NeuroImage 145 (2017) 180–199

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Computational neuroimaging strategies for single patient predictions

K.E. Stephan [a,b,c], F. Schlagenhauf [d,e], Q.J.M. Huys [a,f], S. Raman [a], E.A. Aponte [a], K.H. Brodersen [a], L. Rigoux [a,c], R.J. Moran [b,g], J. Daunizeau [a,h], R.J. Dolan [b,i], K.J. Friston [b], A. Heinz [d,j]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, 8032 Zurich, Switzerland
[b] Wellcome Trust Centre for Neuroimaging, University College London, London, WC1N 3BG, UK
[c] Max Planck Institute for Metabolism Research, 50931 Cologne, Germany
[d] Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité - Universitätsmedizin Berlin, 10115 Berlin, Germany
[e] Max Planck Institute for Human Cognitive and Brain Sciences, 04130 Leipzig, Germany
[f] Department of Psychiatry, Psychosomatics and Psychotherapy, Hospital of Psychiatry, University of Zurich, Switzerland
[g] Virgina Institute of Technology, USA
[h] ICM Paris, France
[i] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK
[j] Humboldt Universität zu Berlin, Berlin School of Mind and Brain, 10115 Berlin, Germany

ARTICLE INFO

Article history:
Received 27 November 2015
Revised 21 May 2016
Accepted 20 June 2016
Available online 22 June 2016

Keywords:
Generative model
fMRI
EEG
Bayesian
Model selection
Model comparison
Model evidence
Generative embedding
Classification
Clustering
Computational psychiatry
Translational neuromodeling

ABSTRACT

Neuroimaging increasingly exploits machine learning techniques in an attempt to achieve clinically relevant single-subject predictions. An alternative to machine learning, which tries to establish predictive links between features of the observed data and clinical variables, is the deployment of computational models for inferring on the (patho)physiological and cognitive mechanisms that generate behavioural and neuroimaging responses. This paper discusses the rationale behind a computational approach to neuroimaging-based single-subject inference, focusing on its potential for characterising disease mechanisms in individual subjects and mapping these characterisations to clinical predictions. Following an overview of two main approaches – Bayesian model selection and generative embedding – which can link computational models to individual predictions, we review how these methods accommodate heterogeneity in psychiatric and neurological spectrum disorders, help avoid erroneous interpretations of neuroimaging data, and establish a link between a mechanistic, model-based approach and the statistical perspectives afforded by machine learning.

Introduction

Despite its potential to provide a non-invasive assay of whole-brain function, neuroimaging has experienced surprising difficulties in delivering diagnostic applications for clinical practice, particularly in psychiatry. While there are several reasons for this disappointing track record, which is shared by other approaches like genetics (for discussions and reviews, see Casey et al., 2013; Kapur et al., 2012; Krystal and State, 2014; Stephan et al., 2015), the perceived failure has triggered important discussions about the most promising avenues for clinical neuroimaging. One particular hope is that the influx of methods from machine learning will realise the translational potential of neuroimaging. For example, in fMRI, multivariate classification and regression schemes have recently yielded impressive successes in clinically relevant domains such as pharmacological or pain research (e.g., Duff et al., 2015; Wager et al., 2013).

However, in direct application to clinical questions, neuroimaging-directed machine learning has mainly been used to discriminate patient groups from each other or from healthy controls. This has been variably successful, as indicated by the diverse outcomes from fMRI-based classification competitions. In these competitions, the accuracies of neuroimaging-based diagnoses have ranged from poor (e.g., attention deficit hyperactivity disorder; Brown et al., 2012) to excellent (e.g., schizophrenia; Silva et al., 2014). More importantly, however, the attempt to replace or augment traditional clinical diagnostics by applying machine learning techniques to neuroimaging data is a strategy of limited long-term clinical utility. This is because a multitude of physiological, genetic and clinical studies over the past decades have made it clear that mental diseases as defined by contemporary classification schemes – such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases (ICD) – are highly heterogeneous. That is, disorders like schizophrenia, depression,

http://dx.doi.org/10.1016/j.neuroimage.2016.06.038

autism etc. group patients with similar clusters of symptoms and signs that are caused by diverse pathophysiological mechanisms (Cuthbert and Insel, 2013; Kapur et al., 2012; Krystal and State, 2014; Owen, 2014; Stephan et al., 2016). This pathophysiological diversity explains why psychiatric diagnoses based on DSM/ICD have little predictive validity; that is, with few exceptions (such as differentiating mono- and bipolar affective disorders) they do not inform the clinician about individual clinical trajectories or treatment responses. As a consequence, even if highly accurate DSM/ICD diagnoses could be derived from machine learning classifiers applied to neuroimaging data, this would simply recapitulate a diagnostic scheme that does not directly inform clinical management – and would do so using a considerably more expensive and less widely available technology compared to classical psychiatric interviews.

This is one reason why the application of machine learning to neuroimaging data has changed direction in recent years and is now being increasingly applied to problems more directly related to clinical management, such as predicting individual disease course or treatment efficacy. This has shown some promising recent results, indicating that it may become possible to predict individual trajectories of patients with schizophrenia (Anticevic et al., 2015) or mood disorders (Lythe et al., 2015; Schmaal et al., 2015) from neuroimaging data, or forecast individual treatment responses to psychotherapy (Mansson et al., 2015), antidepressants (DeBattista et al., 2011; McGrath et al., 2013; Miller et al., 2013) and antipsychotics (Hadley et al., 2014; Nejad et al., 2013).

These are remarkable successes and raise hopes that neuroimaging may finally contribute to clinical decision-making in the not-too-distant future. However, the straightforward application of machine learning to neuroimaging data faces a number of non-trivial technical and conceptual challenges that may impede their long-term clinical utility (for discussions and reviews, see Brodersen et al., 2011; Klöppel et al., 2012; Orru et al., 2012; Wolfers et al., 2015). One central challenge is that neuroimaging data are noisy and very high-dimensional, presenting with a myriad of data features that could inform prediction. Dimensionality reduction and optimal feature selection thus become critical problems. One interesting development in this regard concerns recent advances in feature extraction methods based on restricted Boltzmann machines (Hjelm et al., 2014) and deep neural networks (Plis et al., 2014), which offer novel representations of disease states with potential diagnostic opportunities.

Second, hemodynamic and electrophysiological measurements represent distal and potentially complicated transforms of underlying neuronal mechanisms. This means that conventional machine learning methods, which operate directly on observed features of neuroimaging, do not furnish mechanistic insights into pathophysiology. This can be illustrated with three examples. First, in multivariate classification studies of fMRI, even though the spatial distribution of informative voxels can be determined, this does not disclose a concrete biological process. Similarly, unsupervised learning approaches that exploit multimodal imaging measures can obtain compelling subgroup delineations (Ingalhalikar et al., 2014) but remain descriptive and do not offer a mechanistic link between the structural and functional components of any identified predictor. Finally, while functional connectivity (i.e., statistical dependencies between regional time series) has enabled successful machine learning applications (Arbabshirani et al., 2013; Craddock et al., 2009; Du et al., 2015; Richiardi et al., 2011; Rosa et al., 2015), its characterisation of neuronal processes is restricted to statistical correlations that are agnostic about the physiological causes of network dynamics. In general, machine learning applied to "raw" neuroimaging data does not easily identify mechanisms from which novel therapeutic approaches could be derived.

