FI'). Participants in samples A and B were investigated just once. Participants in sample C (all female) were investigated twice, with 1–25 weeks between sessions (mean = 7 weeks); one measurement took place during the mid-luteal phase and the other during the early follicular phase of the menstrual cycle. Participants in sample D were measured ten times each with the same face perception paradigm. Sessions in which there was no significant activation in each of the three regions of the core system were excluded from DCM analyses (see chapter 2.4.2 [i]). Therefore, we report both the total number of participants and sessions for each study (rows 1 and 2) and the participants and sessions included in the final analyses (remaining rows). In study FI, participants were measured five times with four different paradigms (Fairhall and Ishai, 2007).

Table 1
Sample characteristics of each study.

Sample	*A*	*B*	*C*	*D*	*FI*
total number of participants	25	31	20	16	n.a.
total number of sessions per participant	1	1	2	10	5
number of participants included	23	27	17	16	10
number of sessions included per participant	1	1	1–2	5-10 (Md: 8)	5
number of males	11	13	0	9	5
number of females	12	14	17	7	5
age (years)	24 (Md)	24 (Md)	24 (Md)	n.a.	25 (mean)
minimum age (years)	21	20	20	23*	n.a.
maximum age (years)	29	29	28	37*	n.a.

Abbreviations: Md, median; n.a., Information not available. *Study D: The age range of all 19 participants was included in the online repository. However, at the time of our analysis, the data for only 16 participants were accessible. Therefore, the age range in study D might differ from that shown above.

2.2. Functional paradigms

The paradigms of all the data sets were constructed to tackle questions related to face perception. Participants viewed face stimuli in the experimental conditions and non-face stimuli (i.e., houses or phase-scrambled images) in the control conditions. Studies A–D used photographs of faces, while study FI used different face stimuli (line drawings of faces, famous faces, emotional faces, and unfamiliar faces). Paradigms A–C were set up in a block design similar to study FI, whereas paradigm D used an event-related design. All the paradigms included a simple task, such as a one-back task (paradigm A–C) or symmetry rating task (paradigm D). Study FI did not include any accompanying task. We have presented paradigm A in Fig. 2. More detailed descriptions of paradigms A–C can be found in the supplementary methods. A description of paradigm D is found in the study by Wakeman and Henson (2015). Paradigm FI is described in study FI (Fairhall and Ishai, 2007).

One crucial difference between the paradigms was the inclusion of emotional or famous faces. Paradigm A used four different emotional expressions, namely neutral, fearful, happy, and angry, separated into different blocks (Fig. 2). Paradigm B used two different emotional expressions, neutral and fearful. Paradigms C and D used neutral faces instead of particularly emotional expressions. Paradigms A–C used non-famous faces, whereas paradigm D used non-famous as well as famous faces.

2.3. Data acquisition

High resolution structural images and blood oxygen level-dependent functional images of all four data sets were acquired using Siemens 3T TIM TRIO MR scanners (Siemens, Erlangen, Germany). Study FI used a 3T Philips Intera scanner (Philips, Hamburg, Germany). All measurement volumes for the functional image acquisitions covered the entire core system of face perception. Information on the properties of the scanning sequences is detailed in the supplementary methods.

2.4. Data analysis

2.4.1. Preprocessing and statistical analysis of brain activity

Analyses of the magnetic resonance imaging (MRI) data sets A, B, and C were conducted using Statistical Parametric Mapping 12 (SPM12) (https://www.fil.ion.ucl.ac.uk/spm/). Data set D was processed using FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), and study FI used SPM5. In all the data sets, preprocessing included motion correction, spatial normalization (except study A), and spatial smoothing. Statistical analyses were conducted using a general linear model. We modeled 'faces' as regressors of interest, and the control condition (e.g., houses or scrambled faces) were modeled as separate regressors, following which we contrasted the 'face' vs. control conditions. Here, we did not differentiate between neutral, emotional, or famous faces. Similarly, we did not differentiate between the different control conditions. Nuisance regressors included the six realignment parameters. A more detailed description of the specific analysis pipelines can be found in the supplementary methods. Notably, we could have used the raw data of each data set and implemented an identical preprocessing pipeline for all paradigms. However, to increase the generalizability, we decided to use the preprocessed data sets. All the procedures that were implemented by the respective authors represent valid implementations of preprocessing pipelines.

2.4.2. Dynamic causal modeling

The connectivity pattern of the core system of face perception was assessed with DCM (Friston et al., 2003; Zeidman et al., 2019a). DCM is a framework to disentangle effective connectivity in neuroimaging data. In its original formulation, it models the brain as a deterministic input-output system using the following differential equation:

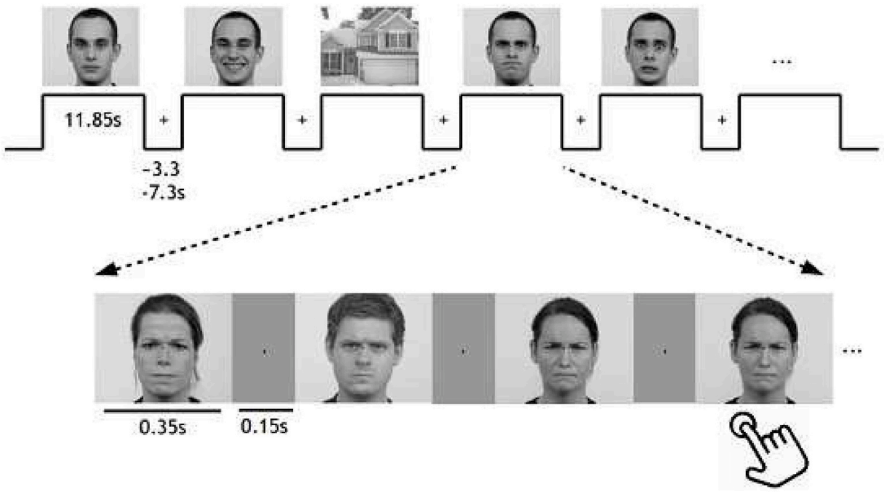


Fig. 2. Experimental paradigm for study A. In paradigm A, pictures of either neutral, happy, angry, or fearful faces (Langner et al., 2010) were shown in the experimental condition, and houses were shown in the control condition. Single stimuli and blocks were intervened by a gray screen. Participants were instructed to maintain the fixation of their gaze throughout the entire experiment. They were further instructed to press a button if a stimulus was presented twice in a row (one to two times per block). The total experiment lasted about 30 min.

$$\frac{dz}{dt} = \left(A + \sum_{j=1}^{m} u_j B^{(j)}\right) z + Cu$$

In this equation, z depicts the neural activations, u is the experimental input or context, A describes the endogenous connection strengths, $B^{(j)}$ models how the experimental context u_j affects connectivity in the network, and C models how the experimental input directly influences the neural activity within the regions of interest. The dynamics of the neural activations are translated into predictions about the blood oxygen level-dependent signal by a hemodynamic forward model (Buxton et al., 1998). The model parameters are then estimated by maximizing the negative free energy.

DCM enables inferences at different levels, such as the inferences on model space and parameter space of any given model. In the following sections, we will describe (i) the extraction of time series from the OFA, FFA, and STS, (ii) the specification of the model space, and (iii) the specific DCM analyses assessing the network parameters within and across the studies.

(i) Identification of the OFA, FFA, and STS

DCMs were constructed for the core system of face perception within the right hemisphere (OFA, FFA, and STS). In the following paragraphs, we describe how we defined regions of the core system and extracted the time series of the respective regions.

Two different approaches were used to identify brain regions at the single-participant level. Regarding the choice of the preprocessing steps, we did not adopt one specific standard for the present study. Instead, to increase the generalizability, we applied the approaches for time series extraction that had been used by the authors in the respective studies. The first approach was used for data set A, in which the MRI data was not normalized. In this data set, we manually identified the peak activation clusters at a single participant level in the native image space (Frässle et al., 2016c). We superimposed the participants' co-registered structural image with the t-map for the contrast "faces > control condition." We then identified the OFA, FFA, and STS as the clusters with the highest activities in the inferior occipital gyrus, posterior fusiform gyrus, and the posterior superior temporal sulcus, respectively. If several clusters were candidates for a particular region, we used the activation strength and symmetry to an analog cluster in the opposite hemisphere as criteria. The second approach was used for data sets B, C, and D, in which the MRI data was normalized (Kessler et al., 2020; Sladky et al., 2015). For each study, we first assessed the brain activity at the group level. The individual contrast images ("faces > control condition") were entered in a random-effects analysis using a one-sample t-test. We identified the group peak activation coordinates for the OFA, FFA, and STS using the same anatomical criteria as described above. Next, we identified participant-specific peak coordinates for these regions. A peak coordinate was defined in each participant as the voxel with the highest t-value within a mask (radius, 12 mm) centered on the group peak coordinate for the respective region.

For all the data sets, the time series were extracted for each region and participant/session as the first principal component of all the voxels activated at a threshold of 0.001, uncorrected for multiple comparisons, located within a radius of 4 mm around the participant-specific peak voxel. Due to the lower overall activation in data set D, we increased the statistical threshold to 0.1 (uncorrected) for this data set. Participants/sessions in which no activity was found at the pre-defined statistical threshold in at least one region were excluded from further analyses. Two participants from data set A, four from data set B, and three from data set C were excluded (Table 1).

(ii) Specification of model space

For all the data sets, we specified models similar to those in study FI (Fairhall and Ishai, 2007). All the models consisted of three regions: the OFA, FFA, and STS. These regions were interconnected differently, varying in the presence or absence of context-independent connections (A-matrix). In total, we constructed 24 models (Fig. 3). A 'face' input regressor was set onto the OFA in all the models (C-matrix). Furthermore, we allowed 'faces' to modulate all available interregional connections within each model (B-matrix). Intra-regional connections (i.e., self-connections) were not modulated in the B-matrix. Our model specification was informed by the models of study FI, as well as by assuming the OFA as an input region and by allowing an input to be distributed to all downstream regions by at least one possible route. However, our model specification deliberately differed from that in study FI with regard to the specification of the influence of face perception. In study FI, the influence of the presentation of faces, in comparison to other objects, was not modeled explicitly (see Section 4.2 for a detailed discussion). To assess the effects of 'emotion' and 'fame,' we further allowed the modulation of all interregional connections by 'emotion' (data set A and B) and 'fame' (data set D). At this point, we decided again to use a different modeling procedure compared to study FI (see discussion 4.2. for a more detailed explanation).

Whereas study B comprised only one emotion (fear, plus neutral expression), the regressor for 'emotion' was interpreted in a

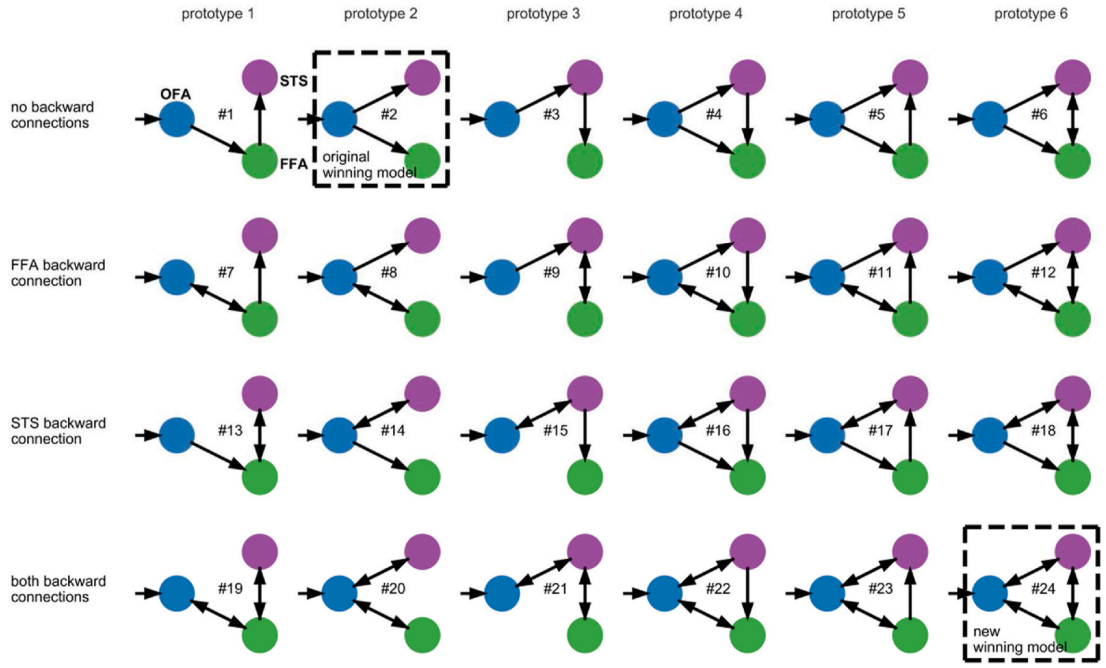


Fig. 3. Model space. Models of the core system of face perception tested with Bayesian model selection (BMS). Connectivity was investigated by modifying the forward, lateral, and feedback connections between the three investigated regions, namely the OFA (blue), FFA (green), and STS (purple). Driving input by faces was set on the OFA (C-matrix, short arrow). All context-independent connections (A-matrix) are displayed with arrows, except the inhibitory self-connections. All interregional connections were modulated (B-matrix) by 'faces' (studies A-D), 'emotion' (studies A and B), and 'fame' (study D). The winning model of the original study FI (#2) and the winning model of our revised model comparisons (#24, see Results section) are marked with dashed rectangles. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

straightforward manner (i.e., as an effect of 'fear'). In study A, however, three different expressions (happiness, fear, anger) were presented alongside neutral facial expressions. We deliberately pooled across all emotional expressions, except 'neutral,' to construct a regressor for 'emotion' to stay consistent with the approach of the original study (Fairhall and Ishai, 2007). In that study, pooling was conducted across two emotions; specifically, fear and happiness. We acknowledge that different emotions may lead to different activity and connectivity.

(iii) DCM Analysis

Our DCM analysis can be divided into three steps. First, we conducted Bayesian model selection (BMS) to assess which model is best supported by the data separately for each participant and study. Second, we used Bayesian model averaging (BMA) to estimate averaged model parameters separately for each participant and study. Last, as the main aim of the present study, we used HLM to assess model parameters across the participants and studies.

Bayesian model selection: First, we compared the different models using random-effects BMS separately for each study (Stephan et al, 2009, 2010). We quantified the models' goodness-of-fit based on the negative free energy, an approximation to the log model evidence (Friston et al., 2007). As a result of BMS, we obtained the posterior and exceedance probabilities for each model, assessed across all the participants within each study. Our objective was not just to assess whether the winning models in our data sets were congruent with the winning model reported in study FI but also to qualitatively assess if the winning model is consistent across all the studies.

Bayesian model averaging: Next, we calculated the averaged model parameters via BMA (Penny, 2012; Penny et al., 2010)). BMA uses the posterior model probabilities of all the models of a particular participant and calculates a weighted average model. The weights were determined by the respective posterior model probabilities. BMA, therefore, accounts for the uncertainty of each model (Stephan et al., 2010). The results are presented at the single participant and group levels. The single participant results allow the visualization of the variance across the participants within one study (see Figs. S1–S4). The group results allow the description of the variability of the results across the studies. Two-sided one-sample t-tests were conducted for each connection per study to assess whether a connection parameter significantly differed from zero. We applied a Bonferroni family-wise error correction within each matrix for a particular study, resulting in a threshold of $\alpha_{Bonf} = \frac{\alpha}{n} = \frac{0.05}{6}$, with n as the number of tests, and α as the native false-positive threshold. We tested inter-regional connections (i.e., off-diagonal elements of the respective matrix). Self-connections were first converted to unit Hertz by applying $a_{Hz} = -0.5^{*}e^{\{a_{logscale}\}}$ to be on the same scale as the inter-regional connections (Zeidman et al., 2019a). We did not test self-connections for significance because those are negative by definition (Fig. S2).

Studies C and D included more than one experimental session per participant. In study C, each participant was measured twice, with the participants' hormone levels differing between the two experimental sessions. Therefore, we have reported the BMS and BMA results for both sessions separately. In study D, we included five to nine experimental sessions per participant depending on the number of sessions in which all the regions could be clearly identified (see 2.4.2.[i]). The division into two separate sessions was not motivated by an experimental manipulation as in study C. For the sake of clarity, we will not report group-BMS and group-BMA results across all nine sessions in study D. However, for the subsequent analysis with HLM, we included each participant and session appropriately.

***Hierarchical linear modeling*:** Third, as the main aim of the present study, we estimated the model parameters across the studies. In the

preceding step, we used the model probabilities of each participant to create an average model for each participant and the respective session. Now, we aimed to quantify the connectivity parameters across all the sessions, participants, and studies.

To assess these group effects, we constructed HLMs using the R (R version 3.6.2, (R Core Team, 2020)) packages lme4 (lme4_1.1) and nlme (nlme_3.1) (Bates et al., 2015; Lindstrom and Bates, 1990). We decided to use hierarchical modeling instead of simple multiple linear modeling to account for the hierarchical structure in the data. Hierarchical structures were introduced by studies C and D, in which participants were measured multiple times.

The present HLM approach evaluates the magnitude of each connectivity parameter between regions. These parameters were nested into studies and further nested into repeated measurements per participant. For HLM, we deliberately used the point estimate of the posterior parameter of each participant and session after BMA.

To describe the magnitude of a particular connectivity parameter, we modeled it as a function of the study and hormone as fixed effects, respectively. Fixed effects are unknown, constant parameters, which are like regression coefficients in multiple regression analysis. We modeled the particular participant as a random effect. Random-effects represent random (unobserved) variables (West et al., 2014) instead of simple regression coefficients. More illustratively, we modeled each participant having a random intercept. Consequently, the participants' intercepts deviated around the fixed-effect, or global, intercept.

We were not interested in the interpretation of the effects of the study, participant, or hormone. We were, however, interested in the shared connectivity across the studies, participants, and sessions. Therefore, it was important to design the model such that the global intercept can be interpreted as an average parameter estimate across the studies. To achieve this interpretation, we used contrast coding or Helmert coding on the study variable and hormone variable (Sundström, 2010). In the first contrast variable ('AvsB'), we assigned a value of +0.5 for all the observations belonging to study A and −0.5 for all the observations belonging to study B. Next, we included study C ('ABvsC') by contrasting studies A and B (+0.25 each) versus study C (−0.5). We continued the same way with study D ('ABCvsD') by assigning +0.16 for the observations of studies A, B, and C and −0.5 for the observations of study D. Similarly, we introduced a one-level Helmert coding for the hormone variable, contrasting mid-luteal vs. early follicular phase ('MvsP').

For each connection of each DCM matrix, we constructed a separate HLM. Of those HLMs, we emphasized the global intercept (i.e., fixed-effect intercept) of the corresponding model. When modeling the DCM parameters of the A-matrix, B-matrix 'faces', and C-matrix, we included all the terms. When modeling the B-matrix 'emotions', we dismissed the explanatory variable 'hormone' because study C and study D did not include emotions in their paradigms. When analyzing the B-matrix 'fame', we did not include 'hormone' or 'study' as we just used study D for this analysis. As an example, a particular B-matrix connectivity parameter for the effect of 'faces' was modeled in the following manner:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma x_{i4} + u_i + \varepsilon_i$$

with y_i being the DCM parameter (response variable) of participant i, β_0 representing the global (fixed effect) intercept, β_1 to β_3 representing the slopes of the contrasts of the study variables x_{i1} to x_{i3}, respectively, γ being the slope of the contrast of the hormone variable x_{i4} (all fixed effects). u_i corresponds to the random effect of 'participant,' and ε_i is the random error, with $\varepsilon_i \sim N(0,{\sigma_\varepsilon}^2)$, and $u_i \sim N(0,{\sigma_u}^2)$. When modeling the parameters of other matrices, such as 'emotion' or 'fame,' particular fixed-effect terms were dismissed according to the logic described above. Using contrast coding, we tested the intercept for significance, applying a Bonferroni family-wise error correction with a threshold of $\alpha_{Bonf} = \frac{\alpha}{n} = \frac{0.05}{6}$, with n as the number of tests on interregional connections per matrix (A matrix, B matrix 'emotion,' and B matrix 'fame'), and α as the native false-positive threshold.

3. Results

The results section is structured as follows: first, we present a comparison of all the neural models using BMS separately for each study (3.1). Second, we describe the weighted parameter estimates after participant-specific BMA for all the data sets (3.2). Last, we present the HLM results showing parameter estimates across the studies (3.3). Based on this, we propose a revised model of the core face perception network.

3.1. Bayesian model selection

First, we conducted a BMS separately for each study. The results for study C are presented separately for both sessions (corresponding to two different phases of the participants' menstrual cycle). Group results are not displayed for study D because of the variable number of sessions included for each participant.

The posterior probability for model #24 (see Fig. 3) was the highest in all the studies (Fig. 4, left panel), ranging from 0.248 (study C1) to 0.417 (study B). Similarly, the exceedance probabilities for model #24 — the probabilities that model #24 is more likely than any of the other models — ranged from 0.915 (study C1) to >0.999 (study B, Fig. 4, right panel). The winning model expressed the highest possible interconnectivity in each analyzed data set. In all the data sets analyzed, we discerned the same winning model with a high posterior and exceedance probability (Fig. 4). Interestingly, our winning model differs from that of study FI (see Fig. 3).

3.2. Bayesian model averaging

In the second step, we calculated an average model for each participant and study using BMA. BMA uses the posterior model probabilities of all the models of a particular participant and calculates a weighted average model. The weights were determined by the respective posterior model probabilities. BMA, therefore, accounts for the uncertainty of each model, as revealed by BMS (Stephan et al., 2010). Kernel density estimates of the participant-specific connectivity parameters after BMA, grouped by the respective study for the A-matrix, C-matrix, and all B-matrices, are illustrated in the supplementary results (Figs. S1–S4). The kernel density plots visualize the variability of the single participant parameter estimates grouped by the respective study.

To calculate a separate model for each study, we applied a one-sample *t*-test onto each connectivity parameter separately for each study. We used a Bonferroni-corrected threshold of $p = 0.05$ per matrix and study (see Methods). The average models for each study are displayed in Fig. 5. The connectivity patterns for each study were similar; although the average connections may have differed in magnitude, they tended to point in the same direction (i.e., positive or negative). Moreover, some connections exceeded the threshold for significance in one study but not in the others. Therefore, naively contemplating each study in the absence of the others could lead one to draw similar conclusions regarding many parameters while disregarding other parameters due to significance thresholds.

As a general pattern, the following was observed: within the A-matrix, the parameters were relatively small and rarely significant. The C-matrix was always significantly positive. Within the B-matrix ('faces'), forward connections from the OFA to the FFA and the OFA to the STS were always significantly positive. Most of the time, the backward connections from the FFA to the OFA and the STS to the OFA were negative (sometimes significantly). Collateral and backward connections between the FFA and STS were always negative (sometimes significantly). The B-matrices ('emotion') showed weaker parameters which were rarely significant.

We tested for statistical significance across the studies in the following step to identify the global effects using HLM.

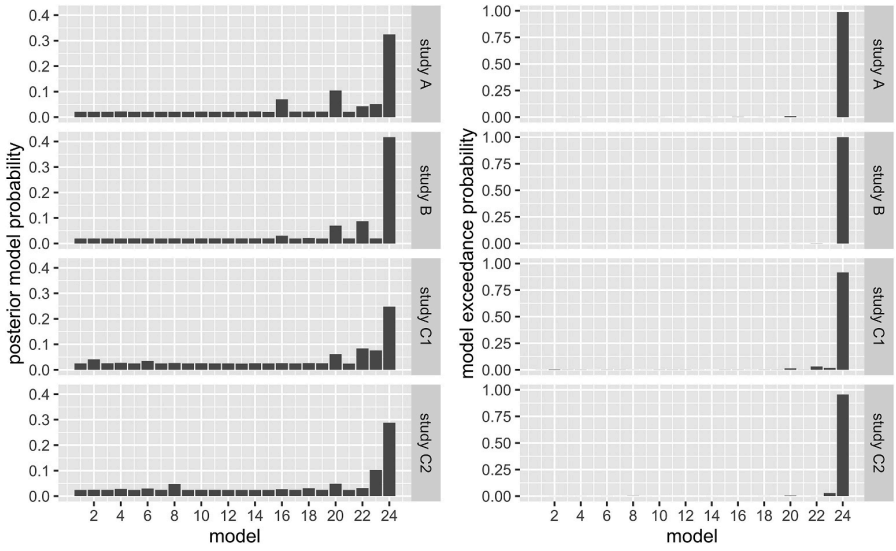


Fig. 4. Bayesian model selection results. Left panel: The posterior model probabilities are displayed. We see that model #24 has the highest relative probability with 0.248 (study C1) to 0.417 (study B). Right panel: The model exceedance probabilities are displayed. In all the data sets, model #24 exhibited a high exceedance probability (>0.9).

3.3. Hierarchical linear modeling

In the final step, we assessed the commonalities in the participant- and session-specific connectivity parameters across the studies to investigate the modulatory influences of 'faces,' 'emotion,' and 'fame' on the network, as well as the interregional, context-independent connection of the A-matrix and the driving input (C-matrix). We used HLM as a tool to quantify the magnitude and significance of each connection. We included all the significant connections in a new and revised model of the core face perception network (Fig. 6).

Using HLM, we identified the intercepts representing the 'average effects across studies' that significantly differed from zero. We have displayed all the connections in pseudo-colors in Fig. S5. Furthermore, we have displayed all the significant connections in a model-like structure in Fig. 6. First, in the context-independent connections (A-matrix), only the forward connection from the OFA to the FFA showed significant positivity (+0.08, $p = 0.0016$). The corresponding backward connection from the FFA to the OFA was significantly negative (−0.19, $p = 3.3*10^{-8}$). Further, the driving input into the system (C-matrix) had a positive value (+1.42, $p = 9.3*10^{-31}$). 'Faces' positively modulated the forward connection from the OFA to the FFA by +0.92 ($p = 1.3*10^{-17}$), and that from the OFA to the STS by +0.77 ($p = 2.8*10^{-13}$). 'Faces' negatively modulated the backward connection from the FFA to the OFA by −1.13 ($p = 3*10^{-14}$), and the collateral connections from the FFA to the STS by −0.31 ($p = 0.002$) and vice versa by −0.4 ($p = 0.0007$). Similarly, 'emotions' positively modulated the forward connection from the OFA to the STS by +0.35 ($p = 7.7*10^{-6}$) and negatively modulated the collateral connection from the FFA to the STS by −0.19 ($p = 0.0005$). However, 'fame' did not significantly modulate any connection.

Our resulting model has some similarities and differences compared with the original study. The similarities include the increase of forward-coupling induced by 'faces.' Differences mainly relate to the connections not included in the winning model of study FI. 'Emotions' modulated the forward connection to the STS instead of those to the FFA. We discuss possible reasons for the differences between our results and those of study FI below.

4. Discussion

In this study, we conducted a conceptual replication of an early network model of face perception using multiple data sets. While we successfully reproduced some aspects of the original model, the revised model was distinct in terms of some other major aspects.

We will first describe our revisited model in terms of single interactions and compare it to the original model (4.1). Secondly, we will discuss the modifications applied to our analysis pipeline compared to that of the original study (4.2). Some of these modifications were introduced by us to remedy issues in the original study, which may have limited its interpretability. Other modifications were merely due to developments within the DCM framework which have been introduced in new software versions. Further, we embed the presented network model within the broader framework of the predictive coding theory and outline some limitations (4.3). Finally, we emphasize the importance of conceptual replications in network neuroscience (4.4).

4.1. The revisited model of face perception

We tested face perception models consisting of the OFA, FFA, and STS, with the OFA serving as a hierarchically early input region that propagates information to the FFA and STS. As stated previously, inference in DCM is possible at different levels; it is possible at the level of the model space (i.e., which model is most likely) and parameter space (i.e., the shape of model parameters) (Stephan et al., 2010). Regarding the model space, we showed that our winning model was fully interconnected. This total interconnectivity was revealed by BMS in all the different samples and paradigms (Fig. 4); it comprised forward, backward, and lateral connections. The model proposed by study FI comprised merely forward connections (Fig. 1). Recently published studies have proposed hemispheric differences in the degree of inter-connectivity. For instance, Wang et al. (2020) quantified structural, functional, and effective connectivity within the core- and extended systems of face perception. They reported higher interconnectivity within the face perception system of the right hemisphere comprising both feed-forward and feedback connections, while the left hemisphere showed a predominantly feed-forward pattern (Wang et al., 2020).

Regarding the parameter space: in all the models, the external input

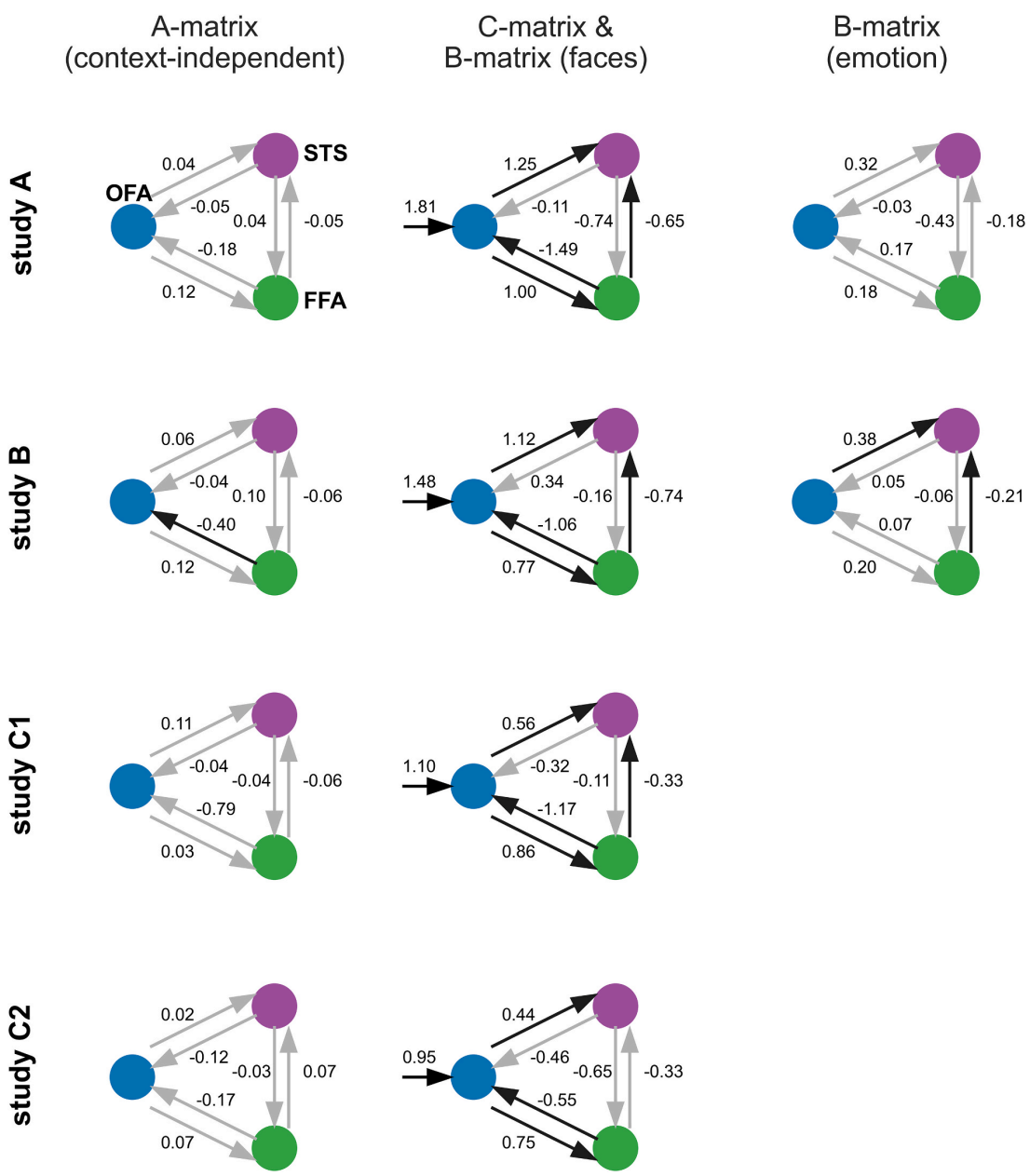


Fig. 5. The average connectivity within each study. Studies A, B (upper panels), and C (lower panels) were divided into two scanning sessions. The connectivity between the following three regions is illustrated: the OFA (blue), FFA (green), and STS (purple). In the left panel, the A-matrix (context-independent coupling) is shown. In the middle panel, the driving input ('faces,' C-matrix) and B-matrix ('faces') are displayed, and in the right panel, the B-matrix ('emotions') is shown. Black arrows indicate significant connections (i.e., significant within-study). Gray arrows indicate non-significant connections. The number alongside each arrow indicates the average connection strength. Self-connections (A-matrix) were omitted in the figures but distributed around −0.5 (see Fig. S2). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

was modeled via the effect of 'faces' in the C-matrix. 'Faces' entered the system via the OFA according to the Haxby model. However, concurring theories propose the FFA as the input region (Rossion, 2008). As a working model, we stick to the OFA as a hierarchically earliest region and, therefore, target region for the driving experimental input, consistent with the Haxby model (Haxby et al., 2000). We further modeled the 'effect of faces' on every interregional connection (B-matrix). Across the studies and participants, we found five significant modulations of 'faces' on interregional connections. 'Faces' positively increased the forward connectivity from the OFA to the FFA and from the OFA to the STS; this supports the prevailing opinion that face perception drives such forward connectivity, as proposed in the original Haxby model (Fan et al., 2020; Haxby et al., 2000). Further, we found a significantly negative backward connectivity from the FFA to the OFA and collateral connectivity between the FFA and STS.

'Emotion' further increased the positive forward connection strength

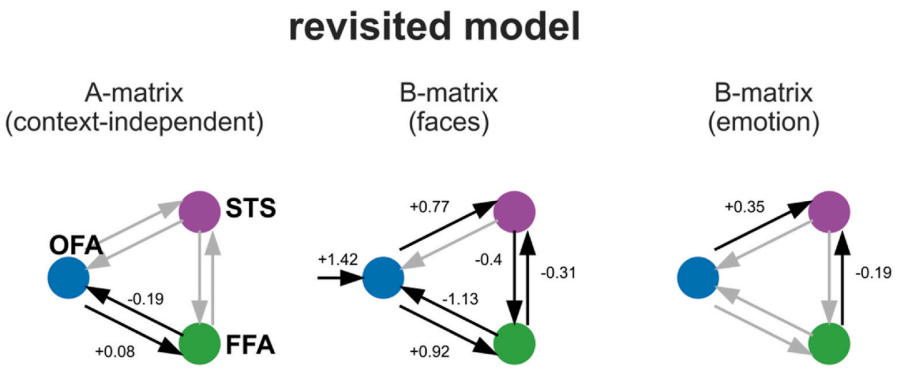


Fig. 6. A revisited model for the core system of face perception. Driving input ('faces') enters the OFA. Significant connections, as revealed by HLM, are displayed with black arrows and depicted with numbers. Non-significant (determined by HLM) but present (determined by BMS) connections are illustrated by gray arrows without numbers. The context-independent connections and modulatory effects of 'faces' and 'emotion' are displayed separately. 'Fame' did not significantly modulate any present connection and is therefore not shown. The final model of the original study is depicted in Fig. 1 for comparison.

from the OFA to the STS and the negative coupling from the FFA to the STS in the revised model (Fig. 6). Here, we differed from the original model ((Fairhall and Ishai, 2007), Fig. 1), which proposed a positive forward modulation by 'emotion' from the OFA to the FFA; based on the single parameters across the studies, we could not clearly discern this across the presently analyzed paradigms. However, previous fMRI studies emphasize the importance of the STS in emotion recognition (Duchaine and Yovel, 2015; Engell and Haxby, 2007; Haxby et al., 2000; Hildesheim et al., 2020; Sliwinska et al., 2020).

Lastly, the effect of 'fame' was not significant, even though it was only modeled in one of our paradigms. According to lesion studies and imaging studies, face familiarity may be processed in more anterior regions, such as the anterior temporal face area in combination with the FFA (Busigny et al., 2014; Evans et al., 1995; Sergent et al., 1992; Williams et al., 2006). To disentangle the effects of 'fame,' models with anterior temporal face regions included might provide better insights.

4.2. Methodological adjustments to the original model

We deliberately introduced some modifications to the original DCM pipeline as described below.

4.2.1. A-matrix and experimental effects

The A-matrix represents the context-independent coupling between regions, i.e., the underlying effective connectivity throughout the entire experiment (control conditions, fixations, etc.). Other effects, such as the effect of 'faces' on a particular connection, specified via the B-matrix, are additive to the context-independent parameters. Deciding which experimental effects to model in which matrices are important in the DCM workflow. We decided to model the effect of 'faces' explicitly in a B-matrix; this allowed us to differentiate the connectivity induced by 'faces' from the residual connectivity at rest or induced by any control condition. In study FI, the effect of 'faces' was not modeled explicitly in a B-matrix (unlike how they modeled the effects of 'emotions' and 'fame' in a B-matrix). Instead, the A-matrix parameters were interpreted as the effect of 'faces,' which were confounded by all the conditions present in the respective experimental runs.

Irrespective of the matrix in which study FI and our study modeled 'faces,' the effect of 'faces' highly overlapped between both studies; the positive forward connectivity from the OFA to the FFA and the OFA to the STS was present in both the original model (Fig. 1) and our revised model (Fig. 6). Backward connections were modeled in the original study but did not survive the model selection (Fairhall and Ishai, 2007).

4.2.2. One-vs. two-step model selection

The effects of 'emotion' and 'fame' in the original study were modeled in a two-step approach. First, the authors assessed the coarse structure of the model by conducting a BMS that only specified the A- and C-matrices. However, it is unclear if the experimental input was properly distributed across the regions without the specification of a B-matrix onto the connections. Due to the control conditions and rest periods, the resulting A-matrix parameters potentially underestimated the true effect of 'faces.' Similarly, the parameters of the A-matrix were provided with more narrow shrinkage priors, much tighter than those of the B-matrix (Zeidman et al., 2019a), which under Bayesian assumptions lead to a weaker posterior parameter estimate.

However, model #2 was selected by BMS in study FI (Fig. 1, left or Fig. 3). Then, the authors added B-matrices for 'emotion' or 'fame' in the appropriate paradigms and reported the significance of the resulting coupling parameters; however, the model selection procedure did not account for these additional regressors. Therefore, the model selection could have yielded different results if these regressors had been included. For this reason, we included all the regressors (A-, B-, and C-matrices in the respective paradigms) from the beginning to avoid biasing the model selection.

4.2.3. The use of different information criteria

Since the original study was published (Fairhall and Ishai, 2007), the DCM framework has undergone significant developments. One implementation was free energy (Friston et al., 2007; Penny, 2012), which became the preferred choice of information criterion. However, in study FI, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were the current standard information criteria that, under certain signal-to-noise ratio conditions, are not sensitive for fully interconnected models. Instead, they deploy a high penalty for the number of parameters (i.e., model complexity) (Penny, 2012). Conversely, free energy incorporates the covariance between the parameters, increasing the sensitivity for fully connected models (Penny, 2012). However, it has also been shown that free energy overemphasizes fully connected models (Litvak et al., 2019). We additionally repeated the BMS analysis with AIC and BIC rather than F. The results are illustrated in Fig. S6 in the supplementary material and demonstrate that the different information criteria have strongly contributed to the differences in the structure of the winning model (Fig. S6). However, none of the BMS results corresponded to the results of the original study FI (Fig. S6).

4.2.4. Modeling across different data sets

We included four different data sets in our analysis; thus, we needed to include covariates to control for specific independent variables of the different studies. A relatively novel method to include covariates in DCM is the parametric empirical Bayes (PEB) framework (Friston et al., 2016; Zeidman et al., 2019b). This framework allows second-level dynamic causal models to assess the effects of covariates across a group or between groups. However, using PEB was not practical in our study, as we dealt with different dependencies and B-matrices for each data set. Further, within the PEB framework, participants are weighted differently according to their respective model fit; we wanted each participant

to be weighted rather equally in a group analysis. Due to these reasons, we decided to use HLM instead of PEB.

4.3. Face perception revisited in the predictive coding framework

In the following, we embed our resulting main model parameters (Fig. 6, Fig. S5) into the broader context of predictive coding as the predictive coding framework generally seems well-suited for such hierarchical models. Despite the oversimplification of the complex predictive coding theory, we integrated our model in the predictive coding framework for a comprehensive and meaningful interpretation at the level of the resulting parameter estimates.

Briefly, in the predictive coding framework, the brain is organized into hierarchical interconnected modules. Each module communicates predictions (i.e., expectations about its input) to the respective lower level. Similarly, each module calculates a prediction error as the discrepancy between the prediction (i.e., the expected signal from the lower level) and input (i.e., received a signal from the lower level). The prediction error is then propagated to the respective higher level, wherein the prediction is updated. The updated prediction is then propagated back to the respective lower level (prediction updating). This iterative process is described on a microscopic scale (Bastos et al., 2012) within the early visual hierarchy (Rao and Ballard, 1999) and on a macroscopic scale in the context of DCM (Chen et al., 2009; Den Ouden et al., 2009).

In this framework, we might interpret the positive parameters from lower regions (OFA) to hierarchically higher regions (FFA and STS) as prediction error signaling, analogous to a forward propagation of the signal along the hierarchy (Fig. 6). Conversely, we might interpret the negative backward connections from higher to lower level areas as prediction updating. Prediction updating in Bayesian networks is equivalent to "explaining away the stimulus" (Gotts et al., 2012), whereby the causes of the sensory input are learned, and the prediction error, which is the neural activation that results in the positive forward-coupling, gets reduced. It is plausible that over the course of an experimental simulation, the presence of a particular input stimulus (either a sequence of faces or a single face, depending on the experimental paradigm) is learned, therefore "explained away," causing the positive and negative couplings on a macroscopic scale.

In the previous paragraphs, we deliberately detailed an interpretation of the positive forward connectivity and negative backward connectivity in the context of the predictive coding theory. However, positive forward connectivity appears to be the most obvious option available. If all three neural regions are activated by faces within the respective fMRI paradigms, and the input regressor of faces (C-matrix) enters the system via the OFA, the obvious explanation in the context of the full model (#24 in Fig. 3) is a positive forward transfer to the other two regions. Alternatively, a positive forward connection to one region and positive collateral connectivity from this region to the remaining regions could also be an alternative pathway to activate all the regions. Analogous effects may be the easiest way to explain the positive activity by face perception within all three regions in the context of other models, such as those evolving from prototypes 1 and 3 (Fig. 3).

The lack of alternatives for the general expression of the parameters can also be seen in the negative backward connectivity by 'faces' from the FFA to the OFA. Usually, negative self-connections within a region induce a decrease in activity within that region over time (e.g., during the whole modeled experiment) and prevent the system from becoming epileptic. However, self-connections in our models were context-independent, as they were only present in the A-matrix, and we did not allow the modulation of those in the B-matrix. Therefore, the inhibitory parameters remained the same throughout the course of the experiment, regardless of whether it was an experimental or control condition. In the experimental condition ('faces'), the activity in all the presently modeled regions was higher (see the definition of the regions for the extraction of the time series). Therefore, allowing only the connections from other regions (instead of self-connections) to downregulate this additional activity may have caused such a manifestation of negative couplings between regions. This concerns the negative backward couplings from the FFA and STS toward the OFA, which downregulate the OFA activity. Further, this concerns the negative collateral connections between the FFA and STS, which downregulate the STS and FFA, respectively.

Experiments and simulations are required to validate these theories in the future. However, such effects, implicitly introduced by the setup of the models, limit any extensive interpretation of our revised model or any similar model. However, complementary imaging techniques such as EEG/MEG, which have a far better temporal resolution, might shed light on the time-sensitive orchestrations between the regions during bottom-up and top-down processing. For instance, a recent study by Fan et al. (2020) investigated response times of the regions of the core system using specialized paradigms to untie top-down and bottom-up processes within the predictive coding framework (Fan et al., 2020). Interpretations using fMRI however can rather be made for long-lasting interactions in the brain.

4.4. The requirement of conceptual replications

As we have already discussed in the introduction, neuroimaging findings are often vulnerable to non-replication (Gorgolewski and Poldrack, 2016). DCM may be particularly vulnerable to this due to the massive number of degrees of freedom a researcher is faced within the analysis. Additionally, changes in the experimental setup, pipeline, statistical methods, and even software versions can cause significant changes in the parameter estimates (Bedenbender et al., 2011; Botvinik-Nezer et al., 2020; Frässle et al., 2016b; Weissenbacher et al., 2009). As we can usually only investigate very narrow hypotheses in a single study, we highly depend on the validity and reproducibility of the previous results being built upon. Therefore, we need more conceptual replications and meta-analyses of models like that in the present study. Most importantly, we need to be critical and mindful while interpreting previously published results.

5. Conclusion

The aim of our study was to conceptually replicate the main findings of Fairhall and Ishai (2007) on the effective connectivity within the core system of face perception. Across four different data sets, we demonstrated that our revised model was more complex than the originally proposed model, with a high degree of interaction between regions.

Funding

This work was supported by funds from the Deutsche Forschungsgemeinschaft (DFG, grant no. JA 1890/11-1) and from the von-Behring-Röntgen Stiftung (grant no. 63-0030).

Data and material availability

Data of study D is available online (https://www.openneuro.org, accession number ds000117, (Wakeman and Henson, 2015)). First-level analyses from study A are available at github.com/kesslerr/efp. From all the studies, data is fully available for the DCM modeling, beginning with the DCM models, model comparison results, model averaging results, hierarchical linear modeling pipeline, and other statistical tests at github.com/kesslerr/coreSysRev. Further data can be provided on request.

Declarations of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ynirp.2021.100045.

References

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711. https://doi.org/10.1016/j.neuron.2012.10.038.