An alternative to machine learning is the use of theory-driven computational models to infer pathophysiological mechanisms in individual patients (Friston et al., 2014; Huys et al., 2016; Maia and Frank, 2011; Montague et al., 2012; Stephan and Mathys, 2014). This strategy has a number of key features, which we discuss in detail below. In short, computational approaches (i) allow one to invoke model comparison procedures for clarifying whether any neurophysiological differences among patients signal real differences in pathophysiology, or simply reflect different cognitive strategies; (ii) provide theory-driven dimensionality reduction; and (iii) can support powerful single-subject predictions based on inferred mechanisms, as opposed to patterns of data features.

This paper provides an overview of computational neuroimaging strategies for single-subject predictions, independently of – or in conjunction with – machine learning techniques. We attempt to illustrate central concepts of generative modelling, and the clinical utility they may afford. To this end, we focus on the general form of model *classes*; by contrast, we do not discuss mathematical properties of any *single* model in detail, but refer the reader to the relevant literature. One area which requires a slightly more detailed mathematical discussion is the framework of Bayesian model comparison. Even here, however, we restrict our treatment to interpreting the general form of key equations.

This paper has the following structure. First, for readers without much background in computational neuroimaging, we provide a brief overview of existing approaches, clarify some nomenclature, and revisit some of its previous successes. Second, we discuss the importance of model comparison for dealing with heterogeneity across individuals and introduce the principles of Bayesian model selection (BMS) and Bayesian model averaging (BMA). Third, we outline how clinical predictions can be derived from computational models, (i) illustrating the use of BMS when theories of disease mechanisms exist and (ii) introducing generative embedding as a link between computational modelling and machine learning, when disease (process) theories are not available. Moreover, we outline how BMS and generative embedding can be deployed in an unsupervised or supervised fashion in order to address problems related to nosology (i.e., detecting subgroups in heterogeneous disorders), differential diagnosis, and outcome prediction.

Computational neuroimaging – what, why and how?

The term "computational" originally derives from the theory of computation, a subfield of mathematics that examines which particular functions are computable. A function is computable if it represents a mapping (from an input to an output set) that can be implemented by an algorithm; i.e., a well-defined sequence of operations. In neuroscience, the term "computational" has been used quite flexibly, ranging from an emphasis on information processing (irrespective of its biophysical implementation), to very broad usage, encompassing any algorithmic investigation of neuronal systems (cf. "computational neuroscience"), in contrast to analytical mathematical treatments.

In neuroimaging, three main computational approaches are presently being pursued (for review, see Stephan et al., 2015). These include biophysical network models (BNMs), generative models, and "model-based fMRI". BNMs are large-scale network models whose nodes represent mean field (or neural mass) models of neuronal population activity. Augmented with a hemodynamic or electrophysiological forward model, neuronal population activity is translated into a predicted fMRI or EEG signal (for reviews, see Deco et al., 2013a; Deco and Kringelbach, 2014; Wang and Krystal, 2014). By connecting the individual nodes in accordance with anatomical connectivity data – obtained from human diffusion-weighted imaging or Macaque tract tracing studies – the dynamics of large-scale networks and ensuing whole-brain neuroimaging signals can be simulated. While the neuronal state equations of BNMs can be rich in biophysical detail, their complexity renders parameter estimation very difficult, and current models allow only the estimation of a single global scaling parameter of connectivity (Deco et al., 2013b). For this reason, this paper focuses on the other two classes of models, generative models and model-based fMRI. These rest on less complex and fine-grained formulations, but allow for estimating model

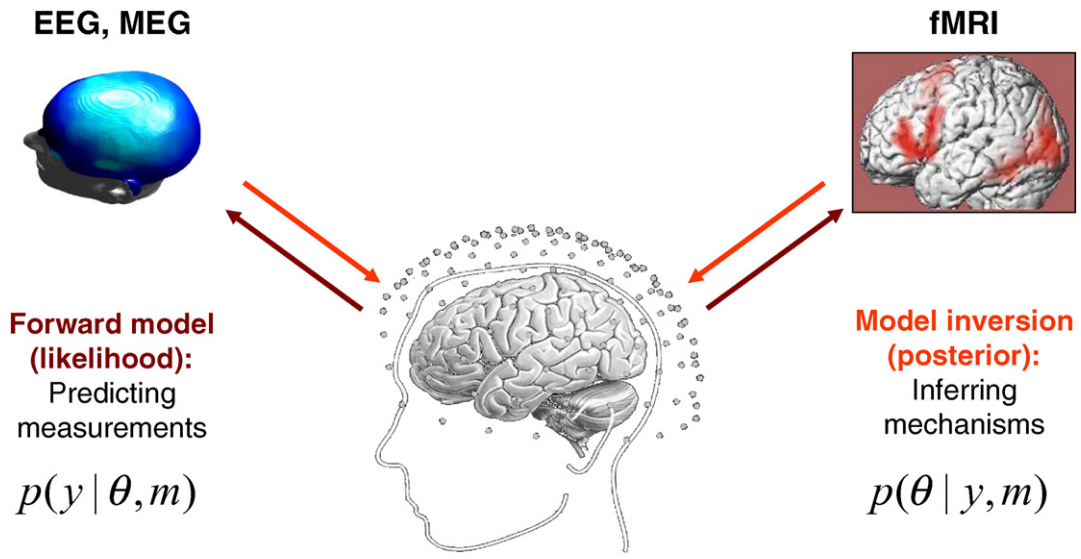

Fig. 1. Summary of a generative model for neuroimaging data. This figure contains graphics that are reproduced, with permission, from Chen et al. (2009), Garrido et al. (2008), and Stephan et al. (2003).

parameters from measured data. In the following, we summarise the principles of these models and clarify some of the technical terms and concepts involved. The following conventions will be used for equations: functions, distributions and scalar variables are represented by lowercase italics; sets, functionals and quantities from probability theory (such as information-theoretic surprise or free energy) by uppercase italics, vectors by lowercase bold, and matrices by uppercase bold letters.