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67. https://doi.org/10.18637/jss.v067.i01.

Bedenbender, J., Paulus, F.M., Krach, S., Pyka, M., Sommer, J., Krug, A., Witt, S.H., Rietschel, M., Laneri, D., Kircher, T., Jansen, A., 2011. Functional connectivity analyses in imaging genetics: considerations on methods and data interpretation. PLoS One 6. https://doi.org/10.1371/journal.pone.0026354.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., Avesani, P., Baczkowski, B.M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R.G., Berkers, R.M.W.J., Bhanji, J.P., Biswal, B.B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K.L., Bowring, A., Braem, S., Brooks, H.R., Brudner, E.G., Calderon, C.B., Camilleri, J.A., Castrellon, J.J., Cecchetti, L., Cieslik, E.C., Cole, Z.J., Collignon, O., Cox, R.W., Cunningham, W.A., Czoschke, S., Dadi, K., Davis, C.P., Luca, A., De Delgado, M.R., Demetriou, L., Dennison, J.B., Di, X., Dickie, E.W., Dobryakova, E., Donnat, C.L., Dukart, J., Duncan, N.W., Durnez, J., Eed, A., Eickhoff, S.B., Erhart, A., Fontanesi, L., Fricke, G. M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J.J., Golowin, S.A.E., González-García, C., Gorgolewski, K.J., Grady, C.L., Green, M.A., Guassi Moreira, J.F., Guest, O., Hakimi, S., Hamilton, J.P., Hancock, R., Handjaras, G., Harry, B.B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.P., Huettel, S.A., Hughes, M.E., Iacovella, V., Iordan, A.D., Isager, P.M., Isik, A.I., Jahn, A., Johnson, M.R., Johnstone, T., Joseph, M.J.E., Juliano, A.C., Kable, J.W., Kassinopoulos, M., Koba, C., Kong, X.Z., Koscik, T.R., Kucukboyaci, N.E., Kuhl, B.A., Kupek, S., Laird, A. R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M.Y.C., Lim, P.C., Lintz, E.N., Liphardt, S.W., Losecaat Vermeer, A.B., Love, B.C., Mack, M.L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J.T., Melero, H., Méndez Leal, A.S., Meyer, B., Meyer, K.N., Mihai, G., Mitsis, G.D., Moll, J., Nielson, D.M., Nilsonne, G., Notter, M.P., Olivetti, E., Onicas, A.I., Papale, P., Patil, K.R., Peelle, J.E., Pérez, A., Pischedda, D., Poline, J.B., Prystauka, Y., Ray, S., Reuter-Lorenz, P.A., Reynolds, R.C., Ricciardi, E., Rieck, J.R., Rodriguez-Thompson, A.M., Romyn, A., Salo, T., Samanez-Larkin, G.R., Sanz-Morales, E., Schlichting, M.L., Schultz, D.H., Shen, Q., Sheridan, M.A., Silvers, J.A., Skagerlund, K., Smith, A., Smith, D.V., Sokol-Hessner, P., Steinkamp, S.R., Tashjian, S.M., Thirion, B., Thorp, J.N., Tinghög, G., Tisdall, L., Tompson, S.H., Toro-Serey, C., Torre Tresols, J.J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A.E., Verguts, T., Vettel, J.M., Vijayarajah, S., Vo, K., Wall, M.B., Weeda, W.D., Weis, S., White, D.J., Wisniewski, D., Xifra-Porxas, A., Yearling, E.A., Yoon, S., Yuan, R., Yuen, K.S.L., Zhang, L., Zhang, X., Zosky, J.E., Nichols, T.E., Poldrack, R.A., Schonberg, T., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88. https://doi.org/10.1038/s41586-020-2314-9.

Busigny, T., Van Belle, G., Jemel, B., Hosein, A., Joubert, S., Rossion, B., 2014. Face-specific impairment in holistic perception following focal lesion of the right anterior temporal lobe. Neuropsychologia 56, 312–333. https://doi.org/10.1016/j.neuropsychologia.2014.01.018.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the ballon model. Magn. Reson. Med. 39, 855–864. https://doi.org/10.1002/mrm.1910390602.

Chen, C.C., Henson, R.N., Stephan, K.E., Kilner, J.M., Friston, K.J., 2009. Forward and backward connections in the brain: a DCM study of functional asymmetries. Neuroimage 45, 453–462. https://doi.org/10.1016/j.neuroimage.2008.12.041.

Den Ouden, H.E.M., Friston, K.J., Daw, N.D., McIntosh, A.R., Stephan, K.E., 2009. A dual role for prediction error in associative learning. Cerebr. Cortex 19, 1175–1185. https://doi.org/10.1093/cercor/bhn161.

Diener, E., Biswas-Diener, R., 2016. The replication crisis in psychology. HKU PSYC2020 Fundam. Soc. Psychol. 6–18. https://doi.org/10.1016/B978-0-12-809324-5.05637-6.

Dima, D., Stephan, K.E., Roiser, J.P., Friston, K.J., Frangou, S., 2011. Effective connectivity during processing of facial affect: evidence for multiple parallel pathways. J. Neurosci. 31, 14378–14385. https://doi.org/10.1523/JNEUROSCI.2400-11.2011.

Duchaine, B., Yovel, G., 2015. A revised neural framework for face processing. Annu. Rev. Vis. Sci. 1, 393–416. https://doi.org/10.1146/annurev-vision-082114-035518.

Elbich, D.B., Molenaar, P.C.M., Scherf, K.S., 2019. Evaluating the organizational structure and specificity of network topology within the face processing system. Hum. Brain Mapp. 40, 2581–2595. https://doi.org/10.1002/hbm.24546.

Engell, A.D., Haxby, J.V., 2007. Facial expression and gaze-direction in human superior temporal sulcus. Neuropsychologia 45, 3234–3241. https://doi.org/10.1016/j.neuropsychologia.2007.06.022.

Evans, J.J., Heggs, A.J., Antoun, N., Hodges, J.R., 1995. Progressive prosopagnosia associated with selective right temporal lobe atrophy. Brain 118, 1–13.

Fairhall, S.L., Ishai, A., 2007. Effective connectivity within the distributed cortical network for face perception. Cerebr. Cortex 17, 2400–2406. https://doi.org/10.1093/cercor/bhl148.

Fan, X., Wang, F., Shao, H., Zhang, P., He, S., 2020. The bottom-up and top-down processing of faces in the human occipitotemporal cortex. Elife 9, 1–21. https://doi.org/10.7554/eLife.48764.

Frässle, S., Krach, S., Paulus, F.M., Jansen, A., 2016a. Handedness is related to neural mechanisms underlying hemispheric lateralization of face processing. Sci. Rep. 6, 1–17. https://doi.org/10.1038/srep27153.

Frässle, S., Paulus, F.M., Krach, S., Jansen, A., 2016b. Test-retest reliability of effective connectivity in the face perception network. Hum. Brain Mapp. 37, 730–744. https://doi.org/10.1002/hbm.23061.

Frässle, S., Paulus, F.M., Krach, S., Schweinberger, S.R., Stephan, K.E., Jansen, A., 2016c. Mechanisms of hemispheric lateralization: asymmetric interhemispheric recruitment in the face perception network. Neuroimage 124, 977–988. https://doi.org/10.1016/j.neuroimage.2015.09.055.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7.

Friston, K.J., Litvak, V., Oswal, A., Razi, A., Stephan, K.E., Van Wijk, B.C.M., Ziegler, G., Zeidman, P., 2016. Bayesian model reduction and empirical Bayes for group (DCM) studies. Neuroimage 128, 413–431. https://doi.org/10.1016/j.neuroimage.2015.11.015.

Friston, K.J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234. https://doi.org/10.1016/j.neuroimage.2006.08.035.

Furl, N., 2015. Structural and effective connectivity reveals potential network-based influences on category-sensitive visual areas. Front. Hum. Neurosci. 9, 1–13. https://doi.org/10.3389/fnhum.2015.00253.

Furl, N., Henson, R.N., Friston, K.J., Calder, A.J., 2015. Network interactions explain sensitivity to dynamic faces in the superior temporal sulcus. Cerebr. Cortex 25, 2876–2882. https://doi.org/10.1093/cercor/bhu083.

Gobbini, M.I., Haxby, J.V., 2007. Neural systems for recognition of familiar faces. Neuropsychologia 45, 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015.

Gorgolewski, K.J., Poldrack, R.A., 2016. A practical guide for improving transparency and reproducibility in neuroimaging research. PLoS Biol. 14, 1–13. https://doi.org/10.1371/journal.pbio.1002506.

Gotts, S.J., Chow, C.C., Martin, A., 2012. Repetition priming and repetition suppression: a case for enhanced efficiency through neural synchronization. Cognit. Neurosci. 3, 227–237. https://doi.org/10.1080/17588928.2012.670617.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cognit. Sci. 4223–233 4, 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0.

He, W., Garrido, M.I., Sowman, P.F., Brock, J., Johnson, B.W., 2015. Development of effective connectivity in the core network for face perception. Hum. Brain Mapp. 36, 2161–2173. https://doi.org/10.1002/hbm.22762.

Herrington, J.D., Taylor, J.M., Grupe, D.W., Curby, K.M., Schultz, R.T., 2011. Bidirectional communication between amygdala and fusiform gyrus during facial recognition. Neuroimage 56, 2348–2355. https://doi.org/10.1016/j.neuroimage.2011.03.072.

Hildesheim, F.E., Debus, I., Kessler, R., Thome, I., Zimmermann, K.M., Steinsträter, O., Sommer, J., Kamp-Becker, I., Stark, R., Jansen, A., 2020. The trajectory of hemispheric lateralization in the core system of face processing: a cross-sectional functional magnetic resonance imaging pilot study. Front. Psychol. 11, 1–15. https://doi.org/10.3389/fpsyg.2020.507199.

Kessler, R., Schmitt, S., Sauder, T., Stein, F., Yüksel, D., Grotegerd, D., Dannlowski, U., Hahn, T., Dempfle, A., Sommer, J., Steinsträter, O., Nenadic, I., Kircher, T., Jansen, A., 2020. Long-term neuroanatomical consequences of childhood maltreatment: reduced amygdala inhibition by medial prefrontal cortex. Front. Syst. Neurosci. 14, 1–11. https://doi.org/10.3389/fnsys.2020.00028.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., van Knippenberg, A., 2010. Presentation and validation of the radboud faces database. Cognit. Emot. 24, 1377–1388. https://doi.org/10.1080/02699930903485076.

Lindstrom, M.J., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 46, 673–687.

Litvak, V., Jafarian, A., Zeidman, P., Tibon, R., Henson, R.N., Friston, K., 2019. There's no such thing as a "true" model: the challenge of assessing face validity. In: Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern. 2019-Octob, pp. 4403–4408. https://doi.org/10.1109/SMC.2019.8914255.

Lohse, M., Garrido, L., Driver, J., Dolan, R.J., Duchaine, B.C., Furl, N., 2016. Effective connectivity from early visual cortex to posterior occipitotemporal face areas supports face selectivity and predicts developmental prosopagnosia. J. Neurosci. 36, 3821–3828. https://doi.org/10.1523/JNEUROSCI.3621-15.2016.

Nagy, K., Greenlee, M.W., Kovács, G., 2012. The lateral occipital cortex in the face perception network: an effective connectivity study. Front. Psychol. 3, 1–12. https://doi.org/10.3389/fpsyg.2012.00141.

Nguyen, V.T., Breakspear, M., Cunnington, R., 2014. Fusing concurrent EEG-fMRI with dynamic causal modeling: application to effective connectivity during face perception. Neuroimage 102, 60–70. https://doi.org/10.1016/j.neuroimage.2013.06.083.

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59, 319–330. https://doi.org/10.1016/j.neuroimage.2011.07.039.

Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comput. Biol. 6 https://doi.org/10.1371/journal.pcbi.1000709.

Rao, R.P.N., Ballard, D.H., 1999. Hierarchical predictive coding of natural images. Nat. Neurosci. 2, 79.

Rossion, B., 2008. Constraining the cortical face network by neuroimaging studies of acquired prosopagnosia. Neuroimage 40, 423–426. https://doi.org/10.1016/j.neuroimage.2007.10.047.

Sato, W., Kochiyama, T., Uono, S., Matsuda, K., Usui, K., Usui, N., Inoue, Y., Toichi, M., 2017. Bidirectional electric communication between the inferior occipital gyrus and the amygdala during face processing. Hum. Brain Mapp. 38, 4511–4524. https://doi.org/10.1002/hbm.23678.

Sergent, J., Ohta, S., Macdonald, B., 1992. Functional neuroanatomy of face and object processing. Brain 115, 15–36. https://doi.org/10.1093/brain/115.1.15.

Sladky, R., Spies, M., Hoffmann, A., Kranz, G., Hummer, A., Gryglewski, G., Lanzenberger, R., Windischberger, C., Kasper, S., 2015. (S)-citalopram influences amygdala modulation in healthy subjects: a randomized placebo-controlled double-blind fMRI study using dynamic causal modeling. Neuroimage. https://doi.org/10.1016/j.neuroimage.2014.12.044.

Sliwinska, M.W., Elson, R., Pitcher, D., 2020. Dual-site TMS demonstrates causal functional connectivity between the left and right posterior temporal sulci during facial expression recognition. Brain Stimul 13, 1008–1013. https://doi.org/10.1016/j.brs.2020.04.011.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025.

Stephan, K.E., Penny, W.D., Moran, R.J., den Ouden, H.E., Daunizeau, J., Friston, K.J., 2010. Ten simple rules for dynamic causal modeling. Neuroimage 49, 3099–3109. https://doi.org/10.1016/j.neuroimage.2009.11.015.

Sundström, S., 2010. Coding in multiple regression analysis: a review of popular coding techniques. Mathematics.

Team, R.C., 2020. R: A Language and Environment for Statistical Computing.

Wakeman, D.G., Henson, R.N., 2015. A multi-subject, multi-modal human neuroimaging dataset. Sci. data 2, 150001. https://doi.org/10.1038/sdata.2015.1.

Wang, Y., Metoki, A., Smith, D.V., Medaglia, J.D., Zang, Y., Benear, S., Popal, H., Lin, Y., Olson, I.R., 2020. Multimodal mapping of the face connectome. Nat. Hum. Behav. 4, 397–411. https://doi.org/10.1038/s41562-019-0811-3.

Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E., Windischberger, C., 2009. Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies. Neuroimage 47, 1408–1416. https://doi.org/10.1016/j.neuroimage.2009.05.005.

West, B., Welch, K., Gałecki, A., 2014. Linear mixed models, linear mixed models. https://doi.org/10.1201/b17198-2.

Williams, M.A., Savage, G., Halmagyl, M., 2006. Abnormal configural face perception in a patient with right anterior temporal lobe atrophy. Neurocase 12, 286–291. https://doi.org/10.1080/13554790601026379.

Zeidman, P., Jafarian, A., Corbin, N., Seghier, M.L., Razi, A., Price, C.J., Friston, K.J., 2019a. A guide to group effective connectivity analysis, part 1: first level analysis with DCM for fMRI. Neuroimage 200, 174–190. https://doi.org/10.1016/j.neuroimage.2019.06.031.

Zeidman, P., Jafarian, A., Seghier, M.L., Litvak, V., Cagnan, H., Price, C.J., Friston, K.J., 2019b. A guide to group effective connectivity analysis, part 2: second level analysis with PEB. Neuroimage 200, 12–25. https://doi.org/10.1016/j.neuroimage.2019.06.032.

Supplementary Material

Supplementary Methods

Paradigms

Paradigm A

To investigate brain connectivity during the observation of different basic emotions, a paradigm was designed to present image sequences of neutral, happy, angry, or fearful faces as well as houses in a block design. The face stimuli (39 individuals) were taken from the Radboud Face Database (RaFD, (Langner et al., 2010)), whereas the house stimuli were freely available pictures taken from the internet. The stimuli were transformed into gray-scale images and cropped to 500*400px. Furthermore, they were matched in mean luminance using the SHINE toolbox for MATLAB (Willenbockel et al., 2010). To avoid lateralization effects due to low-level image properties like asymmetries in the faces or houses, each image was mirrored vertically in one half of its appearances. A fixation cross was shown in the center of each stimulus, as well as during the inter-stimulus-intervals and inter-block-intervals in the center of the screen. The participants were advised to maintain the fixation of their gaze during the entire experiment. Each of the five conditions appeared 20 times, resulting in a total of 100 blocks. Within each block, a sequence of 24 images was shown. A face or house stimulus was shown for 350 ms followed by a fixation cross for 150 ms, resulting in a total block length of 11.85 s (main article, Fig. 2). A jittered inter-block-interval of approximately 3.3–7.3 seconds was introduced to reduce anticipation effects. Additionally, four pause-trials of 25 seconds were included, appearing after the 21st, 40th, 60th, and 80th stimulus block, in which the participants were instructed to relax and close their eyes. To ensure the attention of the participants during the stimulus presentation, a 1-back task was introduced. The participants were instructed to press a button with the index finger of both hands whenever a stimulus was shown twice in a row, which happened 1–3 times in each block. The total duration of the experiment was approximately 30 minutes.

Paradigm B

To investigate the connectivity between the core system of face perception and the amygdala during the observation of fearful faces, stimuli showing neutral faces, fearful faces, or houses were presented in a block design. The face stimuli (30 individuals) were taken from the Karolinska Directed Emotional Faces dataset (http://www.emotionlab.se/resources/kdef) (Lundqvist et al., 1998), whereas the house stimuli were freely available pictures taken from the internet. The stimuli were transformed into gray-scale images and cropped to 600*530px. Furthermore, they were matched in mean luminance using the SHINE toolbox for MATLAB (Willenbockel et al., 2010). The participants were advised to fixate their gaze on the nasion of the faces. For the house stimuli, they were asked to maintain their eyes at about the same height as that of the fixation for the face stimuli. Each of the three conditions (neutral faces, fearful faces, and houses) was repeated 14 times, resulting in a total of 42 blocks. Within each block, a sequence of 20 images was shown. The face or house stimulus was shown for ~310 ms followed by a fixation cross for ~390 ms, resulting in a total block length of 14.5 s. We used an inter-block-interval of ~6.5 s. Additionally, a pause-trial of 30 s in which the participants were instructed to relax and eventually close their eyes for a moment was included after half of the stimuli were presented. To ensure the attention of the participants during stimulus presentation, a 1-back task was introduced. The participants were instructed to press a button with the index finger of both hands whenever a stimulus was shown twice in a row, which happened 1–3 times in each block.

Paradigm C

The participants simply viewed blocks of faces, houses, and scrambled pictures. The pictures of houses and faces were obtained from a standardized database. The scrambled pictures were generated using a Fourier transformation. In total, 14 blocks of every condition appeared in a randomized order, including a 20 s break after half of the experiment. Every block contained 20 stimuli (faces, houses, or scrambled pictures), which were presented for 300 ms in the middle of the screen. Between each block, a fixation cross was shown for 12 s. The stimuli were controlled for brightness and contrast and presented in different grey scales. The participants were asked to push a button on a response box as

soon as they saw the same stimuli in immediate succession. Successive stimuli were presented 3–4 times in every condition. The total length of the paradigm was 13 min.

Paradigm D

See (Wakeman and Henson, 2015) for further information on this paradigm.

Paradigm FI

The authors conducted five experimental runs within one session, scanning 10 participants (Fairhall and Ishai, 2007). They presented line drawings of unfamiliar faces (1 run) and gray-scale photographs of unfamiliar (two runs), famous (one run), and emotional faces (fearful and happy, one run). Each image was presented for 3 seconds. As a visual baseline, scrambled versions of the stimuli were used. The stimuli were presented in a block fashion, with a duration of 36 s for the experimental condition and 24 s for the control condition. The experimental and control conditions were each presented thrice per run.

Data acquisition

Data set A

The MRI data were acquired using a 3.0-Tesla MR scanner (Siemens TIM Trio, Erlangen, Germany) with a 12-channel head matrix receive coil at the Core Unit Brainimaging, Department of Psychiatry and Psychotherapy, University of Marburg. A high-resolution structural data set was acquired using a T1-weighted magnetization-prepared-rapid gradient-echo sequence with the following parameters: acquisition time, 4 min 18 s; repetition time (TR), 1900 ms; echo time (TE), 2.52 ms; field of view, 256 mm; matrix, 256x246; slice thickness (ST), 1 mm; phase encoding direction (PE), anterior » posterior; distance factor (DF), 50 %; flip angle, 9°; parallel imaging generalized autocalibrating partially parallel acquisitions with acceleration factor, 2; bandwidth, 170 Hz/Px; sagittal, ascending acquisition; 176 slices.

Functional images were collected using a T2*-weighted gradient-echo echo-planar imaging sequence (EPI) sensitive to the blood oxygen level-dependent (BOLD) contrast. TR, 1550 ms; TE, 36 ms; matrix,

72x72; phase oversampling, 12 %; ST, 2.7 mm; DF, 15 %; voxel size, 2.8x2.8x2.7 mm (2.8x2.8x3.1 mm incl. gap); PE, anterior » posterior; flip angle, 70 °; bandwidth, 1654 Hz/Px; no parallel imaging; ascending acquisition; 20 slices with the measurement volume aligned to the anterior-posterior commissural line. The volume covered the whole temporal and occipital lobes, and the inferior frontal gyrus.

Data set B

Functional images were collected using a T2*-weighted gradient-echo echo-planar imaging sequence (EPI) sensitive to the BOLD contrast. The parameters were as follows: TR, 1610 ms; TE, 36 ms; matrix, 96 x 128; ST, 2.4 mm; DF, 15 %; voxel size, 2.0x2.0x2.4 mm; PE, anterior » posterior; flip angle, 70 °; bandwidth, 1346 Hz/Px; partial fourier, 7/8; no parallel imaging; ascending acquisition; 18 slices with the measurement volume aligned to the most ventral parts of the temporal and occipital poles. The slab covered the whole temporal and occipital lobes.

Data set C

All MRI data were acquired using a 3-Tesla TIM-Trio MR Scanner (Siemens Medical Systems) at the Department of Psychiatry and Psychotherapy, Philipps-University, Marburg. High resolution T1-weighted anatomical images were acquired from every participant (TE, 2.26 ms; TR, 1.9 ms; flip angle, 9°; matrix, 256 x 256; 176 sagittal slices; ST, 1 mm). To minimize head movements, the participants' heads were fixated with foam pads. Functional images were collected with a T2* weighted EPI sequence sensitive to the BOLD contrast (matrix, 64x64; field of view, 192 mm; 30 slices [descending]; ST, 4 mm [15% gap]; TR, 1450ms; TE, 25ms; flip angle, 90°). Slices covered the whole brain and were positioned in a transaxial parallel direction to the anterior-posterior commissural line. The BOLD responses to different experiments were recorded with this sequence. A total of 565 scans were recorded for the face perception task. The initial four images were excluded. Contrary to data set A and B, the measurement volume covered the whole cortex.

Data set D

See (Wakeman and Henson, 2015) for further information on this paradigm.