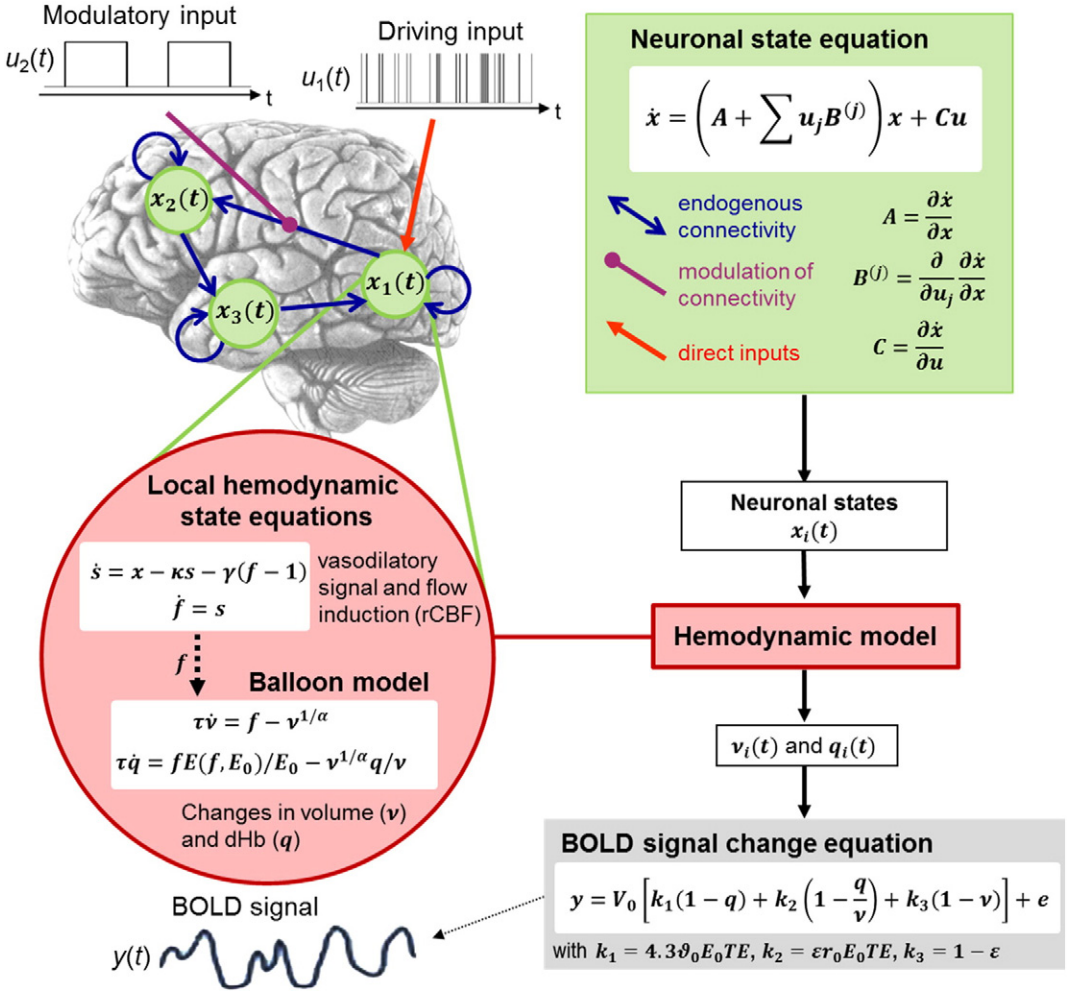

Fig. 2. Overview of DCM for fMRI. Reproduced, with permission, from Stephan et al. (2015).

Generative models of neuroimaging data

In statistics, a "generative model" is defined by the joint probability over all random variables (e.g., observed data and model parameters) that define a system of interest. More intuitively, one can think of a generative model as describing how observed data were generated and hence viewing it as a "recipe" for generating simulated data. A generative model is specified by defining two components: a likelihood function and a prior density. The likelihood function rests on a probabilistic mapping from hidden quantities (parameters $\boldsymbol{\theta}$) of the system of interest to observable quantities (measurements) $\mathbf{y}$:

$$\mathbf{y} = f(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{1}$$

This simply says that the data (feature) vector $\mathbf{y}$ originates from some transformation f, which encodes a putative signal-generating process, plus stochastic noise $\boldsymbol{\varepsilon}$. We deliberately write Eq. 1 in this form, as it will provide a useful reference for the distinction between feature selection methods that do or do not exploit knowledge about the hidden causes of measurements (see section on Generative Embedding below).

The function f encodes how system parameters determine its output (signal); this can range from extremely simple concepts (e.g. a constant term describing mean signal) to complex functions; e.g., a biophysically motivated dynamical system, as in the case of dynamic causal modelling (DCM, see below). Eq. 1 can now be used to specify the likelihood function as quantifying the probability $p(\mathbf{y}|\boldsymbol{\theta})$ of observing a particular measurement $\mathbf{y}$, given a particular parameterisation of the system. For example, assuming identically and independently distributed Gaussian noise $\boldsymbol{\varepsilon}$, the likelihood can be written as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{y}; f(\boldsymbol{\theta}), \sigma^2\mathbf{I}) \tag{2}$$

where $\mathbf{I}$ denotes the identity matrix and σ^2 noise variance.

The prior density encodes the range of values the parameters are expected to take a priori, i.e., before any data are observed. Again, under Gaussian assumptions we can express this as:

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \tag{3}$$

where $\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta$ denote prior mean and prior covariance, respectively.

To generate data, one could simply sample from the prior density and plug the ensuing parameter values into the likelihood function. This approach to simulating data is very general, and traditional non-probabilistic simulations in computational neuroscience (with fixed parameters) can be regarded as a special case of a generative model, where the prior density reduces to a Dirac delta function (point mass) over the chosen parameters.

Using the probabilistic mapping from hidden parameters of the system to observed signals in the "forward" direction is very useful for simulating observable responses, under different parameter values and exploring system behaviour. In neuroimaging, however, we wish to proceed in the opposite direction; i.e., estimate the parameter values from observed data. This reverse mapping is equivalently referred to as "model inversion", solving the "inverse problem", or simply "inference". Formally, this corresponds to computing the posterior probability $p(\boldsymbol{\theta}|y) = N(\boldsymbol{\theta}; \boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$ of the parameters, given the data (Fig. 1). This follows directly from Bayes theorem:

$$p(\boldsymbol{\theta}|\mathbf{y}, m) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|m)}{p(\mathbf{y}|m)} \tag{4}$$

Here, we have made the dependency on a chosen model structure m explicit by conditioning all terms on m. The practical difficulty is that deriving the term in the denominator requires computing an integral (see Eq. 9) which is usually not analytically tractable, except for some very simple cases. Unfortunately, even numerical integration is rarely feasible, since computation time increases exponentially with the number of model parameters. In practice, one has to resort to approximate inference schemes. These comprise two main approaches: Markov chain Monte Carlo (MCMC) sampling and variational Bayes (VB); for in-depth discussions see MacKay (2003) and Bishop (2006). MCMC is computationally expensive and can require very long run times but is guaranteed to converge to the correct solution (in the limit of infinite time). On the other hand, VB is computationally very efficient but is susceptible to local extrema and can be affected by violations of distributional assumptions (Daunizeau et al., 2011).

In this paper, we focus on the first and most widely used generative modelling framework for neuroimaging data, dynamic causal modelling (DCM). This approach was introduced a decade ago for fMRI (Friston et al., 2003). DCM for fMRI uses differential equations to describe the dynamics of neuronal population states $\mathbf{x}(t)$ that interact via synaptic connections and are subject to experimentally controlled perturbations $\mathbf{u}(t)$. These perturbations can either induce neuronal population activity directly; e.g., in terms of sensory stimulation ("driving input") or dynamically modulate the strengths of synaptic connections. The form of these neuronal state equations is given by a low-order (Taylor) approximation to any nonlinear system (Friston et al., 2003; Stephan et al., 2008), where the strengths of synaptic connections and weights of driving and modulatory inputs represent the parameters of interest we wish to infer by model inversion (see Fig. 2).[1]

The activity of each neuronal population is coupled to a regional blood oxygen level dependent (BOLD) signal by a cascade of differential equations describing hemodynamic processes, such as changes in blood flow and blood volume (Friston et al., 2000; Stephan et al., 2007).[2] Technically, this means that DCM represents a hierarchical generative model, where the likelihood function is partitioned into deterministic dynamic state equations f of hidden neuronal and hemodynamic processes (with neuronal and hemodynamic parameters $\boldsymbol{\theta}_n, \boldsymbol{\theta}_h$) and a static observation function g that implements discrete sampling and accounts for measurement noise $\boldsymbol{\varepsilon}$ (for the detailed equations, see Fig. 2 and Friston et al., 2003; Stephan et al., 2007):

$$\begin{aligned} &\text{Neuronal states}: && \frac{d\mathbf{x}_n}{dt} = f_n(\mathbf{x}_n, \mathbf{u}, \boldsymbol{\theta}_n) \\ &\text{Hemodynamic states}: && \frac{d\mathbf{x}_h}{dt} = f_h(\mathbf{x}_n, \mathbf{x}_h, \boldsymbol{\theta}_h) \\ &\text{Measurements}: && \mathbf{y} = g(\mathbf{x}_h) + \boldsymbol{\varepsilon} \end{aligned} \tag{5}$$

Notably, in these equations, noise only enters at the observation level whereas neuronal dynamics unfolds in a deterministic fashion, given external perturbations and system parameters (e.g., synaptic connection strength). That is, in Eq. 5, the dynamic variables of interest (neuronal and hemodynamic states $\mathbf{x}_n, \mathbf{x}_h$) are deterministic functions of designed and known inputs $\mathbf{u}$ and of time-invariant neuronal ($\boldsymbol{\theta}_n$) and hemodynamic ($\boldsymbol{\theta}_h$) parameters. (For ease of reading, we have omitted explicit references to time.) This means that we only need to infer the parameters – the state trajectories follow automatically from any inferred parameter values. This would be different if stochastic differential equations were chosen; in this case, the states are not fully determined by the choice of parameters and would need to be inferred, in addition to the parameters. This is known as stochastic DCM (Daunizeau et al., 2009; Daunizeau et al., 2012; Li et al., 2011).

By specifying plausible prior densities over the neuronal and hemodynamic parameters (see Friston et al., 2003 for details), the generative model is completed. Inverting this generative model allows one to infer the neuronal parameters of interest (e.g., coupling strengths and their modulation by experimental conditions) from empirically measured fMRI data. Notably, the separate representation of neuronal and

[1] Simulations that provide an intuition of neuronal dynamics accounted for by DCM and illustrate how different parameters impact on the resulting signals can be found in several previous papers; for example, see Fig. 1 in Penny et al. (2004b) and Fig. 2 in Stephan et al. (2008).

[2] For simulations illustrating the nature of this hemodynamic model, see Figs. 3, 8, 9 in Friston et al. (2000) and Fig. 5 in Stephan et al. (2007).

hemodynamic mechanisms is crucial for generative models of fMRI; since variability in neurovascular coupling across regions and subjects can otherwise confound inference on connectivity (David et al., 2008).

Following its introduction for fMRI, DCM has been extended to other neuroimaging modalities, including event-related potentials (ERPs; David et al., 2006), induced responses (Chen et al., 2008), and spectral responses (Moran et al., 2009), as measured by electroencephalography (EEG) and magnetoencephalography (MEG). DCM has found widespread use for analysis of effective connectivity between neuronal populations and has furnished insights into circuit-level mechanisms that eluded previous schemes. This includes, for example, physiological characterisations of predictive coding in cortical hierarchies during perceptual inference and learning, both in healthy subjects (e.g., den Ouden et al., 2009; Garrido et al., 2008; Summerfield et al., 2006) and in patients with schizophrenia (Dima et al., 2010; Dima et al., 2009; Ranlund et al., 2015) or altered levels of consciousness due to brain damage. Beyond long-range connectivity, DCM has also proven useful for inferring detailed, low-level physiological (synaptic) mechanisms within local neuronal circuits of the human brain, exploiting the rich temporal information contained by EEG/MEG data. Examples include the detection of conductance changes in AMPA and NMDA receptors under dopaminergic modulation (Moran et al., 2011), changes in postsynaptic gain of supragranular pyramidal cells in auditory cortex under cholinergic stimulation (Moran et al., 2013), or the characterisation of changes in neuronal physiology in individuals with selective mutations of particular ion channels (Gilbert et al., 2016).

Model-based fMRI

Generative models also play a central role in the second computational approach considered in this paper, "model-based fMRI". However, in contrast to the purely physiological DCMs described above, this approach asks whether a (particular component of a) computational process is reflected in BOLD signals (Gläscher and O'Doherty, 2010; O'Doherty et al., 2003). In other words, it tries to explain voxel-wise BOLD signals as a linear mixture of computational processes, which are assumed to be directly encoded by the underlying neuronal activity. The same approach can be applied, of course, to M/EEG responses; for example, in order explain trial-by-trial amplitudes or waveforms of event related potentials (Lieder et al., 2013; Ostwald et al., 2012). However, given its dominance in the present computational neuroimaging literature, we here focus entirely on model-based fMRI.

Model-based fMRI rests on a two-step procedure (see Fig. 2A). First, a generative model of behavioural responses, with computational states $\mathbf{x}_c$ (e.g., trial-wise prediction errors (PEs) or uncertainty) and parameters $\boldsymbol{\theta}_c$, are estimated using the measured behaviour $\mathbf{y}_b$ of an individual:

$$\mathbf{y}_b = g(\mathbf{x}_c, \boldsymbol{\theta}_c) + \varepsilon_b \tag{6}$$

By inverting this model, the computational states $\mathbf{x}_c$ can be inferred (dotted line marked with 1 in Fig. 2A). The subsequent convolution with a standard hemodynamic response function (HRF) then provides explanatory variables or regressors for a standard mass-univariate GLM analysis of voxel-wise fMRI data:

$$\mathbf{y}_{fMRI} = (\mathbf{x}_c \otimes HRF)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{fMRI} \tag{7}$$

The parameters of this GLM, $\boldsymbol{\beta}$, can now be estimated in a second inference step (dotted line marked with 2 in Fig. 2A), either under flat priors (i.e., maximum likelihood estimation) or using empirical Bayesian procedures (Friston and Penny, 2003). Overall, this two-step procedure enables one to search, across the whole brain, for the neuronal correlates of computational variables of interest which had been inferred from the simultaneously measured behaviour.[3]

[3] For a very instructive overview with both simulations and empirical results, please see Figures 1 and 2 in Glascher and O'Doherty (2010).

The model-based fMRI approach was pioneered by O'Doherty and colleagues (O'Doherty et al., 2003) who used a temporal difference (TD) learning model to show that phasic activation of the ventral striatum, a major recipient of dopaminergic projections from the midbrain, correlated with the magnitude of trial-wise reward PEs during an instrumental conditioning task. This was motivated by the seminal work of Schultz, Dayan, and Montague who found that reward PEs correlated with phasic activity of dopaminergic midbrain neurons (Schultz et al., 1997) and that changes in this phasic activity during learning could be predicted under the formalism of a TD learning model (see also (Montague et al., 2004) for review).

Model-based fMRI has subsequently been applied to numerous domains of cognition, accommodating a diversity of modelling approaches and computational themes, such as reinforcement learning (RL) models of behaviour and Bayesian models of cognition (e.g., (D'Ardenne et al., 2013; Daw et al., 2006; Iglesias et al., 2013; Klein-Flügge et al., 2011; Schwartenbeck et al., 2015; Seymour et al., 2004; Vossel et al., 2015). A recent application of model-based fMRI has been the investigation of interactions between learning and decision-making processes which do or do not derive from an explicit model of the environment or task structure. This distinction is commonly referred to as "model-based" vs. "model-free" computations (e.g., Daw et al., 2011; Deserno et al., 2015; Gläscher et al., 2010; Huys et al., 2012); where the latter term induces a terminological twist in the context of model-based fMRI.