Functional imaging data analysis

Data set A

All the fMRI data sets were analyzed using the Statistical Parametric Mapping software (SPM12, release 6685, Welcome Department of Cognitive Neurology, Institute of Neurology, London, United Kingdom) based on MATLAB (version 8.3 R2014a). The initial three functional images were excluded from further analysis due to T1 stabilization effects, as implemented in the protocol of the MR scanner system. The field maps were calculated using the acquired phase and magnitude images from a field map sequence. These were converted to voxel displacement maps to unwarp geometrically distorted EPI images. In a combined realignment and unwarping step, the effects of static and movement-related susceptibility-induced distortions were corrected for, as well as within-participant motion correction through a rigid body (six parameters) spatial transformation. Each participant's functional images were also normalized to the Montreal Neurological Institute space. For an accurate transformation, each participant's T1-weighted image (coregistered to the mean functional image) was segmented, bias-corrected, and spatially normalized using the segmentation algorithm as implemented in SPM12 (formerly called "New Segment" in SPM8). The resulting forward deformation field was used for registering the realigned functional images to the Montreal Neurological Institute space that were subsequently resampled to a resolution of 2 x 2 x 2 mm^3 and blurred with an isotropic Gaussian filter of 6mm full width at half maximum.

Statistical analyses were performed within a general linear model framework to create a 3-dimensional map in relation to the estimated regressor response amplitude. At the single-participant level, the task was modeled in a block design with BOLD responses for each condition (neutral, happy, angry, and fearful faces as well as houses, respectively) convolved with the canonical hemodynamic response function. Inter-block intervals and breaks were not modeled. The six realignment parameters of the motion correction procedure were included in the statistical model as nuisance regressors to correct for residual head movement. For each participant, differences in brain activation between the 'face' and control conditions were calculated. High pass filtering was applied with a cut-off frequency of

130 1/256 Hz to attenuate low-frequency components. At the group level, the weighted ß-images ('face'
131 conditions vs. control condition) were entered into one-sample t-tests.

132 Data set B

133 The preprocessing and statistical analysis of data set B were performed in the same way as those for
134 data set A, but using MATLAB version 7.8 R2009a and a high-pass filter of 1/128 Hz.

135 Data set C

136 All the fMRI data were analyzed with the software package SPM 8 (v4290) (www.fil.ion.ucl.ac.uk/spm)
137 using standard routines and templates running on MATLAB 7.7.0.471 (R2008b) (The MathWorks, Inc.).
138 SPM 8 was used for realignment, normalization, smoothing, and statistical analysis. The functional
139 images were realigned, normalized to a resulting voxel size of 2 x 2 x 2 mm^3, smoothed with a 5-mm
140 isotropic Gaussian filter, and high-pass filtered by a cut-off period of 128 s. After preprocessing, the
141 statistical analysis was performed. BOLD responses for the test and control conditions were modeled
142 by a boxcar function convolved with the canonical hemodynamic response function employed by
143 SPM8. Parameter estimates (β-) and t-statistic images were calculated, describing the activation
144 differences between the test and control conditions. For the face perception tasks, three conditions
145 were modeled (faces, houses, scrambled images; the instruction was not modeled). Additionally, the
146 six realignment parameters were included as covariates in each design matrix. At the group level, the
147 weighted ß-images were entered into one-sample t-tests. The "2* faces > (houses + scrambled)"
148 contrast, describing the differences in the patterns of brain activation between the activation and
149 control conditions, was calculated for each participant.

150 Data set D

151 See (Wakeman and Henson, 2015) for further information on the processing steps.

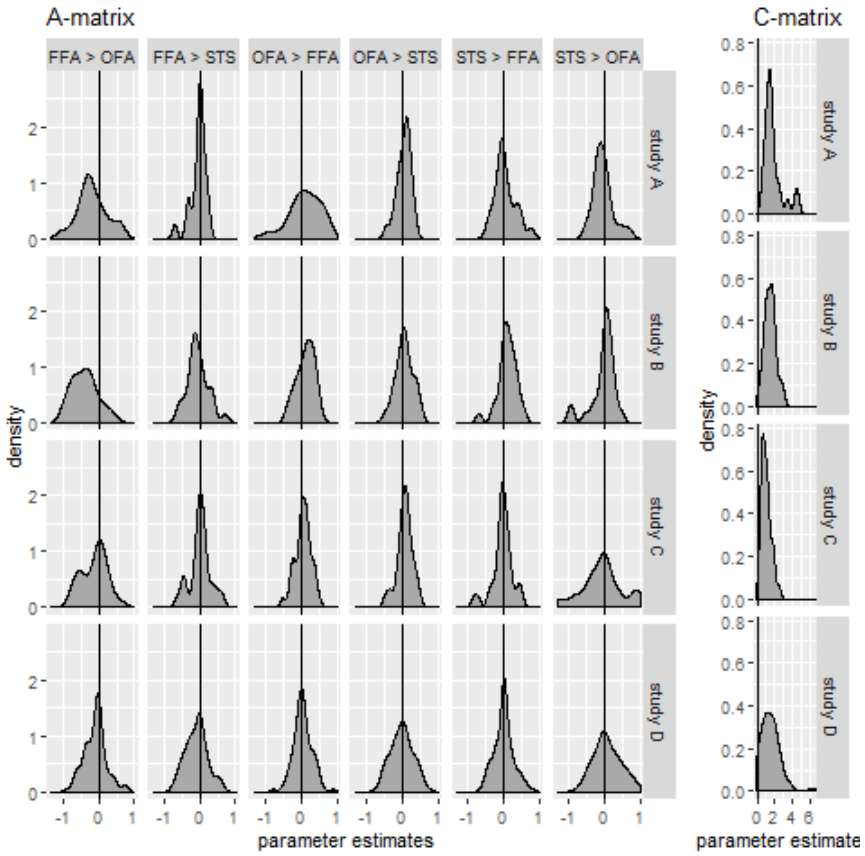
152 Data set FI

153 See (Fairhall and Ishai, 2007) for further information on the processing steps.

Supplementary Results

Single participant BMA parameters

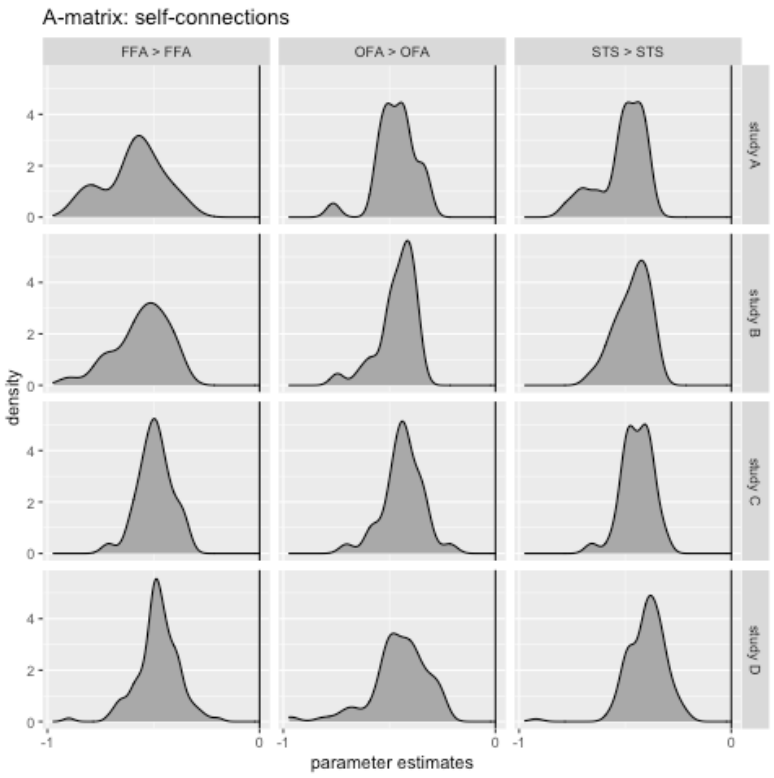
Here, we describe the distribution of the single participant BMA parameters. For the A-matrix (Fig. S1, left panel), we found relatively small parameter estimates ranging from approximately -1 to +1. Most distributions appeared roughly symmetrical and centered around zero. For the C-matrix (Fig. S1, right panel), the parameters were mostly positive and ranged from -0.058 to +6.626. Positive values are expected here, as we extracted the time series for the DCM construction from the voxels that exceeded significance in the contrasts (e.g., 'faces' conditions vs. 'non-face' conditions, depending on the respective study). Therefore, in these voxels, the activation should be higher during the 'faces' conditions, which is consequently a positive input. The connectivity parameters for the 'face' condition (B-matrix, Fig. S3) were spread more widely, i.e., from approximately -4 to +4. The distributions of a parameter over participants indicated that some parameters shifted more into the positive or negative range, respectively. Connectivity parameters for the 'emotions' and 'fame' conditions (B-matrices, Fig. S3) seemed to be more symmetrically accumulated around zero compared to those for the 'faces' conditions.



170

171 **Fig. S1*:* Kernel density estimates of all the connectivity parameters of the interregional (i.e., off-diagonal) connections of**

172 **the A-matrix (left panel) and C-matrix (right panel) after participant-specific BMA.**

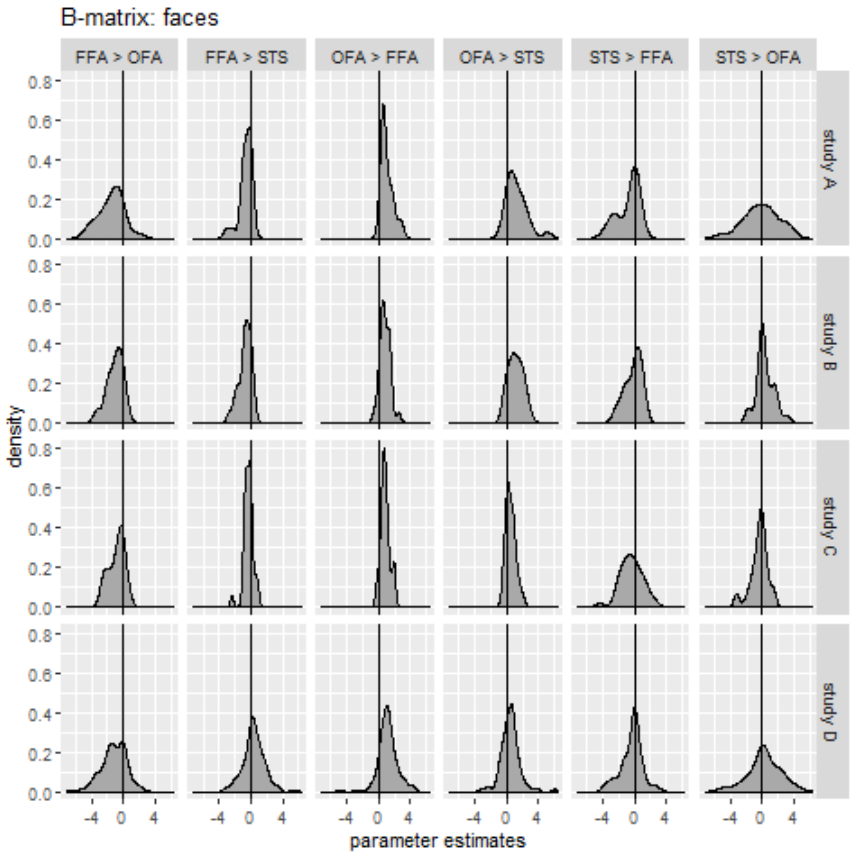


173

174 **Fig. S2*:* Kernel density estimates of all the connectivity parameters of the self-connections (i.e., on-diagonals) of the A-**

175 **matrix after participant-specific BMA.**

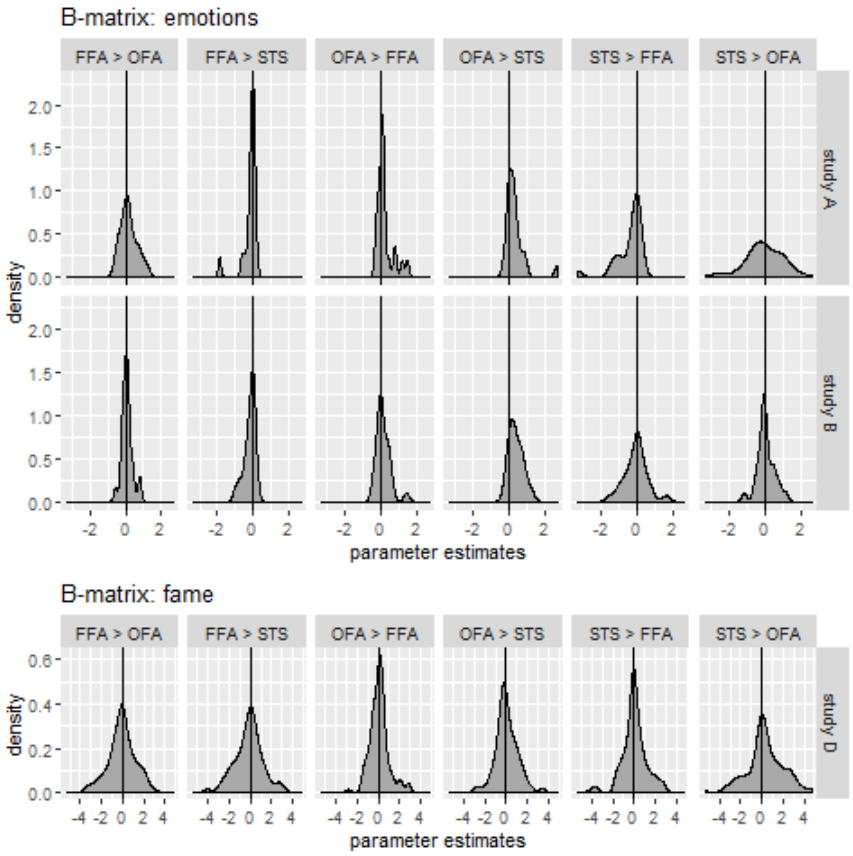
176



177

178 **Fig. S3: Kernel density estimates of all the connectivity parameters of the B-matrix 'faces' after participant-specific BMA.**

179



180

181 **Fig. S4: Kernel density estimates of all the connectivity parameters of the B-matrix 'emotions' (top panel) and 'fame' (lower**

182 **panel) after participant-specific BMA.** Note the different scaling of the axes for the regressors 'emotions' and 'fame'.

183

Hierarchical linear modeling results

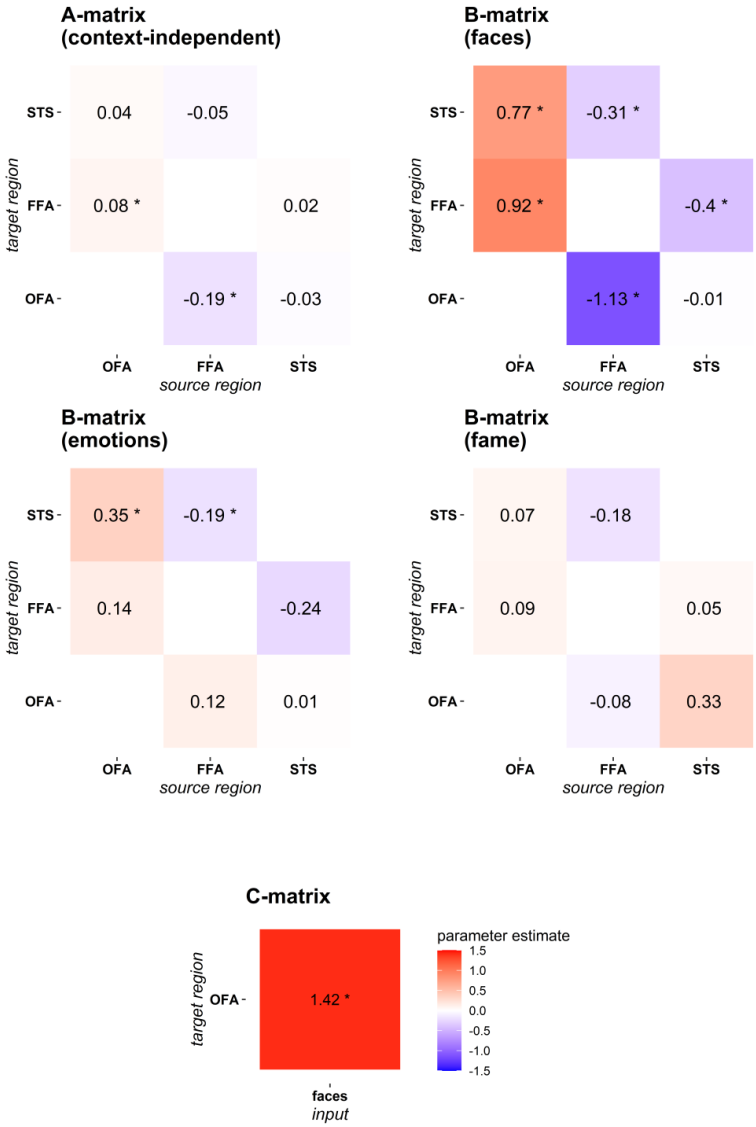
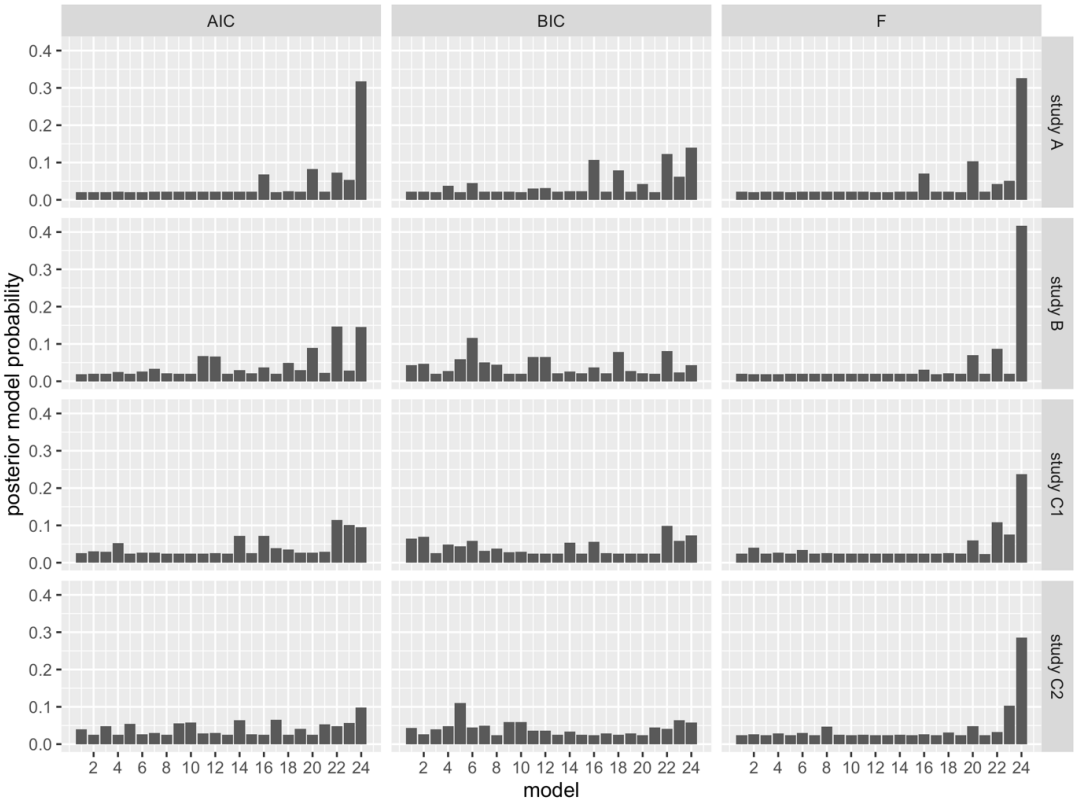


Fig. S5*:* Connectivity parameters across studies as revealed by HLM. For each matrix (A-, B-'faces'-, B-'emotions', B-'fame', and C-matrix), the connection strength is displayed using pseudo colors ranging from -1.5 (blue) to +1.5 (red), as indicated by the color bar at the bottom right and the values in each cell. Significance is indicated by additional asterisks. The source and target regions are labeled on the x- and y-axis of each matrix, respectively. For the C-matrix, there is no source region, but the 'faces' regressor targets the OFA. The significant parameters of this figure are further displayed in a model-like manner in **Error! Reference source not found.**.

Alternative approximations of the log model evidence

Since the publication of SPM8, free energy (F) is the preferred choice as an approximation to the log model evidence. Before (i.e., up to SPM5), Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been largely used for this purpose. The authors of study FI used both criteria as well (Fairhall and Ishai, 2007). In our study using the free energy criterion, model #24 showed striking superiority to the competing models throughout all examined data sets (**Error! Reference source not found.**, **Error! Reference source not found.**). Study FI, which applied the AIC and BIC, found that a sparser model was superior to the competing models (Fig. 3, Fairhall and Ishai, 2007). Some of the advantages and possible disadvantages of the free energy criterion have been outlined in Section 4.2. We hypothesized that the choice of the free energy criterion may largely drive differences in posterior model probabilities, and therefore the winning models between study FI and our study. To test this, we repeated the BMS for study A, B, and the two sessions of study C (C1 and C2) by applying AIC, BIC, and F separately. The results are illustrated in Fig. S6. Using AIC, either model #22 or model #24 have the highest exceedance probabilities, depending on the data set analyzed. Model #22 misses one connection compared to the full model #24, the unidirectional connection from FFA to STS (the corresponding opposite connection is however present, **Error! Reference source not found.**). Using BIC, either model #5, model #6, model #22, or model #24 have the highest exceedance probabilities. Model #6 does not express backward connections from FFA to OFA and from STS to OFA. Model #5 further misses the unidirectional connection from STS to FFA (the corresponding opposite connection is however present, **Error! Reference source not found.**). Altogether the BMS was highly variable across data sets when using AIC/BIC. Most strikingly, none of the single BMS results corresponded to the BMS result of study FI, which found model #2 to have the highest posterior probability.