Model-based fMRI in humans has produced results of high relevance for pathophysiological theories, corroborating, for example, hypothesised links between trial-by-trial activity in neuromodulatory nuclei and the trajectories of specific computational quantities suggested by theoretical accounts and/or animal experiments. Prominent examples include the encoding of reward PEs and precision (inverse uncertainty) by phasic and tonic changes in activity levels of the dopaminergic midbrain (e.g., D'Ardenne et al., 2013; Klein-Flügge et al., 2011; Schwartenbeck et al., 2015), or the reflection of expected and unexpected uncertainty (Yu and Dayan, 2005) by activity in the cholinergic basal forebrain (Iglesias et al., 2013) and noradrenergic locus coeruleus (Payzan-LeNestour et al., 2013). Model-based fMRI has also been applied to patients, for example, in depression (Dombrovski et al., 2013; Gradin et al., 2011) and addiction (Harle et al., 2015; Tanabe et al., 2013). Perhaps most notably, model-based fMRI studies of patients with schizophrenia (Gradin et al., 2011; Murray et al., 2008; Romaniuk et al., 2010) have contributed empirical evidence for the long-standing hypothesis that disturbances in PE signalling by dopaminergic neurons in the midbrain might assign "aberrant salience" to environmental events (Heinz, 2002; Kapur, 2003).

Hybrid and unified models

The relation between the behavioural and neuroimaging domains of model-based fMRI is summarised schematically in Fig. 3A. This highlights the fact (expressed by equations 6 and 7 above) that model-based fMRI essentially represents a correlational approach between two types of measurements, each of which has its own generative process. In the long run, unified models may be developed that explain both neuroimaging signals and behaviour of a given individual from the same underlying state equation (i.e., neuronal process). For example, this could be a state equation describing the biophysical implementation of relevant computations in a circuit of interest; this would require both a mapping from (hidden) neuronal states to behavioural observations that considers the biophysical implementation of relevant computations (e.g., predictive coding), and a mapping from circuit state to neuroimaging data describing how neuronal activity translates into measurable signals (Fig. 3B). This would allow one to infer circuit parameters of interest, simultaneously from both measured behaviour and neuroimaging data and provide a mechanistic (and quantitative) characterisation of neuronal processing that was grounded both in terms of physiology and computational function. Moreover, unified

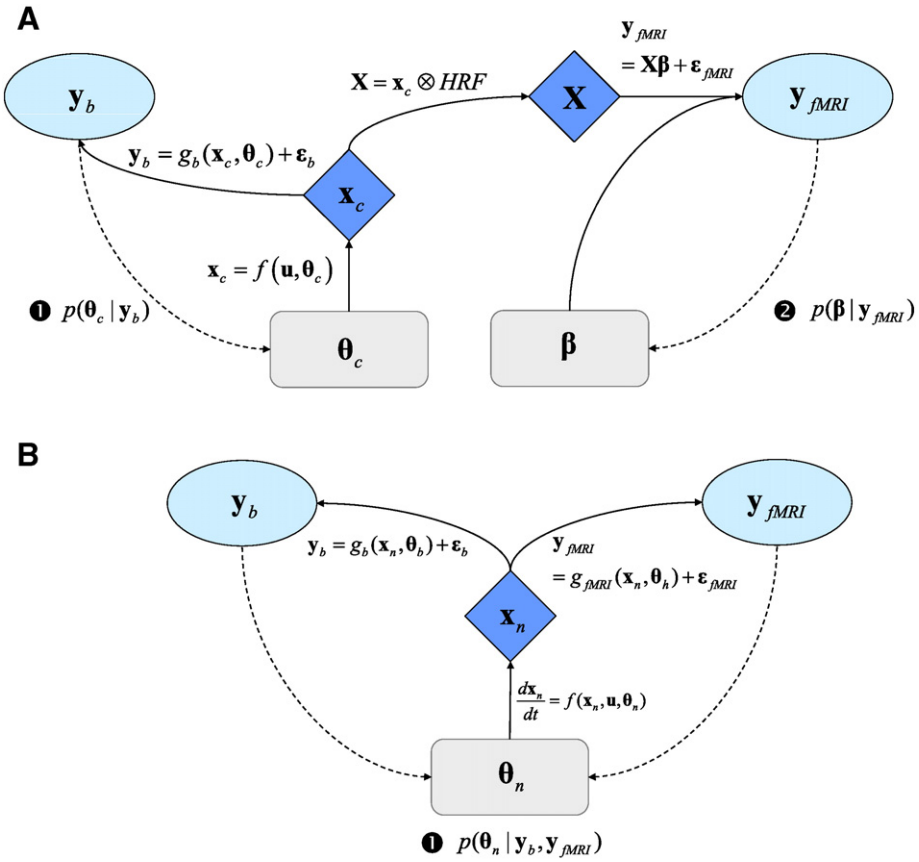

Fig. 3. A. Summary of the two-step procedure in "model-based" fMRI. Grey plates denote random variables; light blue ellipses represent observed variables (measurements); darker blue diamonds represent states (which follow deterministically, given the parameters and inputs). Solid arrows represent dependencies among variables; dashed arrows represent inference. B. Summary of a unified model in which both behavioural and neuroimaging data are predicted from the same underlying state equation. In this model, the unknown parameters can be inferred in one step, based on both behavioural and neuroimaging data. See main text for details and definition of variables.

models of this sort can help identifying which parts of a circuit are particularly critical for maladaptive actions. This might be particularly useful when addressing the problem of multiple potential causes or strategies underlying an observed behaviour; cf. (Schlagenhauf et al., 2014).

A significant step towards such a unified model has been made recently (Rigoux and Daunizeau, 2015). They proposed a mapping from a single state equation of circuit dynamics (based on the formalism of DCM for fMRI) to simultaneously acquired behaviour and fMRI signals (see their Figs. 3–5 for simulations that demonstrate the principles of this model). While this relatively coarse state equation only allows for relatively simplistic relations between circuit dynamics and behaviour, the conceptual advance of this model is considerable since it helps identifying which parts of the network (nodes or connections) are crucial for funnelling inputs (stimuli or task instructions) into behavioural outputs.

Prior work towards integrating generative models of neurophysiology and computation (information processing) have mainly examined the notion of PEs as "teaching signals" that regulate the amount of

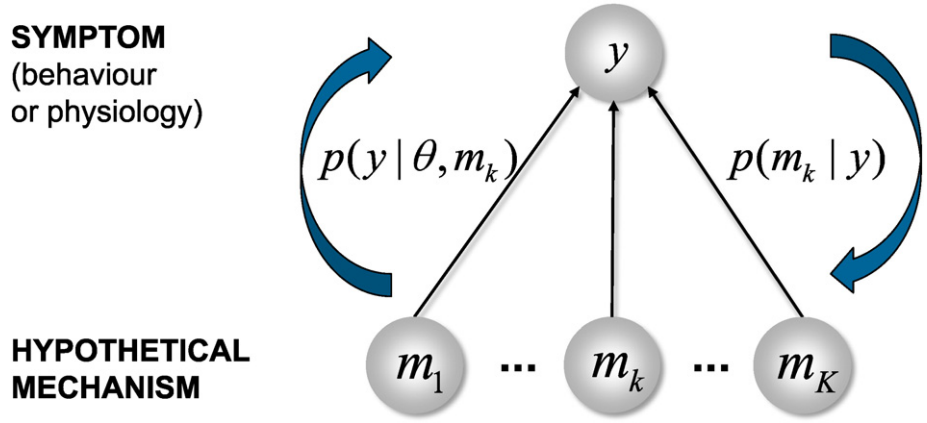

Fig. 4. Illustration that model selection can provide a formal basis for differential diagnosis. Here, the relative plausibility of a set of competing models, representing alternative mechanisms how the observed data could have been generated, is evaluated in terms of the posterior model probability. In the typical case of a flat prior on model space, the latter is identical to the model evidence (see main text for details).