AIC
BIC
F
study A
study B
study C1
study C2
posterior model probability
0.0
0.1
0.2
0.3
0.4
2 4 6 8 10 12 14 16 18 20 22 24
model

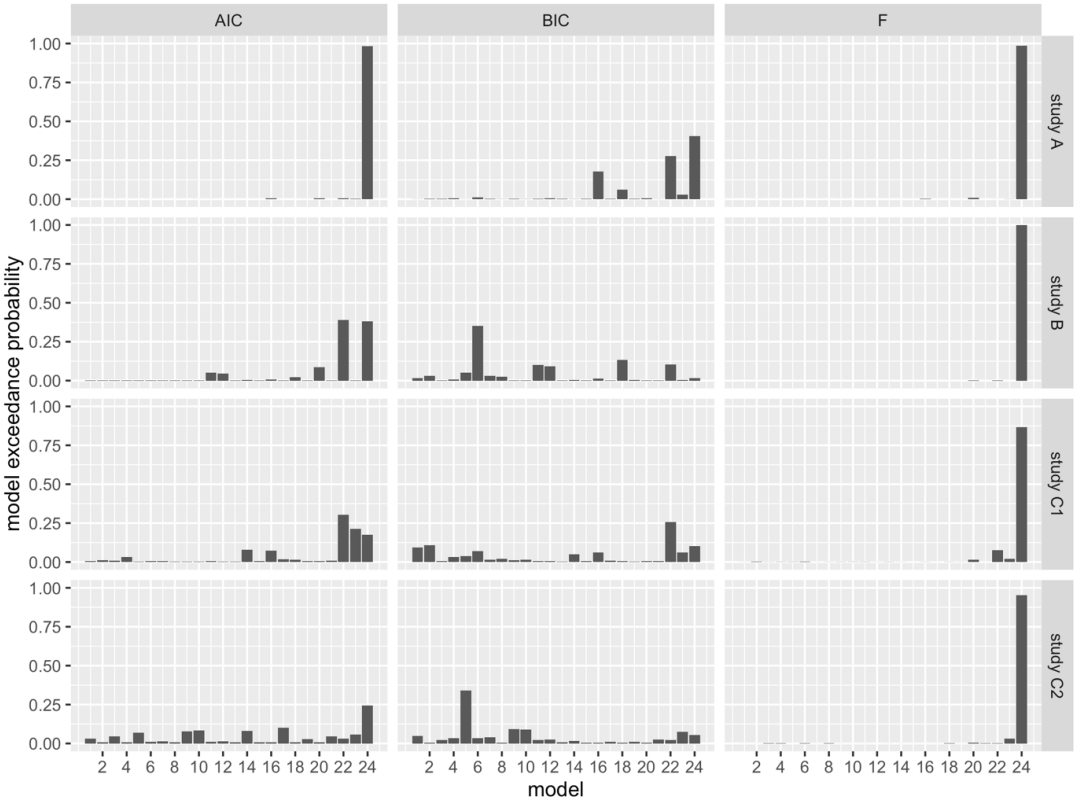

AIC
BIC
F
study A
study B
study C1
study C2
model exceedance probability
0.00
0.25
0.50
0.75
1.00
2 4 6 8 10 12 14 16 18 20 22 24
model

Fig. S6: BMS results using different approximations for log model evidence. Upper panel: posterior model probabilities. Lower panel: Model exceedance probabilities. The different data sets (A, B, C1, and C2) for which a BMS was possible are separated along the vertical axes. The BMS results of these studies are separated for AIC, BIC, and F as information criteria along the horizontal axis. The panels for F correspond to the results displayed in **Error! Reference source not found.**.

Fairhall, S.L., Ishai, A., 2007. Effective connectivity within the distributed cortical network for face perception. Cereb. Cortex 17, 2400–2406. https://doi.org/10.1093/cercor/bhl148

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., van Knippenberg, A., 2010. Presentation and validation of the radboud faces database. Cogn. Emot. 24, 1377–1388. https://doi.org/10.1080/02699930903485076

Lundqvist, D., Flykt, A., Ohman, A., 1998. The Karolinska directed emotional faces (KDEF). CD ROM from Dep. Clin. Neurosci. Psychol. Sect. Karolinska Institutet.

Wakeman, D.G., Henson, R.N., 2015. A multi-subject, multi-modal human neuroimaging dataset. Sci. data 2, 150001. https://doi.org/10.1038/sdata.2015.1

Willenbockel, V., Sadr, J., Fiset, D., Horne, G., Gosselin, F., Tanaka, J., 2010. The SHINE toolbox for controlling low-level image properties. J. Vis. https://doi.org/10.1167/10.7.653

Appendix C

Study 3

How function is bound by structure in models of effective connectivity

Roman Kessler[1,2,*], Andreas Jansen[1,2,3]

1 Laboratory for Multimodal Neuroimaging (LMN), Department for Psychiatry, University of Marburg, Germany
2 Centre for Mind, Brain, and Behaviour (CMBB), University of Marburg and University of Giessen, Germany
3 Core-Unit Brainimaging, Department for Psychiatry, University of Marburg, Germany

*** rkesslerx@gmail.com**

Abstract

Dynamic Causal Modeling (DCM) is a widely used method to analyze effective connectivity between brain regions based on functional magnetic resonance imaging (fMRI) data. The major advantage of DCM over other connectivity analyses is mainly the putative possibility to make statements about directions of information transfers. Once one is familiar with the rough basics, the application of the method is relatively straightforward and allows - also due to the integration of the analysis steps into a simple graphical user interface - a broad mass of researchers to apply this analysis method to their own dataset.

In this article we want to dispel some misconceptions about the interpretability of DCM parameters. In particular, we want to show that many parameters of a stereotypical DCM model are not interpretable, using real measured but also simulated fMRI data. More specifically, we show that within these experiments, much of the qualitative output of the DCM model is determined solely by the way it is modeled. As a result, their expression (e.g., whether the parameters become positive or negative) is determined solely by the structure of the model, and does not have much to do with the cognitive processes that one would like to interpret into the model parameters.

Among other things, we show that forward connections almost always become positive, while backward connections almost always become negative. We provide simple but valid evidence for this pattern. The results emphasize the following: by pure plausibility considerations, a preponderance of the expressions of connections - i.e., whether they are estimated to be positive or negative - can already be predicted before an experiment is even conducted. Thus, only a few connectivity parameters remain for free interpretation. We argue that the explanatory power of a DCM model is therefore severely limited.

Introduction

Modeling the interactions between brain regions can offer invaluable insights on brain function in health and disease. Numerous techniques have been developed over the past decades, aiming at quantifying information transfer between cortical regions based on functional magnetic resonance imaging (fMRI) data. Some techniques offer insights on correlative basis (e.g., seed-based correlation, independent component

analysis) – and are termed measures of *functional connectivity* [25]. Other techniques however promise to disentangle causal interactions between cortical regions (e.g., psychophysiological interaction, structure equation models, Dynamic Causal Models) – and are termed measures of *effective connectivity* [12]. One very popular framework of effective connectivity is Dynamic Causal Modeling (DCM, [7]). DCM at the one hand aims at determining the nature of neural connections, i.e., whether it is excitatory (positive) or inhibitory (negative). On the other hand, DCM aims at estimating the connection magnitudes either during particular cognitive computations [7] or at rest [23].

The core part of the vanilla DCM implementation is the neural state equation (eq. 1), which describes the rate of change in connectivity $\dot{z} \in \mathbb{R}^n$,

$$\dot{z} = Az + (\sum_{j=1}^{k} u^{(j)} B^{(j)})z + Cu \tag{1}$$

with n being the number of modeled brain regions. The rate of change is assembled as a sum of several components. First, the product between a binary onset-offset vector u multiplied with the magnitude of the experimental input $C \in \mathbb{R}^{n \times k}$. Second, the product of the magnitudes of connections between regions $A \in \mathbb{R}^{n \times n}$ and the current state (i.e., activation) $z \in \mathbb{R}^n$ of the respective regions. Third, connection between regions can be further modulated by experimental perturbation, which is reflected in a sum across k different experimental manipulations, for each of which the magnitudes of connections specific to the respective experimental manipulation $B^{(j)} \in \mathbb{R}^{n \times n}$ is multiplied with the current states z and also $u^{(j)}$ as a binary onset-offset vector specifying the presence or absence of the respective experimental manipulation at a particular timepoint.

The neural state equation is the point above which a scientist typically defines the structure of their model, by allowing (i.e. *enabling*) or not allowing (i.e. *disabling*) particular cells of the A, B and C matrices to be estimated. Other parts such as the input vectors u are mostly given by the setup of the (e.g., fMRI) experiment. Having agreed for the particular connections which they wants to be estimated, a sequence of steps is rather generic for the remaining modeling procedure. This includes the specification of the prior means, which are all set around 0. Next, the prior variances are set to 0 for all connections, which were 'disabled', and to 1 for 'enabled' connections of the B and C matrices, or to $\frac{1}{64}$ for 'enabled' connection within the A matrix [28, 29]. The modeled neural states z are then translated into a hemodynamic signal using an empirically inspired forward model [3] with its own internal priors [29]. The neural state equation alongside with the forward model is then iteratively estimated using a variational Laplace approximation [8].

A straightforward experiment using fMRI and DCM includes a activation condition (e.g., visual input, or faces) and a control condition (e.g., void screen, or houses). In many cases, the experimental condition activates all of the modeled brain regions, whereas in the control condition these regions are less active (e.g., see [6, 11, 14, 15, 18, 26]). Next, one or several DCMs are constructed encompassing those very regions. Those DCM then gets fitted to disentangle *causal* interactions between the brain regions, for instance to quantify the information transfer due to a neural process such as processing a face.

In the present work, we will illustrate, that a researcher using DCM can infer many of the resulting model parameter estimates without even estimating the model, by plain plausibility considerations on the level of the neural state equation (eq. 1). Critically, by being able to anticipate many of the outcomes of an DCM estimation process – i.e., the expression of parameter estimates of the A, B and C matrices – the interpretability of these parameter estimates simply cease to apply.

Motivating example

In the following illustrations we ignore the non-linearities in the signal induced by the hemodynamic forward model to keep it traceable, and we instead stay at the neuronal rather than the hemodynamic level to describe interactions between cortical regions. The introduced non-linearities of the hemodynamic forward model would not change any of the qualitative conclusions we draw from the plausibility considerations, but keep the argumentation more straightforward.

We will illustrate the general problem with the following example. Picture a model comprising two regions. The fMRI experiment conducted consisted of a task condition and a control condition. Both modeled brain regions are activated by the task condition, but not by the control condition. Both regions are reciprocally interconnected, and each region has a (per definition negative) self connection, leading to a fully connected A matrix (eq. 2) where τ denotes a parameter with non-zero variance (i.e., enabled), and 0 (not present) would denote a parameter with zero prior variance (i.e., disabled). Further, the indices of each element of an A or B matrix (e.g. $a_{i,j}$) denote the index of the target region i and the source region j of the connection.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} = \begin{pmatrix} \tau & \tau \\ \tau & \tau \end{pmatrix} \tag{2}$$

No B matrix is specified in this example. We further set the experimental input (i.e., a region gets targeted by an experimental input, meaning the experimental input directly changes the regions state over time) to the first region, resulting in the following C matrix (eq. 3):

$$C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \tau \\ 0 \end{pmatrix} \tag{3}$$

Having specified A and C matrices, and rearranged the neural state equation (eq. 1), we can describe the rate of change in activity of region 1 and region 2 by equations 4 and 5, respectively:

$$\dot{z}_1 = c_1 \cdot u + a_{1,1} \cdot z_1 + a_{1,2} \cdot z_2 \tag{4}$$

$$\dot{z}_2 = a_{2,1} \cdot z_1 + a_{2,2} \cdot z_2 \tag{5}$$

In the following, we will disentangle the different model terms of equations 4 and 5: The current activation $z_\cdot$ is always non-negative. The self connections $a_{1,1}$ and $a_{2,2}$ are always negative per definition. Region 1 gets activated by the task, as the experimental condition leads to higher activity in the modeled regions (see definition of the experiment). The input parameter c_1 therefore has to become positive to excite the system and to render $\dot{z}_1$ and consequently z_1 positive (eq. 4). Next, due to the fact that region 2 is more active during task condition than control condition, region 2 needs a positive slope of activation $\dot{z}_2$. The only non-negative term in eq. 5 encompasses the connection from region 1 to region 2 $a_{2,1}$. Therefore, $a_{1,2}$ needs to be positive, to render $\dot{z}_2$ positive.

Determined by the activation properties of the both regions (i.e., both regions get activated by the experimental task), and by the connectivity between the regions, we were able to already infer that two major parameters c_1 and $a_{2,1}$ should be positive. Parameters $a_{1,1}$ and $a_{2,2}$ are negative by definition, and the only still to be inferred parameter would be parameter $a_{1,2}$. Therefore, the remaining three parameters are of minor value for a range of interpretations, as their quality (i.e., sign) is already predetermined.

In the present article, we will demonstrate this very effect using real and artificial fMRI data. Furthermore, we will demonstrate related effects, which again can vastly

be inferred by plausibility considerations based on the neural state equation. Those effects are anything but rocket science. However, we argue that most researchers generously interpret the quality (i.e., sign) of their DCM parameters, while this quality should be highly predictable by the model structure, assuming the empirical time series are not dominated by noise. The purpose of the present article is to show the user of such models, that given these predictabilities, those very models are in some cases not useful tools to answer given research hypotheses.

Material and Methods

In the following, we will describe the principle data and experimental procedures which will form the basis of all subsequent analyses. Afterwards, we will assemble Methods and Results individually for each claim about the behavior of DCMs. Each claim will be evaluated on empirical data from a prototypical fMRI experiment (e.g., visual input versus rest) in the early visual cortex. Furthermore, the claims will be supported by simulations, in which the ground truth (i.e., the model which generated the data) is well known.

Experimental data

For the purpose of this study, we decided to use a rather simple fMRI data set. The only prerequisites were an experimental and a control condition. We then aimed at identifying regions, which were more active during experimental conditions than during control conditions. Therefore, we decided for a simple visual stimulation paradigm, freely available in the internet (`openneuro.org/datasets/ds001553`) [10].

The data set comprises three healthy, right-handed, female participants (25, 28, and 30 years), each measured around 100 times with the identical visual checkerboard paradigm during a full k-space echo planar imaging sequence (time of repetition (TR): $2.0s$, echo time (TE): $0.03s$) in a $3T$ MRI scanner [10]. We only made use of the first participant's data (25 years), comprising 105 functional runs. Each run began with 30 seconds of rest, followed by 5 repetitions of alternating task blocks (20 seconds each) and rest blocks (40 seconds each), and ending with an additional rest block of 10 seconds. During rest blocks, a black screen with a fixation cross was displayed. During task blocks, a checkerboard stimulus was flickering at a frequency of $7.5Hz$. To control for attention and ensure compliance, participants were instructed to discriminate between letters and numbers with corresponding button presses. Letters and numbers were pseudo-randomly displayed in the center of the screen [10].

fMRI data processing

The downloaded raw data (`openneuro.org/datasets/ds001553`) was processed using SPM12 (r7771, `fil.ion.ucl.ac.uk/spm`) for Matlab (R2020b, The MathWorks Inc.). All functional images were realigned and cross-registered across sessions using a 6 parameter rigid body registration. Then, all functional images were normalized to MNI space using a 12 parameter affine transformation. No additional spatial blurring was performed. We discarded the first 5 volumes to allow for magnetic stabilization. Thereafter, we fitted a general linear model (GLM) using the onset-offset vectors of the checkerboard stimulation and convolved it with a hemodynamic response function for each session separately. Mass-univariate one-sample t-tests were performed for each voxel to reveal brain activation induced by visual stimulation.

Region identification

Our aim was to model interactions between regions. All regions were activated by the experimental conditions, compared to the control condition. Such a pattern should clearly be identifiable in early visual areas [10]. We used anatomical masks to narrow down early visual cortical regions. We decided for a mask for Brodmann area 17 (BA17), Brodmann area 18 (BA18), and Brodmann area 19 (BA19, Fig. S1), extracted from the Talairach atlas [16,17] implemented in *nilearn* (`nilearn.github.io`) [1]. BA17 is supposed to encompass cortical region V1, BA18 is supposed to encompass V2, and BA19 is supposed to encompass V3, V4, and V5 [4,24]. We extracted time

series within the three disjunct masks as the first principle component of all voxels within each mask at a statistical threshold of $p < 0.001$ (uncorrected for multiple comparisons). Untypical to many other DCM studies, the regions were spatially closely aligned, and we did not limit the voxels to a sphere with small radius (see e.g. [9, 14, 15, 26]). The extracted time series of all modeled brain regions increased following the start of the experimental stimulation, and decreased after stimulation ended (Fig. 1).

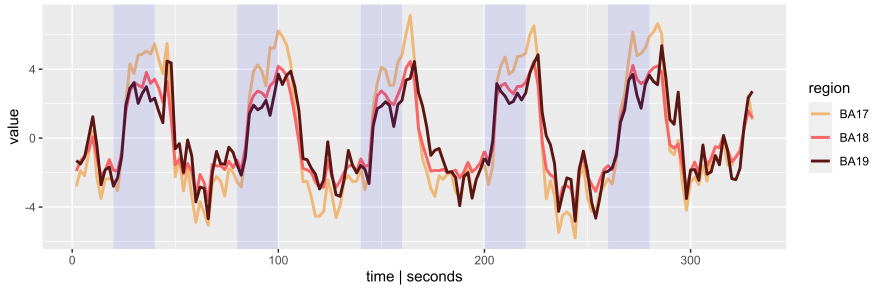


Figure 1. Exemplary time series of functional run #1. The extracted first principle component for each region is displayed. Hemodynamic activation (mean-centered) is displayed on the y axis (arbitrary numbers), and the time (1 TR = 2.0 seconds) on the x axis. The shaded regions mark time intervals of visual stimulation., i.e., in which the onset-offset vector shapes to 1. The hemodynamic pattern is the same throughout all functional runs: Activity increases shortly after stimulation onset, and decreases shortly after stimulus offset.

Modeling

We then constructed DCMs encompassing three regions for each of the 105 functional runs (Fig. 2). All models share several common properties:

- within the A matrix, BA17 was connected to BA18, and BA18 was connected to BA19;
- beginning from an input region (red), all downstream regions were targeted by at least one forward connection along this route;
- within the A matrix, a self connection (per definition negative) was deployed on each region (i.e., on-diagonal elements of the matrix).

For instance, the prior variances of the A matrix of model 'A1' (Fig. 2) were assembled as displayed in equation 6,

$$Var_{prior}(A) = \begin{pmatrix} \frac{1}{64} & \frac{1}{64} & 0 \\ \frac{1}{64} & \frac{1}{64} & \frac{1}{64} \\ 0 & \frac{1}{64} & \frac{1}{64} \end{pmatrix} \tag{6}$$

The regions comprise BA17, BA18, and BA19. Source regions are represented in columns, and target regions in rows of matrices A and B. Therefore, parameter $Var_{prior}(A_{1,2})$ represents the A matrix (in this case: prior variance) originating at region 2, targeting region 1. Priors variances of enabled connections are then set to either 1 or $\frac{1}{64}$, depending on the matrix type (see above), and set to 0 for disabled

connections. Throughout this article, the anatomical regions BA17, BA18, and BA19 are used interchangeable with the terms region 1, regions 2, and region 3.

Naïvely, a biophysically plausible model could be constructed in a hierarchical fashion [22]. The experimental input would enter the system via BA17, and then gets propagated to BA18 and BA19, with reciprocal connections between regions (Fig. 2, model 'A1'). To tackle some questions about information transfer between regions, such a base model could easily be accepted by the reader community. However, to test different hypotheses about the behavior of a model, we systematically varied the definition of the A, B, and C matrices, as visualized in Figure 2. For an better understanding of the graphical representations of the models used in Figure 2 and for the terms used to describe particular regions and connections over the curse of this article, we further refer to Figure S2.

In contrast to the commonalities of all models, the following model parameters were varied systematically:

- input either to BA17 ($Var_{prior}(C = \begin{pmatrix}1\\0\\0\end{pmatrix}$), BA18 ($Var_{prior}(C) = \begin{pmatrix}0\\1\\0\end{pmatrix}$), or BA19 ($Var_{prior}(C) = \begin{pmatrix}0\\0\\1\end{pmatrix}$);
- experimental modulation of connections (i.e., B matrix parameters) either enabled or disabled. If disabled, the prior variances of the B matrix parameters comprised only zeros. If enabled, the prior variances in the B matrix has non-zero prior variance at the same off-diagonal matrix elements as the A matrix (i.e., interregional connections), but with prior variances of 1 instead of $\frac{1}{64}$, and contrary to the A matrix comprises prior variances of 0 on all on-diagonal elements (i.e., self connections);
- BA17 and BA19 either not connected, or also reciprocally connected;
- regions were either reciprocally connected or only unidirectionally connected, so that only forward connections were enabled

By those variations in our 3-region DCM, we estimated $3 \times 3 \times 2$ models (Fig. 2) for each experimental session. We analyzed posterior model parameters to verify different claims about their quality (i.e., positive or negative). Therefore, we extracted both posterior means and variances of each single parameter estimate of all models and connections across functional runs and analyzed them further. For self connections of the A matrices (e.g., $a_{1,1}$) we transformed the estimated parameters to unit Hertz to be consistent with the interregional parameters (eq. 7).

$$a_{Hz} = -\frac{e^{a_{log}}}{2} \tag{7}$$

In contrast, for self connections of the B matrices (e.g., $B^{(1)}_{1,1}$) we did not convert to unit Hertz, for ease of interpretation (see [28] for more details). Consequently, a more negative parameter estimate (in log scale) of the self connection of the B matrix indicates a weaker self-inhibition, i.e., shifting the total self inhibition towards zero Hertz. In contrast, a more positive parameter estimate (in log scale) of the self connection of the B matrix indicates a stronger self-inhibition, i.e., shifting the total self inhibition (in Hertz) stronger negative.

Simulations

One might argue, that the models of e.g., the leftmost column of Figure 2 can be biophysically motivated to some degree. Unfortunately, the ground truth, i.e., the

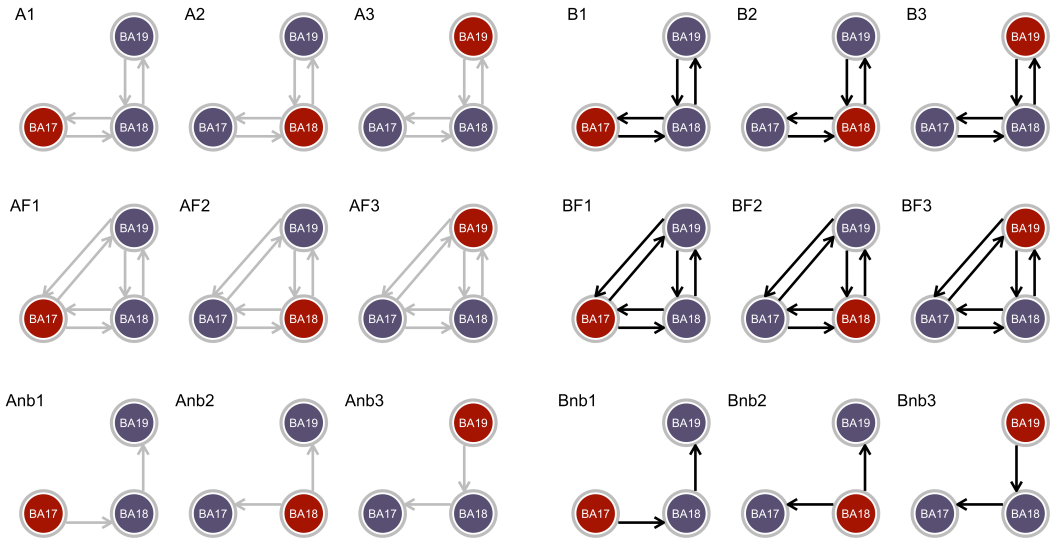


Figure 2. Overview of constructed models. Three regions were modeled, BA17 (region 1), BA18 (region 2), and BA19 (region 3). The input region (C matrix) is colored in red, whereas non-input regions are colored in purple. For each type of model (A, B, AF, BF, Anb, and Bnb), the input region is systematically varied. Grey arrows depict parameters modeled in the A matrix only (i.e., only non-negative prior variances are allowed within the A matrix). Black arrow depict parameters modeled in the B matrix, which requires the same parameter also being modeled within the A matrix. Grey circles around a region represent the (negative) self connection modeled within the A matrix. See also Figure S2.

model that generated the data we observe, is not known. Therefore, any validity of inference about the ground truth model is – among other things – bound by the limited world of the tested hypotheses (i.e., the model space). To partially finesse the fact that the ground truth model is unknown, we simulated hemodynamic time series. Therefore, we constructed several – in our opinion – plausible ground truth models, generated the neural and hemodynamic time series, added some noise, and then continued the modeling procedure similar to above, by constructing and estimating DCMs.