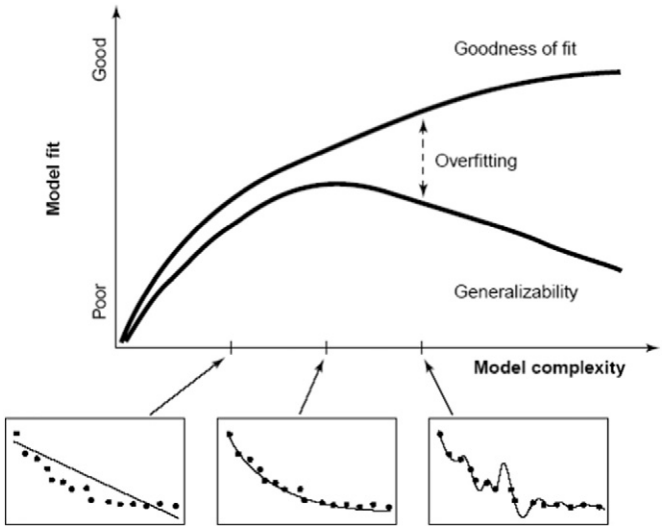

Fig. 5. An illustration of the trade-off between model fit and model complexity, and an example of overfitting. Here, models of increasing complexity are fitted to data that were generated from an exponential function, plus added observation noise. It can be seen that a highly complex model fits the data perfectly but, because it is trying to explain the noise as well, makes predictions (such as the pronounced bumps in the middle of the data series) which will not generalise across future instantiations of the data from the same underlying process ("overfitting"). Reproduced, with permission, from Pitt and Myung (2002).

synaptic plasticity needed to update neural circuits during learning. Specifically den Ouden et al. (2009, 2010) demonstrated that short-term plasticity during sensory learning could be measured by inferring how effective connection strengths were modulated by trial-by-trial prediction errors obtained from RL and hierarchical Bayesian models, respectively. A similar demonstration was provided in the context of learning under aversive outcomes (Roy et al., 2014). Most recently, Vossel et al. (2015) showed how attentional shifts were accompanied by changes in cortical-subcortical network connectivity that evolved according to trial-wise estimates of certainty (precision) of target predictions, where the latter were inferred from saccadic eye movement data using a hierarchical Gaussian filter (Mathys et al., 2011).

Bayesian model selection

As outlined in the Introduction, many conventionally defined neurological and probably all psychiatric diseases are highly heterogeneous: patients with similar symptoms and behaviour may differ considerably in terms of (patho)physiological mechanisms and/or cognitive processes. A central goal for neuroimaging approaches with the ambition of clinical utility is thus to identify, in any given individual patient, the most likely mechanism that underlies a particular observation (brain activity pattern). This is simply the challenge of differential diagnosis, which is ubiquitous throughout medicine. Differential diagnosis maps directly onto hypothesis testing which, in parametric statistics, corresponds to the formal comparison of different models of how observed data could have been generated (Fig. 4). In other words, an elegant approach to establishing differential diagnoses in psychiatry based on neuroimaging would be to formalise competing pathophysiological theories in terms of alternative generative models. The relative plausibility of these models (hypotheses) would then be evaluated by formal model comparison procedures, given empirical measurements of neuroimaging and/or behaviour.

As a concrete example, many pathophysiological concepts of schizophrenia converge on the notion of dysregulation of dopaminergic midbrain neurons in patients with schizophrenia (Heinz, 2002; Kapur, 2003; King et al., 1984; Winton-Brown et al., 2014). This dysregulation could be caused by at least three different mechanisms (for details, see (Adams et al., 2013; Stephan et al., 2009a): (i) altered prefrontal inputs that target midbrain neurons via NMDA receptors; (ii) enhanced inputs from cholinergic brainstem nuclei (PPT/LDT), or (iii) altered autoregulation of dopaminergic midbrain neurons (by paracrine release of dopamine and activation of dopaminergic autoreceptors). Disambiguating between these possibilities, by comparing models that embody the above mechanisms (given measurements from the midbrain and the areas it communicates with), would have tremendous relevance for delineating schizophrenia into pathophysiological subgroups – and for guiding individual treatment decisions.

In what follows, we unpack the statistical basis of differential diagnosis by model selection. We hope to familiarise the reader with Bayesian techniques for comparing and selecting models that are used frequently in the current literature and provide a powerful way to deal with individual variability in physiology and/or computation. These techniques are equivalently referred to as Bayesian model comparison (BMC) or Bayesian model selection (BMS); while the former is the more general term, the latter describes the common situation of selecting a single (most plausible) model from a set of alternatives.[4]

Model evidence

Generally, the first step of any model-driven (hypothesis-led) investigation is to decide which class of explanation accounts for the observations. This is what all scientists implicitly do when testing hypotheses – although this step might not always represent an explicit choice. Technically, hypothesis testing or model comparison corresponds to defining a hypothesis set or model space M of competing explanations that are deemed plausible *a priori*. This is equivalent to specifying a prior over models; where, typically, all models within M are considered equally likely and all other possible models have zero prior probability:

$$p(m) = \begin{cases} 1/|M| \text{ if } m \in M \\ 0 \text{ if } m \notin M \end{cases} \quad (8)$$

(Here, $|M|$ refers to the cardinality of the hypothesis set.) The challenge then is to find the model m within M that provides the best explanation of the observed data. Importantly, selecting a model that "best explains" the data is not simply a statement about model fit. Indeed, it is trivial to find, for any data set, models with excellent or even perfect fit; for example, for any observation consisting of t data points, a polynomial function of order $t-1$ will fit the data perfectly. These overly accurate models simply explain noise or random fluctuations that are specific to the particular measurement and do not generalise to other (e.g., future) measurements of the same process. This tendency of an overly flexible model to recognise spurious patterns in noise is referred to as "overfitting" (see Fig. 5 of this paper and Fig. 1.4 of Bishop (2006) for examples). On the other hand, the simplest model possible, which would consist of a constant term only and explains no signal variance (i.e., $R^2 = 0$), can indeed be the best explanation of a time series – when the time series contains no signal and only noise. In summary, measures of fit alone are inadequate to judge model goodness (Pitt and Myung, 2002). Instead, the challenge is to select models that generalise best; these are the models that provide an optimal balance between fit (accuracy) and complexity.

This balance is implicit in the Bayesian model evidence used during Bayesian model selection (BMS). The model evidence is the probability of observing the data y given the model m. This probability is also referred to as the marginal or integrated likelihood and corresponds to the denominator from Bayes theorem (see Eq. 4). It can be computed

[4] While this paper focuses on the conceptual and mathematical foundations of BMS, previous papers have provided toy examples (simulations) and step-by-step BMS analyses of single subject data which may be useful for the interested reader. For example, for simulations, please see Figures 4–6 and Tables 2–5 in Penny et al. (2004a) and Figure 2 in Stephan et al. (2009b); for detailed single subject BMS analyses, please see Figures 7–9 and Tables 6–13 in Penny et al. (2004a) and Figures 3–4 and Table 1 in Stephan et al. (2005).

by integrating out (or marginalising) the parameters from the joint probability:

$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|m)d\boldsymbol{\theta} \tag{9}$$

This is why the Bayesian model evidence is also referred to as the marginal likelihood. As a simplifying intuition, the evidence can be understood as providing an answer to the question: "If I randomly sampled from my prior and plugged the resulting value into the likelihood function, how close would the predicted data be – on average – to my observed data?"

In practice, model comparison does not utilise the model evidence directly but typically employs its logarithm (log evidence). Given the monotonic nature of the logarithmic function, ranking models based on either model evidence or log evidence yields identical results. However, the log evidence is numerically easier to deal with (the logarithm of a small number between zero and one is a large negative number) and results in more intuitive equations, some of which we encounter below. It also offers an additional nice intuition derived from information theory. Specifically, given that (Shannon) surprise S is defined as negative log probability, for an agent operating under a given model m, the log evidence corresponds to the negative surprise about observing the data y:

$$\log p(\mathbf{y}|m) = -S(\mathbf{y}|m) \tag{10}$$

Put simply, log evidence – and hence model goodness – increases when we are less surprised about the data encountered.

While the statistical procedure of model comparison typically rests on the log evidence, the result of comparing two models is understood more intuitively when reported as a Bayes factor (BF); this is simply the ratio of two model evidences. As with p-values in frequentist statistics, conventions exist about which thresholds are meaningful for Bayes factors (Kass and Raftery, 1995). For example, a Bayes factor larger than 20 (equivalent to a log evidence difference larger than 3) would be considered as "strong" evidence in favour of one model relative to another.

An alternative option, when reporting the results of model comparisons in an intuitively accessible form, is to compute, for each model m_i, its posterior probability. In the typical case where the prior on models is uninformative or flat (cf. Eq. 8), this simplifies to normalising the evidence for each model by the sum of all model evidences:

$$p(m_i|\mathbf{y}) = \frac{p(\mathbf{y}|m_i)p(m_i)}{\sum_{j=1}^{|M|} p(\mathbf{y}|m_j)p(m_j)} = \frac{p(\mathbf{y}|m_i)}{\sum_{j=1}^{|M|} p(\mathbf{y}|m_j)} \tag{11}$$

This makes it easy to see that across all models, the posterior model probability sums to unity.

Approximations to the log evidence

One major barrier to computing the model evidence is that the integral in Eq. 9 can rarely be evaluated analytically; furthermore, numerical integration is typically prohibitively expensive. Therefore, one usually resorts to approximations of the log evidence, such as the Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), or negative free energy (Friston et al., 2007; Neal and Hinton, 1998; Penny et al., 2004a). These approximations decompose model goodness into a balance of two terms – accuracy and complexity. All of them agree in the definition of accuracy as log likelihood. By contrast, they differ considerably in their approximation of complexity.

AIC and BIC have a seemingly straightforward approximation of complexity. In AIC complexity simply corresponds to the number of free parameters; in BIC, this is additionally scaled by the log number of data points:

$$\begin{aligned} AIC &= \log p(\mathbf{y}|\boldsymbol{\theta}, m) - k \\ BIC &= \log p(\mathbf{y}|\boldsymbol{\theta}, m) - k\frac{\log n}{2} \end{aligned} \tag{12}$$

The additional scaling factor in BIC means that, once that more than $n > 8$ data points are available, BIC entails a stronger complexity penalty than AIC. The simplicity of their complexity approximations makes AIC/BIC easy to compute, but has two significant disadvantages: AIC/BIC ignore interdependencies among model parameters (which are ubiquitous in biological systems; Gutenkunst et al., 2007) nor can they capture differences in prior variance across parameters.

These issues are resolved by a third approximation to the log evidence, the negative free energy F. Its name derives from close connections between free energy concepts in statistical physics and variational approaches to probability theory (see Friston et al., 2007; Neal and Hinton, 1998). Variational free energy represents a lower bound approximation to the log evidence, where the tightness of the bound depends on how well the true (but unknown) posterior can be matched by an approximate posterior q (of known form):

$$\log p(\mathbf{y}|m) = F + KL[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathbf{y}, m)] \tag{13}$$

Here, KL refers to Kullback-Leibler divergence or relative entropy, an information theoretic measure of the dissimilarity between two probability densities. The KL divergence is zero when the densities are identical and becomes increasingly positive the more the two densities differ (Kullback and Leibler, 1951). Importantly, since we do not know the true posterior, the KL term cannot be evaluated directly. However, by maximising F one implicitly minimises the KL term, thus tightening the lower bound approximation to the log evidence (see Fig. 6). This is achieved by optimising the approximate posterior q (e.g., when q is Gaussian, finding the mean and variance of q that maximises F according to Eq. 13 above). In other words, by maximising F we can both obtain an approximation to the log evidence and the posterior densities of the parameters.

The negative free energy (and hence model evidence) can be decomposed into the following balance between model fit and model complexity (for details, see Penny et al., 2004a; Stephan et al., 2007):

$$F = \langle \log p(y|\theta, m) \rangle_q - KL[q(\theta)\|p(\theta|m)] \tag{14}$$

In this expression, the first term represents accuracy: the expected log likelihood, under a chosen approximate posterior q. The second term represents complexity and is given by another KL divergence; this time between the approximate posterior and the prior. When the form of the approximate posterior is the same as the true posterior, the complexity is exactly the difference between the posterior and prior and inference becomes exact. Put simply, this means that a model has high complexity if it is sufficiently flexible to allow for a substantial belief update, i.e., a pronounced divergence of the posterior from the prior belief. Another heuristic is that the complexity reflects both the effective number of parameters that need to be displaced from their prior values to provide an accurate explanation for data and the degree of their displacement.

This heuristic can be turned into a more formal perspective by examining the analytical expression of the complexity term under an assumed distributional form for the approximate posterior (see discussions in Stephan et al. (2009b) and Penny (2012)). For example, under Gaussian assumptions:

$$\begin{aligned} &KL[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|m)] = \\ &\frac{1}{2}\log(\det(\mathbf{C}_\theta)) - \frac{1}{2}\log(\det(\mathbf{C}_{\theta|y})) + \left(\boldsymbol{\mu}_{\theta|y} - \boldsymbol{\mu}_\theta\right)^T \mathbf{C}_\theta^{-1}\left(\boldsymbol{\mu}_{\theta|y} - \boldsymbol{\mu}_\theta\right) \end{aligned} \tag{15}$$

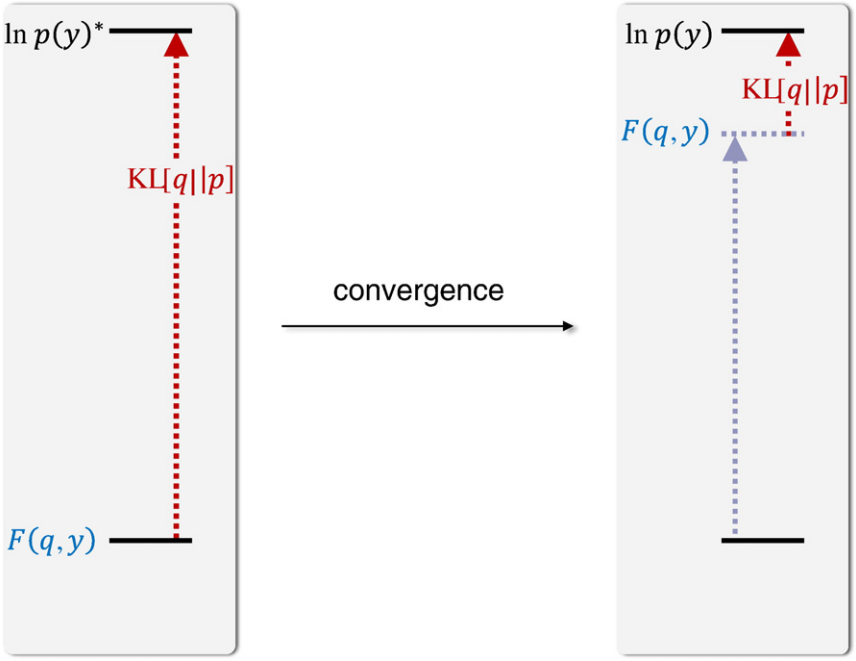

Fig. 6. A graphical illustration of the negative free energy approximation to the log model evidence, and its evolution during model inversion in the context of variational Bayes. Here, by adjusting the parameters of the approximate posterior such that the negative free energy *F* is maximised, one implicitly minimises the KL divergence between the approximate and true posterior and tightens the bound on the log evidence. See main text for details.