To be able to generate plausible hemodynamic time series, we needed plausible parameters for the data generating model in the first place. To obtain such, we calculated posterior parameter estimates for some of the biophysically most plausible models, for example for model 'A1' (Fig. 2, top left) from all 105 experimental sessions. We then averaged the parameter estimates per connection. The average estimates formed the ground truth model, which was then used to generate hypothetical time series. The generation of the hypothetical time series was repeated 100 times to simulate 100 sessions. Furthermore, the signal-to-noise ratio (SNR) was systematically varied to be either 1, 3, or 5. The generated time series were in turn used to feed and estimate other models, for instance models 'A2' or 'B1' depending on the respective research question. See Figure S3 for an exemplary simulation procedure.

This detour was motivated by two reasons. First, the shape of the data generating models were not completely far-fetched, but were based on some empirical motivations. Second, we needed the simulations to draw conclusions in a hypothesis space, where

the ground truth is known. By having time series generated by a 'true' synthetic model (e.g., 'A1'), and by using these times series to estimate different models (e.g., 'B1'), we can attribute the differences in resulting model parameters rather to the perturbations in model structure, than to the likelihood of the model in the first place, because the shape of the model which generated the data is known.

Statistical inferences

To test hypotheses about the resulting DCM parameter estimates, we applied several statistical tests. In cases in which we tested if parameter estimates across sessions differed from zero, we deployed two-tailed one sample t tests for each posterior parameter estimate of a particular matrix. To test the difference between two distributions, e.g., to compare the parameter estimates of an A to a B matrix, or the respective A matrices of two kinds of models, we used paired sample t tests.

To adjust for multiple comparisons, we applied Benjamini-Yekutieli correction for false discovery rate (FDR) [2]. We particularly used Benjamini-Yekutieli approach to account for arbitrary interdependencies between tests, i.e., interdependencies between different connections of a model. We applied a corrected threshold of $\alpha_{BY} = 0.05$, and corrected for the number of tests. For example, the number of tests is 10 when testing if each kind of connection per model (i.e., forward $\times 3$, backward $\times 3$, downstream forward $\times 2$, downstream backward $\times 2$) of model space 'A' is different from zero. The number of tests is 20, when testing the same for model space 'B', or 9 respective 18 when testing the same for model spaces 'AF' respective 'BF' (see Fig. 3,5,9).

1 Forward connections are positive

1.1 Hypothesis

Connections leading away from the input region (i.e., forward connections and downstream forward connections) are always characterized by positive parameters. If not, downstream regions showing a similar time course (i.e., being activated in the same contrast) could not be upregulated.

Therefore, when the input is given to a different region, all connections within the models are adjusted so that they always follow the same pattern. Namely, all forward connections become positive. This applies both to the direct forward connection starting from the input region and to the subsequent (i.e., downstream) forward connections starting from the next region through which the signal travels (Fig. S2). This would continue until all regions have been activated by at least one positive (forward) connection.

1.2 Methods

To test the hypothesis, we created three models. The models were identical except for the input region. We expected in all models both positive forward connections and positive downstream forward connections.

Model structure: Each model included three regions (BA17, BA18, and BA19). *A* matrix: Endogenous connections were set bidirectionally between BA17 and BA18 and between BA18 and BA19. *B* matrix: No modulatory connections were allowed in any model. *C* matrix: Exogenous input was modeled by an 'activation' regressor (modeling activation task) and was set to BA17 (model 'A1'), BA18 (model 'A2') and BA19 (model 'A3'), respectively (for an overview, see Fig. 2). Model parameters were estimated both for real data and for synthetic data. For synthetic data, we generated data from model 'A1' (see Fig. S3 for more details).

For illustration, we classified the model parameters into different groups, forward, backward, downstream forward, downstream backward, and self connections according to the rationale outlined in Figure S2. Each connection of a model was classified into one of these five groups and then evaluated together. We tested significant (FDR corrected) differences from zero by one sample t tests.

1.3 Results

As expected, forward connections and downstream forward connections were always positive (Fig. 3 for real data, Fig. S4 for synthetic data). The sign of the endogenous parameters was therefore not dependent on the specific data, but on the structure of the model. In other words, already the choice of the input region was sufficient to determine the sign of the endogenous forward and downstream forward connections of all models. Interestingly, parameter estimation also revealed that the backward connections as well as the downstream backward connections were always negative (all $p << 0.001$). This pattern was again present in all models, independent of where the experimental input entered the model. We will refer to that behavior in claim 3.

One might argue that we could have additionally used a Bayesian model selection (BMS) procedure. If the BMS favored one model (i.e., the winning model), only parameters from this model might have been interpreted. BMS, however, did not yield any evidence that the rather 'plausible' model 'A1' was superior to the others, rather the opposite (see claim 6).

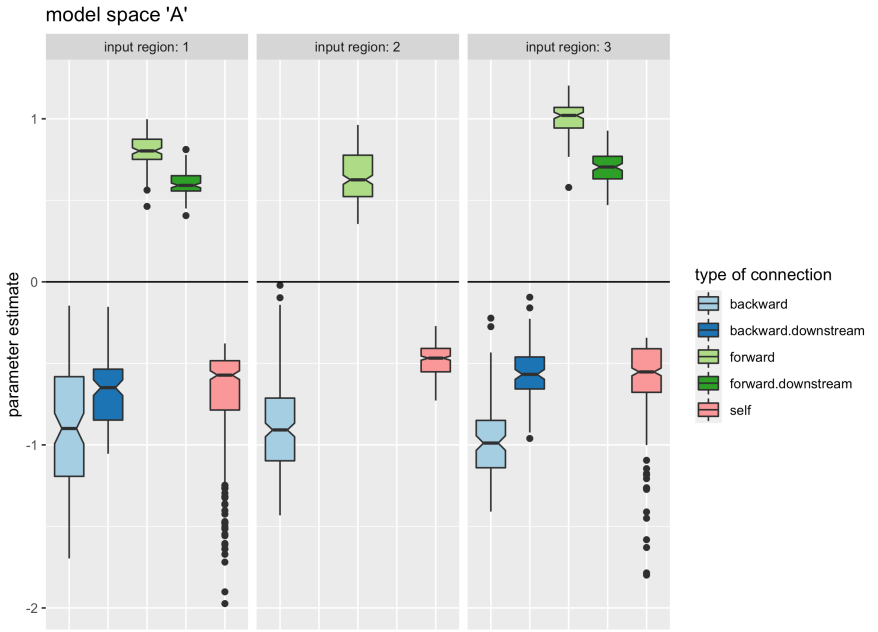


Figure 3. Magnitude of DCM parameters in the A matrix of model space 'A'. Illustrated are parameter estimates of three different models of model space 'A'. The models differed in their respective input region. Driving input (C matrix) either entered the model via BA17 (input region 1), BA18 (input region 2) or BA19 (input region 3). Across 105 sessions posterior parameter estimates were extracted from all connections of each model. The parameter estimates were classified as forward, backward, downstream forward, downstream backward, or self connection. See Figure S2 for a more detailed explanation of this classification. Across all different models, the following pattern was present: Forward connections and downstream forward connections became positive, and backward and downstream backward connections became negative. Self connections became negative by definition.

2 Positivity shifts to the B matrix, if enabled

2.1 Hypothesis

The previous models were rather naïve with respect to the matrix specifications, as we totally neglected a modulatory influence of the experimental condition on the connections (i.e., B matrix), but rather assumed a context-independent transfer between regions (i.e., A matrix). However, one might argue that the interregional connections – examplary in the early visual cortex – behaved differently during the experimental condition than during the control condition. More specifically, one would expect stronger coupling in the task condition than in the control condition, that is, when a visual stimulus is processed.

Consequently, the same effect visible in claim 1 should also be present, when a B matrix is specified (i.e., enabled) in addition to an A matrix. The B matrix is multiplied with the onset/offset vector $u^{(j)}$ of the experimental condition (see eq. 1). In theory, it is exactly where we would expect the positive forward connections to manifest itself, as visual processing is unsurprisingly dependent on a visual input.

In addition, parameters of the A and B matrices do have different prior variances (Fig. 4, top row) [28, 29]. Due to the higher variances in the priors of the B matrix, the B matrix parameters can be adjusted more easily during the estimation procedure. This in turn makes it simpler to capture the variance in the data by adjusting a particular connection in the B matrix rather than the corresponding connection in the A matrix. Furthermore, by the higher prior variances within the B matrix, larger maximum a posteriori parameter estimates become more likely.

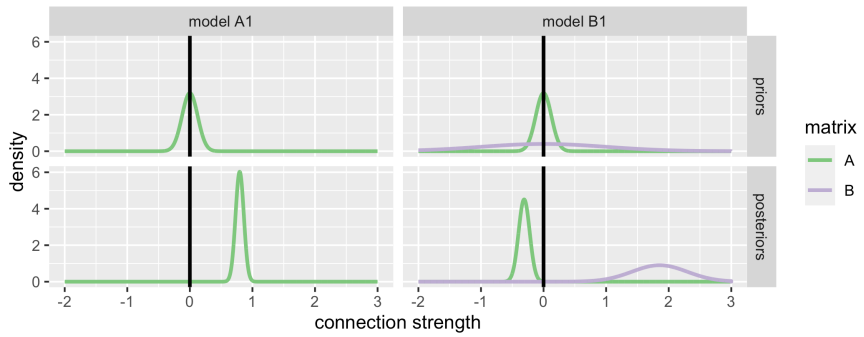


Figure 4. Exemplary prior and posterior distributions of model parameters. A interregional connection of the A matrix gets assigned a prior mean of $\mu = 0$ and prior variance of $\sigma^2 = \frac{1}{64}$. A interregional connection of the B matrix gets assigned a prior mean of $\mu = 0$ and prior variance of $\sigma^2 = 1$. Displayed are priors and posteriors of a particular connection from BA17 to BA18 in both A and B matrix ($a_{2,1}$ and $b_{2,1}$). The connections were extracted from the models 'A1' and 'B1', estimated by data of the first experimental session. Top row: prior distributions of the parameters are displayed. Bottom row: posterior distributions of the parameters are displayed. Left column: In model 'A1', in which only the A matrix is enabled for this particular connection, the posterior becomes positive. Right column: When a B matrix parameter is enabled for this connection (as in model 'B1'), the posterior becomes more positive, than the A matrix parameter of the corresponding model 'A1'. Conversely, the respective A matrix parameter of model 'B1' becomes even negative.

We therefore hypothesized that the pattern of the A matrix seen in claim 1 now

rather manifests in the connections of the B matrix. Namely, we expected positive forward connections and downstream forward connections in the B matrix.

2.2 Methods

First, we adapted both models (real and synthetic data) of the previous claim and added a B matrix to these models (Fig. 2, model space 'B'). Only interregional connections were enabled in the B matrix, corresponding the the off-diagonal elements of the matrix. As before, we varied the C matrix (i.e., input region), and analyzed the magnitude of all resulting model parameter estimates. We tested significant (FDR corrected) differences from zero by one sample t tests. We further tested for differences in the magnitude of parameter estimates of the B matrix parameters of the models of model space 'B', compared to those of the A matrix parameters of model space 'A' by applying FDR corrected paired samples t tests.

In addition, we simulated time series from an artificial ground truth model 'B1', and, like before, estimated all models of model space 'B' (Fig. 2).

2.3 Results

DCM estimation revealed positive forward connections, and positive downstream forward connections in the B matrix (Fig. 5, $p << 0.001$). This pattern was independent of the input region. Therefore, if the input was set to a different region, the parameter estimates changed according to fit this very pattern. Furthermore, backward connections as well as downstream backward connections were for the most part negative (all $p << 0.001$), comparable to the corresponding A matrix parameters in the claim 1. These patterns were present in the vast majority of models, and independent of where the experimental input entered the model.

In addition, the pattern was significantly stronger in the B matrix (in 8 out of 10 tested types of connections, $p < 0.05$), than it has been in the A matrix in claim 1, easily recognizable on the different scaling of the y axes between Figure 3 and Figure 5 (bottom row). With a B matrix enabled, the parameters of the A matrix tended to be closer to zero compared to the models with only A matrix enabled (all $p < 0.05$). The pattern of positive forward and negative backward connections has can not be seen in the A matrix (Fig. 5).

Moreover, with simulated data, one can clearly identify the same pattern for the B matrix as with real data (Fig. S5). In turn, the pattern was not longer manifested in the A matrix (Fig. S6).

We illustrated the parameter estimates (prior and posterior means and variances) of the forward connection from BA17 to BA18 of experimental session 1 (out of 105) as an example in Figure 4. The posterior A matrix parameter of the forward connection was positive, when the B matrix was disabled (i.e., claim 1, Fig. 3, bottom left). When the a B matrix was enabled, this very connection from BA17 to BA18, now in the B matrix became even stronger positive, then the A matrix forward connection of the corresponding model with only A matrix enabled (Fig. 4, bottom right). Conversely, the corresponding A matrix parameter in the model with B matrix enabled however was slightly negative. Whereas this is just an exemplary connection of session 1, Figure 5 in comparison with Figure 3 indicates that this pattern is present across the vast majority of sessions.

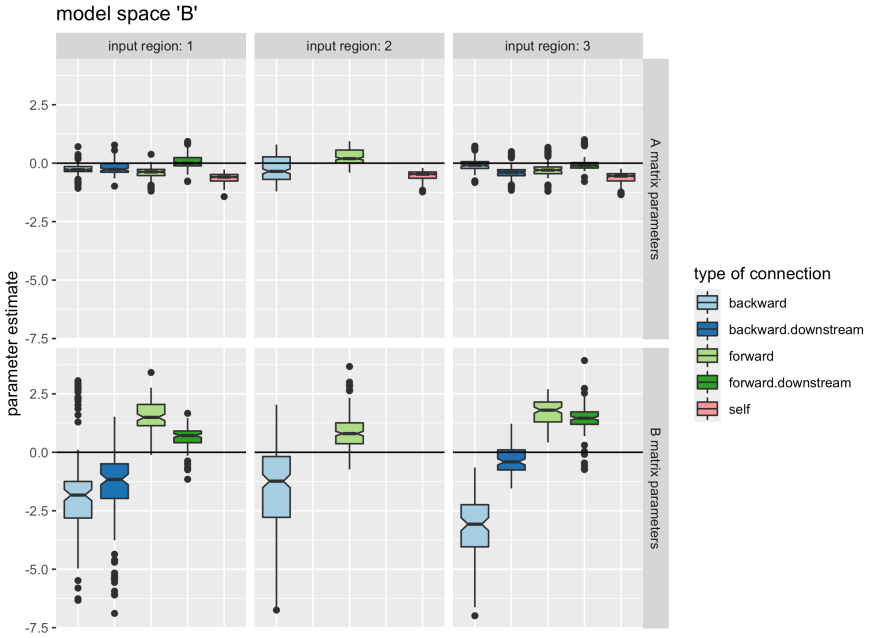


Figure 5. Magnitude of DCM parameters, when a B matrix was enabled. Illustrated are parameter estimates of three different models of model space 'B'. See Figure 3 for details. Across all different models, the following pattern was now rather present in the B matrix than in the A matrix: Forward connections and downstream forward connections were estimated positive, and backward and downstream backward connections were estimated negative. Self connections (within the A matrix) were estimated negative by definition.

3 Negative backward connections support self inhibition

3.1 Hypothesis

Experiments of claim 1 and claim 2 have confirmed a positive forward coupling, and positive downstream forward coupling. This pattern has either be seen in the A matrix, if a B matrix was *disabled*, or in the B matrix, if the B matrix was *enabled*. However, a negative backward coupling and negative downstream backward coupling has been demonstrated to accompany the positive forward and downstream forward couplings. As the negativity in (downstream) backward connection seems similarly predetermined as the positivity in (downstream) forward connections, the question regarding the origin or purpose of this negativity arises.

One possible role of the negative backward connection would be to support the inhibitory self connections in down-regulating the activity in the regions, and therefore to subserve as antagonist to the positive (forward & downstream forward) connections in maintaining a stable system. It might therefore be computationally more cost effective to move the priors of both the negative self connection, and the backward connection to just a certain degree, rather than moving just the prior of the self connection alone. By shifting just one of both priors, that very shift would need to be stronger to exceed the same effect (of downregulating the regions) than if both priors had to be shifted to a less severe degree.

We hypothesized, that in models with backward connections disabled, the self connections get larger negative than in models with enabled backward connections. This effect should rather be seen in self connections of those regions, which are targeted by a backward connection (e.g., region BA17 and BA18 in model 'A1', Fig. 2), rather than regions that are not targeted by a backward connection (e.g., region BA19 in model 'A1', Fig. 2).

3.2 Methods

We replicated the models of model spaces 'A' and 'B' to form model spaces 'Anb' and 'Bnb', respectively (Fig. 2, bottom row). In these model spaces, we disabled all backward and downstream backward connections. So for instance, from model 'A1', we disabled the backward connection from BA18 to BA17 in the A matrix, and the downstream backward connection from BA19 to BA18 in the A matrix, and named it model 'Anb1'. From model 'A2', we disabled the backward connection from BA17 to BA18, and the downstream backward connection from BA19 to BA18 in the A matrix, and named it model 'Anb2'. We proceeded with model 'A3' in an analogue fashion to create model 'Anb3'. In model space 'B' we disabled the very same connections to create models of the model space 'Bnb', but both in the respective A matrix, and the respective B matrix.

We already illustrated in claims 1 and 2, that backward connections render negative in most of the cases in the A or B matrix, respectively. Here, we were interested in how the omission of the (negative) backward connections changes the self-inhibition in each region. We hypothesized, that in all regions, which were initially targeted by a (negative) backward connection, the self connection will become stronger negative after disabling the backward connection to compensate for the 'missing' backward connection.

3.3 Results

In Figure 6, we illustrate the parameter estimates of the self connections, pair-wise for models with backward connections enabled, and models with backward connection disabled. One can clearly discern, that if a region was targeted by a backward connection, the self connection became less negative than if the region was not targeted by a backward connection. For instance, in models which got the input in region 1 (Fig. 6, top row), both region 1 and region 2 were targets of a backward connection. These both regions showed a clear difference in negativity of the self connection, when the backward connections were disabled. On the opposite, in regions 3, which was not targeted by a backward connection, this effect is not seen as strongly or not seen at all. Similarly, when the input enters region 2 (Fig. 6, middle row), only region 2 was targeted by backward connections. Therefore, only region 2 showed a strong difference in self connections after disabling the backward connections, whereas regions 1 and 3 remain less unaffected. The same logic applies for models with region 3 as input region (bottom row). Strong differences in the distribution of the magnitudes of self connections are seen in regions, which were targeted by a backward connection in model spaces 'A' and 'B', but not in model spaces 'Anb' and 'Bnb'.

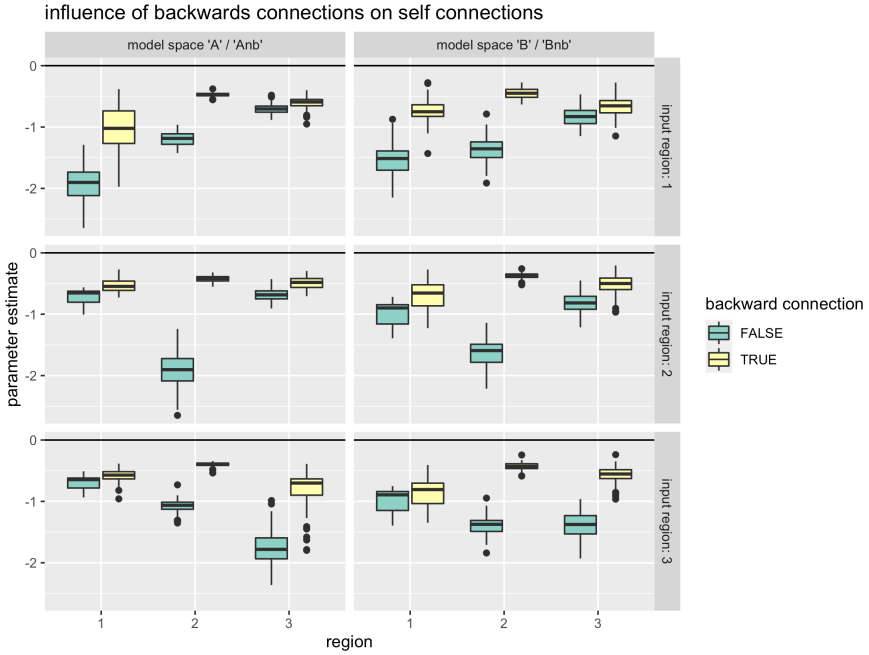


Figure 6. Magnitude of self connections when backward connections are disabled vs. enabled. Displayed are self connections (unit Hertz) of regions 1, 2, and 3 (x axis). The input region varies across rows. The left column displays self connections of the three regions within models which are part of model spaces 'A' and 'Anb', whereas the right column displays self connections of regions of models which are part of model spaces 'B' and 'Bnb'. Yellow boxes represent parameter estimates of models with backward connections enabled (model spaces 'A' and 'B'). Green boxes represent parameter estimates of models with backward connections disabled (model spaces 'Anb' and 'Bnb').

4 The pattern manifests independently of the underlying true model

4.1 Hypothesis

All of the previous analyses were limited to only include models belonging to one particular model space. Still open is the behavior of parameters, if the true (data generating) model is outside the respective model space, and if the very same pattern would still be persistent. To answer this question we conducted several *cross simulations*, by generating data by one model and estimating models of a disjunct model space.

In the previous claims we have demonstrated, that the positive and negative connections manifest themselves in the B matrix, if enabled, or in the A matrix, if B matrix is disabled. However, the ground truth model which generated the data remained unknown. In this claim we argue, that even if the time series are generated by a model with positive and negative connections in the A matrix (with B matrix disabled), the pattern manifests itself in a B matrix, if enabled during model estimation. Likewise, if the time series is generated by a model with positive and negative connections in the B matrix, the pattern will manifest itself in the A matrix, if the B matrix is disabled during model estimation.

4.2 Methods

For this line of simulations, we created models similar to the procedure shown in Figure S3. First, we created an average model 'A1' from empirical data. Then we simulated time series with different SNRs by this model, and in turn used these time series to estimate models of model space 'B', i.e., 'B1', 'B2', and 'B3' (Fig. 2). Likewise, we created an average model 'B1' from empirical data. Then we simulated time series with different SNRs by this model, and in turn used these time series to estimate models of model space 'A', i.e., 'A1', 'A2', and 'A3' (Fig. 2). We extracted the parameter estimates of the different estimated models for inference.

4.3 Results

The posterior parameter estimates distributed according to the very same pattern as before. If the time series was generated by model 'A1', and this time series was used to estimate models of model space 'B', the pattern now shifted into the B matrix (Fig. 7). More specifically, the forward and downstream forward connections in the B matrix became positive, and the backward and downstream backward connections became negative (Fig. 7). This was seen in a vast majority of models where the experimental input entered region 1 or 3. In models in which the experimental input entered region 2, the negativity of the backward connection only manifested in about 50% of cases. On the contrary, in the A matrix, where this pattern was initially manifested during data generation, it largely vanished (Fig. S7), closely corresponding to the insights revealed by real data in Figure 2.

On the other hand, we simulated the opposite direction. If the time series was generated by model 'B1', and this time series was used to estimate models of model space 'A', the pattern now shifted from the B matrix towards the A matrix (Fig. 8). More specifically, the forward and downstream forward connections in the A matrix became positive, and the backward and downstream backward connections became negative (Fig. 8).

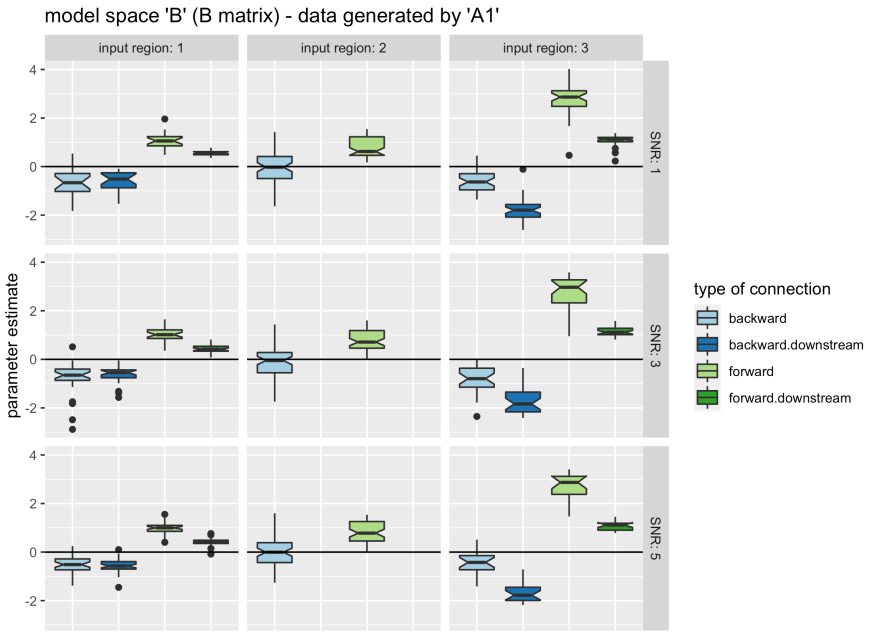


Figure 7. Magnitude of DCM parameters in the B matrix of models with B matrix enabled, estimated using time series generated by model 'A1'. Illustrated are parameter estimates of the B matrix of three different models (columns) of model space 'B' for three different SNRs (rows). The times series used to estimate the models however were generated by model 'A1'. The pattern of positive forward and negative backward connections manifested in the B matrix, although the data generating model comprised no B matrix. See Figure S4 as reference for the simulations. See Figure S7 for the corresponding A matrix parameters of the models.

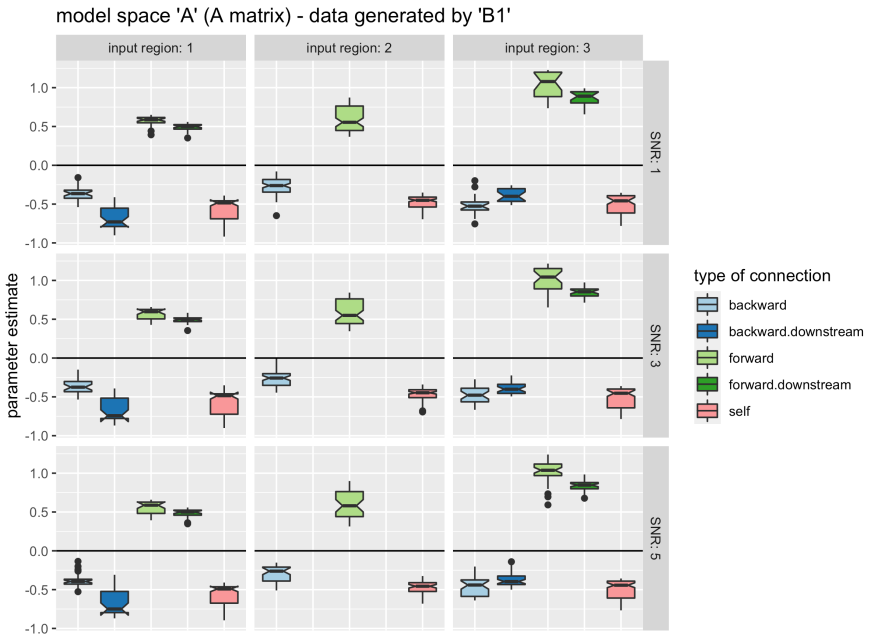


Figure 8. Magnitude of DCM parameters in the A matrix of models with B matrix disabled, estimated using time series generated by model 'B1'. Illustrated are parameter estimates of the A matrix of three different models (columns) of model space 'A' for three different SNRs (rows). The times series used to estimate the models however were generated by model 'B1'. The pattern of positive forward and negative backward connections manifested in the A matrix, whereas the data generating model showed this pattern in the B matrix. See Figure S4 as reference for the simulations.

5 The pattern vanishes when enabling lateral connections

5.1 Hypothesis

The models used to evaluate claims 1-4 were of hierarchical structure. In particular, a signal needed to propagate all three regions in a continuous fashion. For instance, in model 'A1', it first entered region 1 (BA17), then wanders to the second (BA18), and finally to the third region (BA19) (Fig. 2). Likewise, in the case the input entered the second region (BA18), the signal propagated to BA17 and BA19 in a parallel fashion (Fig. 2).

However, it is not yet clear, if the pattern (forward positive, backward negative) still persists, if *all* regions of the model were fully interconnected. Hypothetically, the pattern may vanish, because of additional degrees of freedom by the additional connections. The signal might travel along the regions in several possible ways, even differently per experimental session, leading to the pattern getting untraceable across sessions. In other words, with the higher amount of possible connections, the model becomes simply more flexible to fit the data.

5.2 Methods

To test this, we enabled reciprocal connections between the two regions, which were not connected in the models of claim 1 and claim 2. Therefore, we compared the parameter estimates of models from before (Fig. 2, 'A' and 'B') to the new models with the additional connections (Fig. 2, 'AF' and 'BF'), which we will term *fully connected*.

5.3 Results

The distribution of model parameters is displayed in Figure 9. The pattern was still present, i.e., positive forward connections, and negative backward connections. New were however the lateral connections. Those rendered either positive or negative. In model space 'AF' with disabled B matrix, the lateral connections rendered either positive or negative, with a slight tendency of getting negative (Fig. 9). The same accounts for model space 'BF' with enabled B matrix. In those models, the lateral connections rendered either negative or positive, with a slight tendency to getting negative.

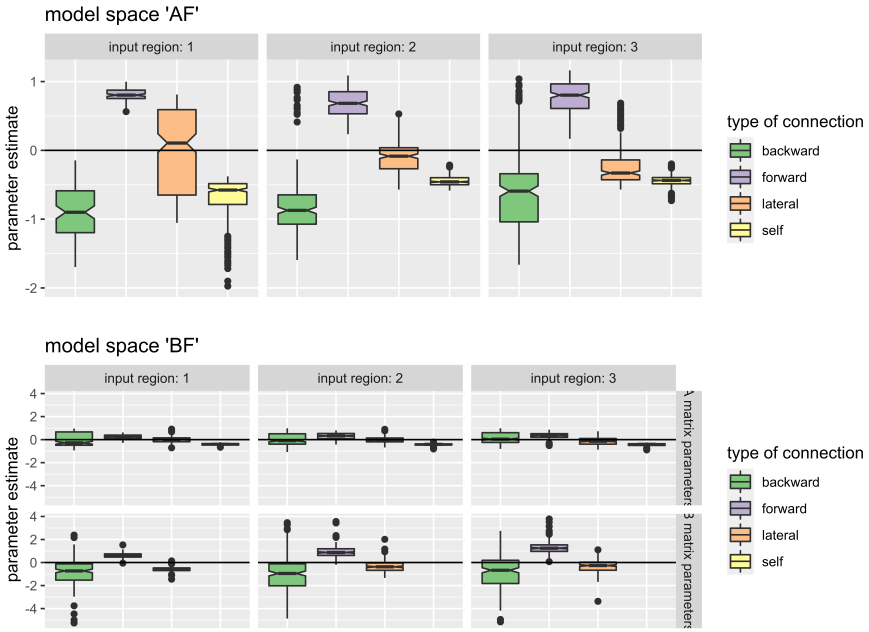


Figure 9. Magnitude of DCM parameters, when the regions are fully connected. Illustrated are parameter estimates of models of model space 'AF' (top) and 'BF' (bottom). The models differed in their respective input region. Driving input (C matrix) either entered the model via BA17 (input region 1), BA18 (input region 2) or BA19 (input region 3). All regions were fully interconnected. Across 105 sessions posterior parameter estimates were extracted from all connections of each model. The parameter estimates were classified as forward, backward, lateral, or self connection. See Figure S2 for a more detailed explanation of this classification. For model space 'BF', both A and B matrices are illustrated separately. For model space 'AF', only A matrix is illustrated. Across all different models, the following pattern was still present: Forward connections were estimated positive, and backward connections were estimated negative. Self connections were estimated negative by definition. Lateral connections were either estimated positive or negative.

6 The input region is not identifiable with uninformative priors

6.1 Hypothesis

The human early visual system is likely one of the best-studied cortical areas. Therefore, it is common sense, that the visual input enters the human cortex via the Lateral Geniculate Nucleus entering the V1 region (i.e., BA17), and hierarchically and reciprocally processes to higher regions, such as V2 (BA18), and other visual regions such as V3, V4, or V5/MT (BA19) [13]. However, the model structure is not always as easy to derive from the literature as in the present example. For example, it is often unclear between which regions connections should be enabled (A and B matrices) and into which region the experimental input should enter directly (C matrix).

The DCM framework encourages to set up multiple models and to use data to evaluate their likelihood via model comparison, i.e., BMS [20]. However, in the distributed DCM software package, a uniform distribution of prior model probabilities is set upon a model space, so that $P(M_1) = P(M_2)$, and so forth. However, given differently strong (prior) evidence from the literature for different model structures, this assumption is oversimplified for nearly every application. Therefore, model selection results must be distorted, as equal prior probabilities are implausible. In other words, this restriction somewhat contradicts the Bayesian nature of the analysis and hence further limits its possibilities. Most researchers however do not take into account such different prior probabilities on the models, but assume a uniform distribution of prior probabilities. As a consequence, the posterior odds of a model, compared to another model, completely depends on the Bayes Factor (BF) and therefore the data, and not the prior probabilities of each model (see eq. 8- 10).

$$P(M_i|y) = \frac{P(y|M_i) \cdot P(M_i)}{P(y)} \tag{8}$$

$$\frac{P(M_1|y)}{P(M_2|y)} = \frac{P(y|M_1)}{P(y|M_2)} \cdot \frac{P(M_1)}{P(M_2)} \tag{9}$$

$$\text{posterior odds} = \text{Bayes Factor} \cdot \text{prior odds} \tag{10}$$

Equation 8 depicts the Bayes theorem [5]. The conditional probability of model i, given the data y is the posterior probability of a model $P(M_i|y)$. It is the quotient of the conditional probability of the data, given that model i is true $P(y|M_i)$ – often termed likelihood – times the prior probability of model i $P(M_i)$, and the marginal probability $P(y)$ for normalization. When comparing two models, simply the fraction between two posterior probabilities can be calculated, as for example with equation 9, which is named the posterior odds (eq. 10). The posterior odds is nothing than the BF times the prior odds. With uniform prior probabilities, and therefore prior odds of $\frac{1}{1}$, the posterior probabilities simply correspond to BFs.

6.2 Methods

We implemented a straightforward BMS based on BFs. We then compared three models within a model space with each other in a pairwise fashion. The respective three models varied in their input regions (input into BA17, BA18, or BA19). Therefore, we only took the models of model spaces 'A' and 'B' separately (Fig. 2). We first took each of the 105 experimental sessions, and calculated a BF between each pair of models i and j, as derived by the model evidences $p(y|M_i)$, which were approximated by the negative free energy F_i [27] of a model:

$$BF_{i,j} = \frac{p(y|M_i)}{p(y|M_j)} \approx exp(F_i - F_j) \tag{11}$$

We further calculated a Group Bayes Factor $GBF_{i,j}$ [27] between each pair of model across functional runs:

$$GBF_{i,j} = exp(\sum_{s=1}^{S} F_i - F_j) = \prod_{s=1}^{S} BF_{i,j}^{(s)} \tag{12}$$

with s the sessions and S the total number of sessions. Because the proposed GBF depends on the number of sessions included, and therefore easily converges to 0 or ∞ for high number of sessions, we further introduce the root-n-Group Bayes Factor:

$$rnGBF = \sqrt[S]{GBF_{i,j}} \tag{13}$$

Further to calculating Bayes Factors, we converted the Bayes Factors to posterior probabilities. Therefore, the posterior probability p_{M_1} of model 1 (within a comparison to model 2) is – given the uniformity assumption on priors, therefore prior odds of 1 – derived by [19]:

$$p_{M_1} = \frac{\text{posterior odds}}{1 + \text{posterior odds}} = \frac{\frac{p(y|M_i)}{p(y|M_j)}}{1 + \frac{p(y|M_i)}{p(y|M_j)}} = \frac{\frac{p(M_i|y)}{p(M_j|y)}}{1 + \frac{p(M_i|y)}{p(M_j|y)}} = \frac{\text{BF}}{1 + \text{BF}} \tag{14}$$

According to Raftery [21], a Bayes Factor of > 150 – corresponding to a posterior probability of > 0.99 – can be seen as *very strong evidence* in favor of a model. Similar accounts for a Bayes Factor of $\frac{1}{150}$ – corresponding to a posterior probability of < 0.01 – speaking strongly against a particular model.

6.3 Results

The pair-wise model comparisons between models with different input regions are displayed in Figure 10, both for model group 'A' (left column) and model group 'B' (right column). In both groups, the following pattern is visible:

- model 1 (i.e., input into BA17) is outperformed by model 2 (i.e., input into BA18), with $rnGBF_{1,2}^{(A)} = 2 \times 10^{-40}$, and $rnGBF_{1,2}^{(B)} = 4 \times 10^{-50}$
- model 1 outperforms model 3 (i.e., input into BA19), with $rnGBF_{1,3}^{(A)} = 4 \times 10^{34}$ and $rnGBF_{1,3}^{(B)} = 2 \times 10^{12}$
- model 2 outperforms model 3, with $rnGBF_{2,3}^{(A)} = 2 \times 10^{74}$ and $rnGBF_{2,3}^{(B)} = 5 \times 10^{61}$

Therefore, according to Bayes Factors and $nrGBF$ – i.e., under uniformity assumptions on priors – model 2 has highest posterior probability, with $p_{M_2} >> 0.99$ when comparing both to models 1 and 3.

The model selection presented in this chapter was performed to emphasize the relevance of the patterns elaborated in the previous claims. For example, in claim 1, one could argue that it is nothing one should be concerned about when model parameters change after the researcher changing the input region, because a biophysically illogical model, such as model 'A2', would be inferior in a model comparison compared to a biophysically more plausible model such as model 'A1'. Inferior means, it should be inferior both by smaller prior probability (which is

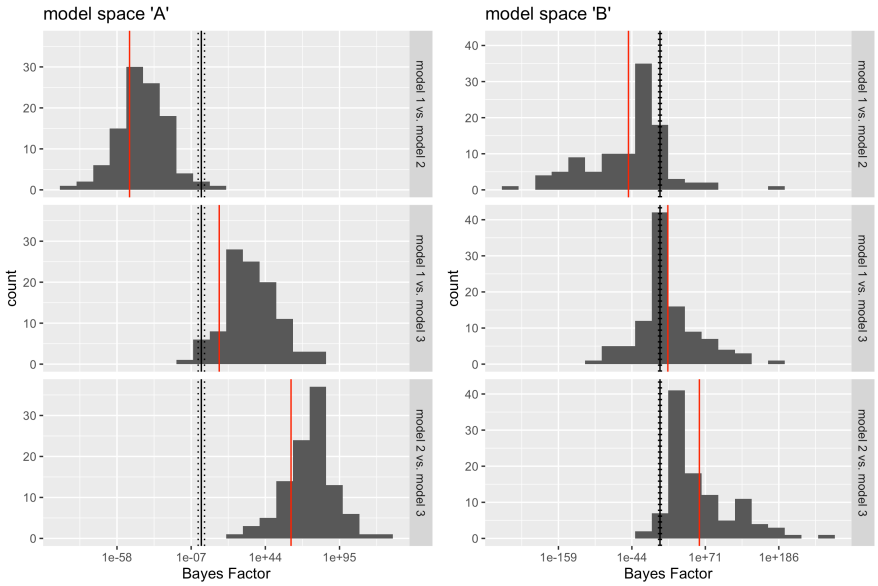


Figure 10. Distributions of Bayes Factors for pairwise model comparisons. Left column: models of type 'A' (see Fig. 2). Right column: models of type 'B' (see Fig. 2). The Bayes Factors (in log10 scale) are visualized on the x axes, and the corresponding counts (out of 105 total comparisons each) are displayed on the y axes. The solid vertical black lines marks a Bayes Factor of 1, i.e. the situation where both models of a comparison are equally likely. The dotted vertical black lines mark Bayes Factors of 150 and $\frac{1}{150}$, the defined thresholds for *very strong evidence* in favor or against a particular model, respectively [21]. In the figures of the right column, those thresholds are very close to the solid vertical black line. The red vertical line indicates the respective $rnGBF$ (see eq. 13).

typically not adjusted in DCM), and smaller likelihoods and therefore BFs (which was not the result of the presented model comparison). Thus, if one were to include a model selection in one's analysis pipeline, and then interpret only the parameters of the winning model, then it need not matter that the inferior models have unreasonable parameter expressions. With the above results we were able to show, using two different model spaces as examples, that the more plausible model (from a biophysical point of view, e.g., 'A1' or 'B1') would have been inferior in the model comparison, and a rather implausible model such as 'A2' or 'B2' would have won. If – in a study with a substantive question about connectivity in early visual areas – a researcher had subsequently interpreted the parameters of these winning models, they would have interpreted parameters which were merely determined by the structure of the model (which is also true for the winning model), but are not very plausible from the outset.

Discussion

Complementing the motivating example

Throughout the different claims we have evaluated possible explanations for the pattern we have seen in all models, namely, positive forward and negative backward connections. We were able to demonstrate, that if we changed the input region and therefore if for instance a forward connection turned into a backward connection, its parameter estimate switched accordingly. In our introductory example (page 3) we introduced a hypothesis about the nature of the positivity of the forward couplings, namely to spread the neural activation across all regions. We have then carved out a possible explanation for the emergence of negative backward couplings, namely that those might play a supportive role in the down-regulation of each region, a task which is usually assigned to the self connections.

With the findings we are now in a position to close the last gap from our motivating example on page 3. In that example we argued, that following plausibility considerations on the level of the neural state equation, one is able to predict the quality (i.e., being positive or negative) of most connections without even estimating the model. Now having gained some insights about the nature of the backward connection, which seems to turn negative in most cases, we were able to close the remaining gap, rendering the model fully predictable.

Putting it more precisely, from the elements of the A matrix in our motivating example (eq. 2), one is now able to not only predict the on-diagonal elements, which are by definition negative, but also both of the off-diagonal elements $a_{1,2}$ and $a_{2,1}$. Whereas we already proposed the forward connection $a_{2,1}$ to become positive from plausibility considerations, we have then shown with empirical data and simulations of claims 1, 2, 4, and 5 that this holds in the vast majority of cases. Furthermore, we were able to demonstrate that the backward connection $a_{1,2}$ should render negative, by demonstrating the negativity of backward connections with claims 1, 2, 4, and 5, and identifying a possible concept for the negativity in claim 3.

In addition, we were able to show that the data are generously explained by model structures that actually had nothing to do with the generation of the data. For example, the B matrix had caught most of the effects when modeled, even though these effects were actually attributable to the A matrix. On the other hand, the A matrix explained effects when no B matrix was available for modeling (claim 4).

Now, one could argue that all this is not so dramatic if one just follows the guidelines of the developers and performs model selections first. We have shown the opposite here with a very comprehensible model selection. Even with this intuitive fMRI dataset, model selection did not spawn the model that would have been easily accepted by a wide audience of experts.

Limitations of the present claims

Despite using a simple paradigm, we have also used a simple model for all plausibility considerations, empirical calculations and simulations. For instance, one could argue that the pattern vanished if the number of regions becomes sufficiently high, and if the regions got more interconnected than in the presently used models. In all claims but claim 5, the signal had only one possible route to pass through the system. Enabling many connections between several regions might open more degrees of freedom for the signal to be propagated, and the pattern may vanish across sessions or subjects. However, in a three region model, fully interconnected, the pattern remains still consistent (claim 5). Furthermore, the pattern may become less traceable if the covariance of the regions' time series is not high enough, or if some regions activate in

an statistically orthogonal fashion. When more experimental conditions would be combined, perhaps in an fully orthogonal design, plausibility considerations leading to a fully predictable outcome might get more cumbersome.

Frontiers of model interpretability

Many models in the existing literature (e.g., [6, 11]), including our own (e.g., [9, 14]) can become victim of the logic introduced during this article. Whereas DCM is a highly elegant method to describe neural interactions, we want to emphasize, that the interpretation of many model parameters just do not add any value to some research, if the outcome is predictable anyway. Therefore, before applying Dynamic Causal Models in the first place, we highly recommend to rehearse possible outcomes, which is feasible because at that point preliminary inference has usually already been done on the level of neural activation.

References

1. A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014.

2. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), Aug. 2001.

3. R. B. Buxton, K. Uludağ, D. J. Dubowitz, and T. T. Liu. Modeling the hemodynamic response to brain activation. *NeuroImage*, 23:S220–S233, Jan. 2004.

4. D. F. Cechetto and J. C. Topolovec. Cerebral Cortex. In V. S. Ramachandran, editor, *Encyclopedia of the Human Brain*, pages 663–679. Academic Press, New York, Jan. 2002.

5. B. Efron. Bayes' theorem in the 21st century. *Science (New York, N.Y.)*, 340(6137):1177–1178, 2013. Publisher: American Association for the Advancement of Science.

6. S. L. Fairhall and A. Ishai. Effective Connectivity within the Distributed Cortical Network for Face Perception. *Cerebral Cortex*, 17(10):2400–2406, Oct. 2007.

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

8. K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, Jan. 2007.

9. S. Frässle, F. M. Paulus, S. Krach, S. R. Schweinberger, K. E. Stephan, and A. Jansen. Mechanisms of hemispheric lateralization: Asymmetric interhemispheric recruitment in the face perception network. *NeuroImage*, 124:977–988, 2016. Publisher: Elsevier Inc.

10. J. Gonzalez-Castillo, Z. S. Saad, D. A. Handwerker, S. J. Inati, N. Brenowitz, and P. A. Bandettini. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proceedings of the National Academy of Sciences*, 109(14):5487–5492, Apr. 2012.

11. J. D. Herrington, J. M. Taylor, D. W. Grupe, K. M. Curby, and R. T. Schultz. Bidirectional communication between amygdala and fusiform gyrus during facial recognition. *NeuroImage*, 56(4):2348–2355, June 2011.

12. B. Horwitz, B. Warner, J. Fitzer, M.-A. Tagamets, F. T. Husain, and T. W. Long. Investigating the neural basis for functional and effective connectivity. Application to fMRI. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1093–1108, 2005. Publisher: The Royal Society London.

13. E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, S. Mack, and others. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

14. R. Kessler, K. M. Rusch, K. C. Wende, V. Schuster, and A. Jansen. Revisiting the effective connectivity within the distributed cortical system for face perception. *NeuroImage: Reports*, 2021.

15. R. Kessler, S. Schmitt, T. Sauder, F. Stein, D. Yüksel, D. Grotegerd, U. Dannlowski, T. Hahn, A. Dempfle, J. Sommer, O. Steinsträter, I. Nenadic, T. Kircher, and A. Jansen. Long-Term Neuroanatomical Consequences of Childhood Maltreatment: Reduced Amygdala Inhibition by Medial Prefrontal Cortex. *Frontiers in Systems Neuroscience*, 14:28, June 2020.

16. J. L. Lancaster, L. H. Rainey, J. L. Summerlin, C. S. Freitas, P. T. Fox, A. C. Evans, A. W. Toga, and J. C. Mazziotta. Automated labeling of the human brain: A preliminary report on the development and evaluation of a forward-transform method. page 5.

17. J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox. Automated Talairach Atlas labels for functional brain mapping. page 12.

18. K. Nagy, M. W. Greenlee, and G. Kovács. The Lateral Occipital Cortex in the Face Perception Network: An Effective Connectivity Study. *Frontiers in Psychology*, 3, 2012.

19. O'Hagan, Tony. Bayes Factors. *Significance*, 3(4):184–186, 2006.

20. W. Penny. Comparing Dynamic Causal Models using AIC, BIC and Free Energy. *NeuroImage*, 59(1):319–330, Jan. 2012.

21. A. E. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163, 1995.

22. R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, Jan. 1999. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group.

23. A. Razi, M. L. Seghier, Y. Zhou, P. McColgan, P. Zeidman, H.-J. Park, O. Sporns, G. Rees, and K. J. Friston. Large-scale DCMs for resting-state fMRI. *Network Neuroscience*, 1(3):222–241, Oct. 2017.

24. L. A. Remington. Chapter 13 - Visual Pathway. In L. A. Remington, editor, *Clinical Anatomy and Physiology of the Visual System (Third Edition)*, pages 233–252. Butterworth-Heinemann, Saint Louis, Jan. 2012.

25. B. P. Rogers, V. L. Morgan, A. T. Newton, and J. C. Gore. Assessing functional connectivity in the human brain by fMRI. *Magnetic resonance imaging*, 25(10):1347–1357, 2007. Publisher: Elsevier.

26. R. Sladky, M. Spies, A. Hoffmann, G. Kranz, A. Hummer, G. Gryglewski, R. Lanzenberger, C. Windischberger, and S. Kasper. (S)-citalopram influences amygdala modulation in healthy subjects: a randomized placebo-controlled double-blind fMRI study using dynamic causal modeling. *NeuroImage*, 108:243–250, Mar. 2015.

27. K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston. Comparing hemodynamic models with DCM. *NeuroImage*, 38(3):387–401, Nov. 2007.

28. 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*, 200:174–190, Oct. 2019.

29. 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*, 200:12–25, Oct. 2019.

Supporting Information

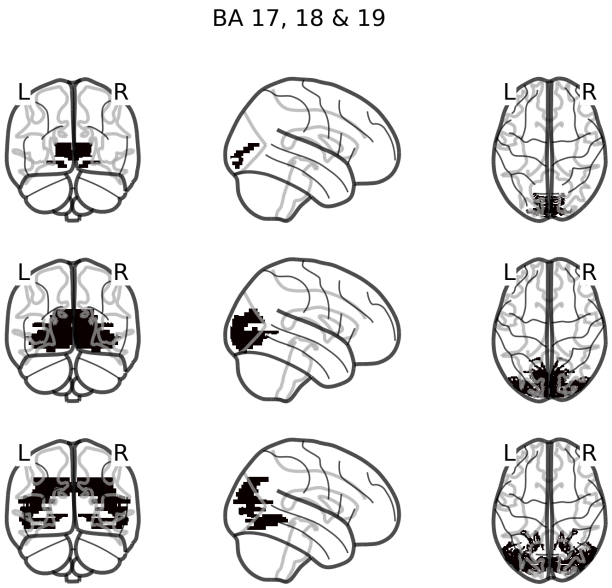


Figure S1. Masks to identify visual regions in occipital cortex. **Top row**: Brodmann area 17 (BA17) encompassing region V1. **Middle row**: Brodmann area 18 (BA18) encompassing region V2. **Bottom row**: Brodmann area 19 (BA19) encompassing region V3, V4, and V5.

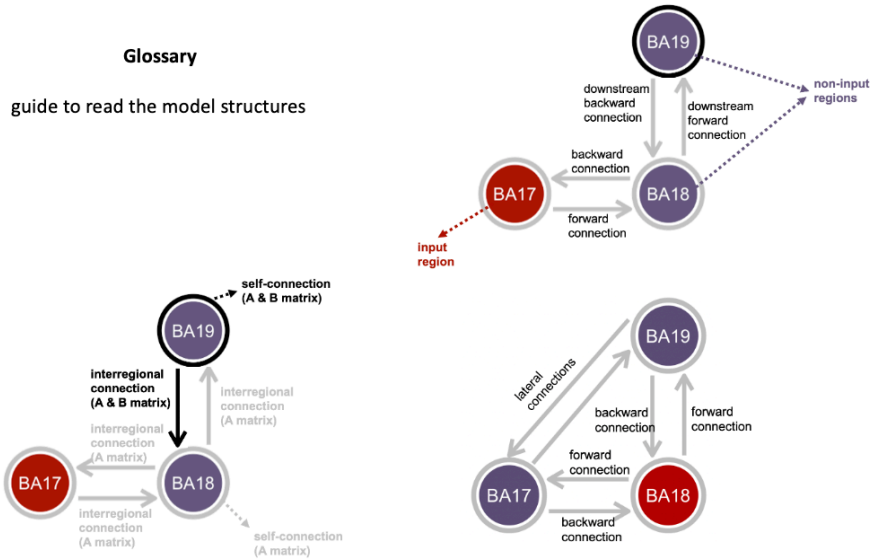


Figure S2. Glossary models. Different parts of these hypothetical models were labeled to better understand the differences in the models of Figure 2, and to get an intuition about the terms used when describing particular connections of a model.

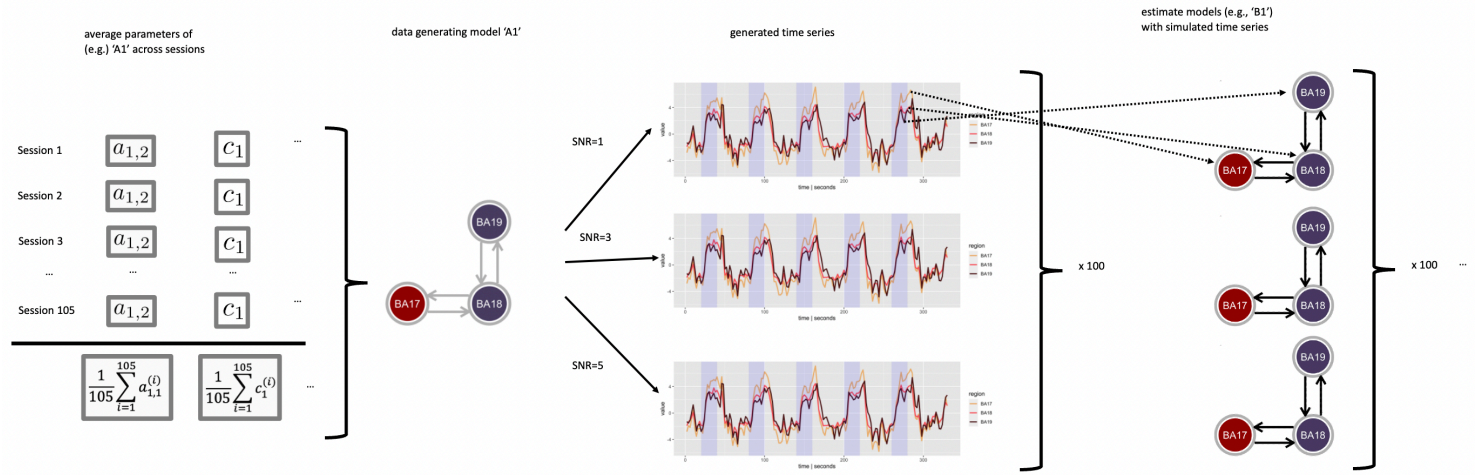


Figure S3. Exemplary creation of a data generating model and simulation of time series. First, estimated DCM parameters for each connection of a particular model (e.g., 'A1') are averaged across sessions. Thereafter, a data generating model is created by the average model parameters. Then, time series are generated by that model with different signal-to-noise ratios (SNR). For each SNR and region, 100 different time series were generated. These generated time series are in turn used to estimate particular models of interest, e.g., models of type 'B1'. Inference is done on the resulting model parameters.

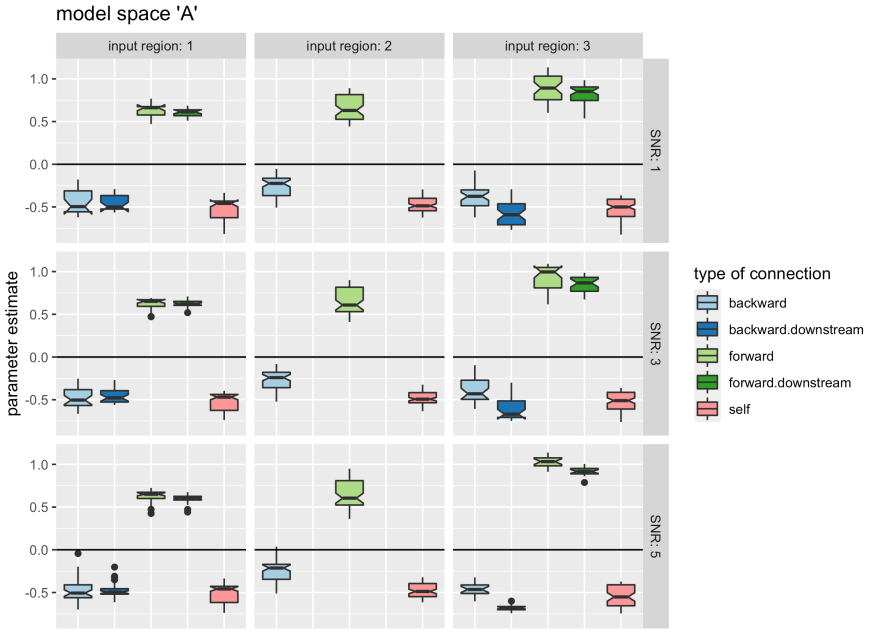


Figure S4. Magnitude of DCM parameters in the A matrix with simulated time series. Illustrated are parameter estimates of three different models of model space 'A'. The models differed in their respective input region. Driving input (C matrix) either entered the model via BA17 (input region 1), BA18 (input region 2) or BA19 (input region 3). Across 100 simulated sessions posterior parameter estimates were extracted from all connections of each model. The parameter estimates were classified as forward, backward, downstream forward, downstream backward, or self connection. See Figure S2 for a more detailed explanation of this classification. Throughout all different models, the following pattern was present: Forward connections and downstream forward connections were estimated positive, and backward and downstream backward connections were estimated negative. Self connections were estimated negative by definition. This pattern was present for different signal-to-noise ratios (SNRs).

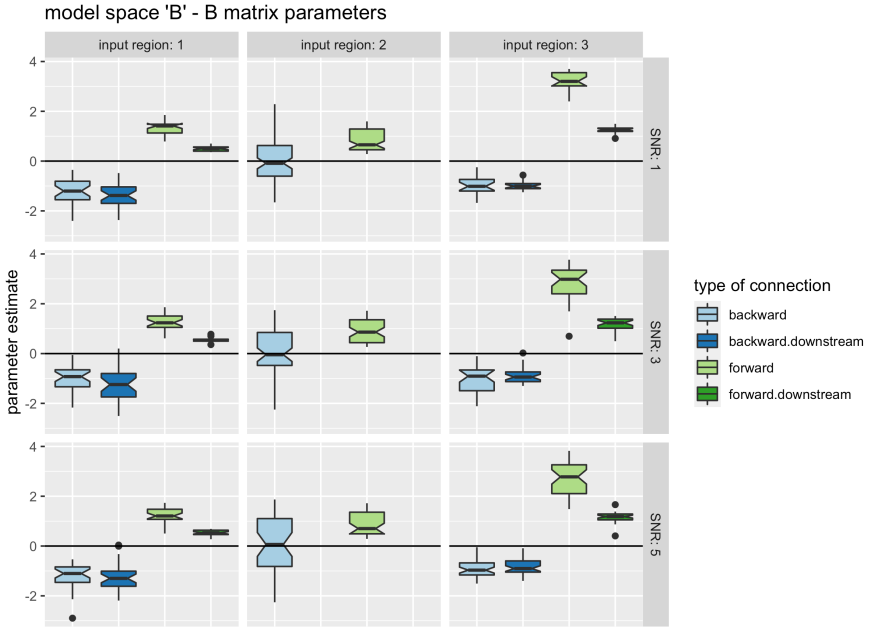


Figure S5. Magnitude of DCM parameters in the B matrix of models with B matrix enabled with simulated time series. Illustrated are parameter estimates of the B matrix of three different models of model space 'B'. See Figure S4 as reference. When a B matrix is enabled, the pattern (positive forward, negative backward) is now manifested in the B matrix.

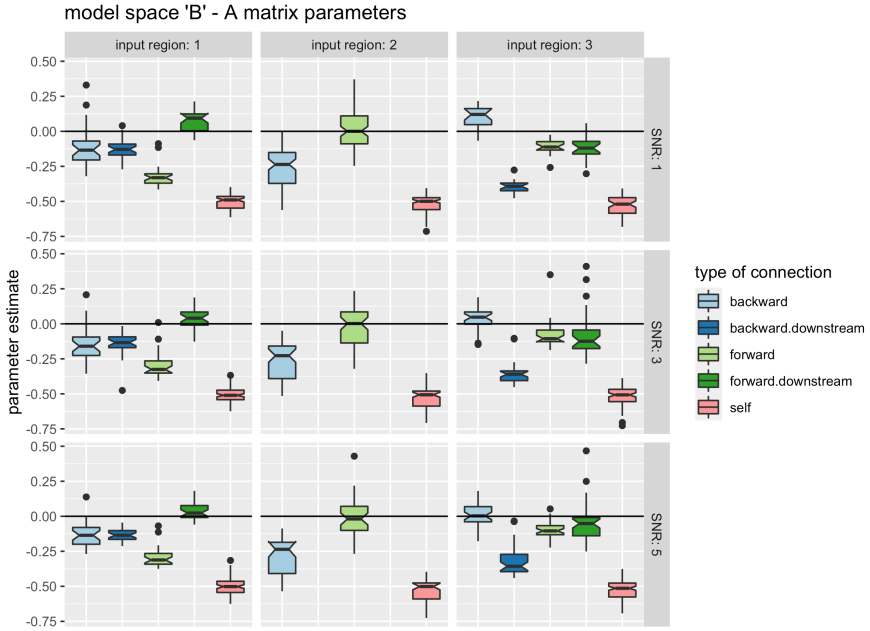


Figure S6. Magnitude of DCM parameters in the A matrix of models with B matrix enabled with simulated time series. Illustrated are parameter estimates of the A matrix of three different models of model space 'B'. See Figure S4 as reference. When a B matrix is enabled, the pattern (positive forward, negative backward) is not longer manifested in the A matrix.

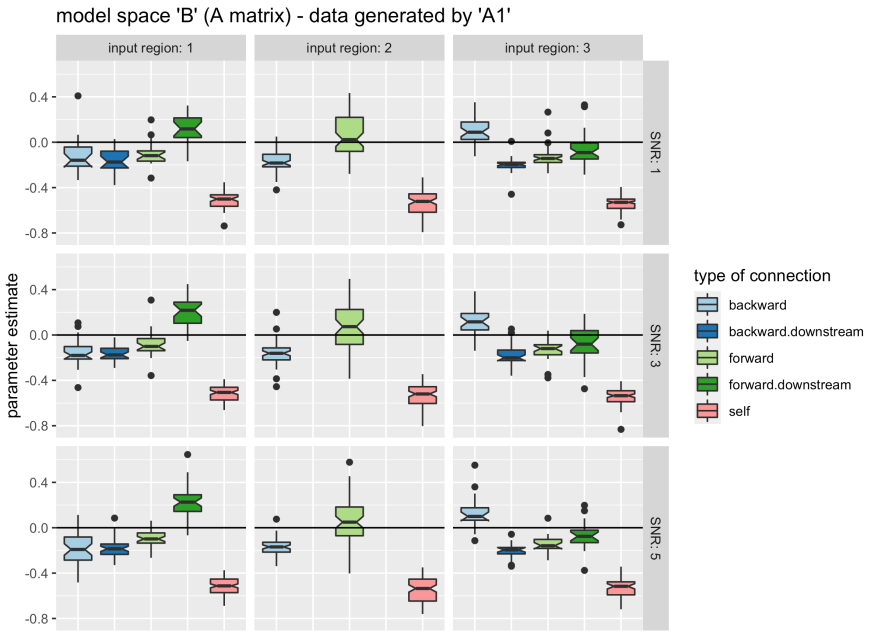


Figure S7. Magnitude of DCM parameters in the A matrix of models with B matrix enabled, fed by time series generated by model 'A1'. Illustrated are parameter estimates of the A matrix of three different models (columns) of model space 'B' for three different SNRs (rows). The times series used to estimate the models however were generated by model 'A1'. The pattern of positive forward and negative backward connections manifested in the B matrix (Fig. 7), although the data generating model comprised no B matrix. See Figure S4 as reference for the simulations. See Figure 7 for the corresponding B matrix parameters of the models.

Eigener Anteil der vorliegenden Arbeit

Laut §8, Absatz 3 der Promotionsordnung der Philipps-Universität Marburg (Fassung vom 15.07.2009) müssen bei den Teilen der Dissertation, die aus gemeinsamer Forschungsarbeit entstanden sind, „die individuellen Leistungen des Doktoranden deutlich abgrenzbar und bewertbar sein“. Der eigene Anteil wird im Folgenden detailliert erläutert.

1. Manuskript 1

   - Konzeption der Auswertung / Hypothese
   - Datenvorverarbeitungen (Programmierung & Durchführung)
   - Konnektivitätsanalysen (Programmierung & Durchführung)
   - Statistische Analysen (Programmierung & Durchführung)
   - Visualisierung der Ergebnisse
   - Interpretation der Ergebnisse (Zusammen mit Prof. Dr. Andreas Jansen, und z.T. anderen Koautoren)
   - Schreiben des Manuskriptes (Zusammen mit Prof. Dr. Andreas Jansen, Korrektur durch z.T. anderen Koautoren)

   Anteil gesamt: 70%

   Dieses Manuskript wurde in der vorliegenden Form im Journal *Frontiers in Systems Neurosciences* veröffentlicht [Kessler et al., 2020].

2. Manuskript 2

   - Konzeption des experimentellen Paradigmas (Paradigma 1)
   - Programmierung des Paradigmas (Paradigma 1)
   - Rekrutierung und Messung der Probanden (Paradigma 1)
   - Konzeption der Auswertung / Hypothese
   - Datenvorverarbeitungen (Programmierung & Durchführung)
   - Konnektivitätsanalysen (Programmierung & Durchführung)
   - Statistische Analysen (Programmierung & Durchführung)
   - Visualisierung der Ergebnisse
   - Interpretation der Ergebnisse (Zusammen mit Prof. Dr. Andreas Jansen)
   - Schreiben des Manuskriptes (Zusammen mit Prof. Dr. Andreas Jansen, Korrektur durch andere Koautoren)

   Anteil gesamt: 80%

   Dieses Manuskript wurde in der vorliegenden Form im Journal *Neuroimage: Reports* veröffentlicht [Kessler et al., 2021b].

3. Manuskript 3

   - Konzeption der Hypothese

- Programmierung & Durchführung Simulationen
- Konzeption der Auswertung / Hypothese
- Datenvorverarbeitungen (Programmierung & Durchführung)
- Konnektivitätsanalysen (Programmierung & Durchführung)
- Statistische Analysen (Programmierung & Durchführung)
- Visualisierung der Ergebnisse
- Interpretation der Ergebnisse (Zusammen mit Prof. Dr. Andreas Jansen)
- Schreiben des Manuskriptes (Korrektur durch Prof. Dr. Andreas Jansen)

Anteil gesamt: 90%

Dieses Manuskript ist noch nicht veröffentlicht [Kessler and Jansen, 2022].

Ort, Datum, Unterschrift Roman Keßler

Ort, Datum, Unterschrift Prof. Dr. Andreas Jansen

Norwegian University of Science and Technology

Institutt for informasjonssikkerhet og kommunikasjonsteknologi

Busch

Fakultet for informasjonsteknologi og elektroteknikk

Raja

Hochschule Darmstadt

Fachbereich Mathematik

Döhler, Gross, Jahn, Zisgen

Fachbereich Informatik

Andelfinger, Brucherseifer, Busch, Döhring, Grieser, Hergenröther, Roth, von Rüden, Tapia

Fachbereich Wirtschaftspsychologie

Hansen

Philipps Universität Marburg

Fachbereich Medizin

Jansen, Kircher, Nenadic, Straube

Fachbereich Biologie

Culmsee, Homberg, Oberwinkler

Fachbereich Psychologie

Endres, Schubö, Schwarting, Wöhr

Fachbereich Physik

Bremmer

Albert-Ludwigs Universität Freiburg

Fachbereich Biologie

Baumeister, Driever, Hess, Kirsch, Neubüser, Schulze, Speck, Weber

Fachbereich Medizin

Bach, Heinrich