Here, $\mathbf{C}_\theta$ and $\mathbf{C}_{\theta|y}$ denote prior and posterior covariance matrices, and det refers to the determinant, a matrix property that can be interpreted as a measure of "volume" (the space spanned by the eigenvectors of the matrix). This volume increases with the number of dimensions (the rank of the covariance matrix), and with the length and orthogonality of the basis vectors. With this in mind, the first term in Eq. 15 means that complexity increases with the number of free parameters, the more flexible these parameters are (the higher their prior variance), and the more orthogonal they are. The second term means that complexity decreases with increasing orthogonality of the posterior parameter estimates (a desirable property of an interpretable model) and with increasing posterior variances (highly precise posterior estimates result in brittle model predictions which are unlikely to generalise). Finally, the third term captures our heuristic above and expresses that complexity grows the more the posterior mean diverges from the prior mean.

While all of the above approximations have proven useful in practice, they come with different pros and cons.[5] AIC and BIC are easy to compute since the log likelihood is always available and estimating complexity boils down to simply counting the number of free parameters. On the downside, AIC and BIC are agnostic to several important aspects of complexity, such as the prior variance of and interdependence among parameters. By contrast, the free energy approximation provides a more informed measure of complexity that is generally more appropriate for real-world biological systems which are imbued with parameter interdependencies (Gutenkunst et al., 2007). However, distributional assumptions may have greater impact than for BIC and AIC (since they concern not only the accuracy, but also the complexity term), and evaluating the tightness of its bound approximation requires computationally expensive sampling schemes (Aponte et al., 2016). On the other hand, the negative free energy approximation was shown to exhibit better model comparison performance than AIC/BIC in the context of regression models and DCM (Penny, 2012) and also proved superior to BIC for model comparison of directed acyclic graphical models (Beal and Ghahramani, 2003).

An alternative to the above approximations are sampling-based approaches, typically based on MCMC. One (highly simplified) way to think of MCMC – in this particular context – is of reconstructing an integral by an approximation (essentially like a histogram) that is "experienced" by a random walk. Here, each step only depends on the previous one (Markov chain) and tends to move in a direction that is likely to provide a meaningful contribution to the integral. Depending on how much computation time one is willing to invest, different options exist. A computationally less expensive approach is to use a single chain for obtaining samples from a particular distribution and using these samples to evaluate Eq. 9, in order to obtain an approximation to the model evidence. For example, using the chain to sample from the prior leads to the prior arithmetic mean (PAM) approximation (which tends to underestimate the model evidence), whereas sampling from the posterior distribution leads to the posterior harmonic mean (PHM) approximation (which tends to overestimate the model evidence).

A more robust alternative is multi-chain sampling. The key idea here is to build a sequence (path) of probability distributions that connect the prior to the posterior distribution by using a temperature parameter on the likelihood part of the model (Lartillot and Philippe, 2006). Independent single chains can be used to obtain samples from each distribution from this sequence; joining the samples from multiple chains yields an asymptotically exact estimate of the log evidence. An additional improvement is to use population MCMC (Calderhead and Girolami, 2009) which imposes a dependency between neighbouring distributions (chains) in the sequence and improves the samples obtained from each single chain.

While sampling-based approaches to the log evidence are a promising directing for future developments of BMS, they have usually been prohibitively expensive (in terms of compute time) so far. However, recent advances in exploiting the power of graphics processing units (GPUs) are now beginning to turn sampling approaches, including multi-chain and population MCMC methods, into a viable alternative for computing accurate log evidence estimates (Aponte et al., 2016).

[5] Several simulation studies have examined the validity of these approximations in the context of the models discussed in this paper; for example, see Figure 4 in Penny et al. (2004a), Figures 3–5 in Stephan et al. (2008), and Figures 6–8 in Penny (2012).

Inferring model structure in individual subjects vs. groups

Computational modelling studies of neuroimaging and behavioural data have largely applied model comparison at the group level. By contrast, this paper focuses on the need for disambiguating and quantifying pathophysiological or psychopathological processes in individual patients. For this reason, we will only briefly touch on group-level model comparison methods and refer the interested readers to previous publications and reviews (Friston et al., 2016; Huys et al., 2011; Penny et al., 2010; Rigoux et al., 2014; Stephan et al., 2009b; Stephan et al., 2010).

Group-level model comparison faces the same general challenge as any other statistical inference procedure at the group level, namely does the quantity of interest constitute a fixed or a random effect in the population? Under a fixed-effects perspective, one assumes that the variable of interest is constant across subjects, and any variability in the measured data arises from observation noise. In the context of model comparison, this means that the model is assumed to be identical across subjects. This allows for a straightforward statistical procedure: because the data obtained from different subjects are independent, one can simply pool the evidence across the sample and multiply individual Bayes factors, resulting in a group Bayes factor (Stephan et al., 2007).

Fixed-effects model comparison is simple to perform and has great sensitivity. However, the underlying assumption is usually incompatible with the heterogeneity of patient groups and it therefore rarely has a place in clinical studies. Even in the healthy population, however, variability in cognitive and physiological mechanisms can be substantial, and a fixed-effects approach is typically reserved for situations where it can be assumed that the mechanism of interest is the same across subjects, e.g., basic anatomical or physiological phenomena, such as the relation between structural and effective connection strengths (Stephan et al., 2009c). In all other applications, a random effects perspective is usually more appropriate.

However, this statement has to be qualified by asking where random (between-subject) effects arise. This can be at the level of parameters or models. In other words, one could assume that different subjects have the same basic architecture but each subject has unique and unknown (random) model parameters. In other words, group data could be generated by sampling from subject-specific distributions over all model parameters but under the same model.[6] Alternatively, one could assume that the parameters of a model are sampled from different models, where some models preclude various parameters.[7] The first approach (random parametric effects) calls for hierarchical or empirical Bayesian models of group data; while the second (random model effects) approach can be implemented using just the model evidences over different models for each subject. A simple frequentist approach to random parametric effects is to use parameter estimates from each subject, obtained under the same model, as subject-wise summary statistics that enter a second-level (parametric or non-parametric) test which probes the null hypothesis of no parametric effects.

More sophisticated, fully hierarchical approaches extend this summary statistic approach to provide so-called "empirical Bayesian" models of group data (Friston et al., 2016; Huys et al., 2011). They are "empirical" because their hierarchical structure allows for estimating subject-level priors from the group data (Huys et al., 2011). The advantage of this is that subject-wise estimates are less noisy; however, the parameter estimates are no longer independent across subjects. To repeat, in this setting, all subjects are assumed to be drawn from the same model, i.e., this is a random effects approach in parameters but not models (random parametric effects). A simple parametric manner to relax this is to fit a mixture of models to each subject and infer individual and group weighting parameters. This assumes a within-subject multiplicity of generative processes.

The equivalent hierarchical treatment of random model effects uses just the log evidences for each model (as opposed to the parameter estimates). A hierarchical model of the log evidences across the population was introduced by Stephan et al. (2009b). This model accommodates model heterogeneity in the studied sample and computes, for each model m considered, the expected model frequency (i.e., the prevalence of model m in the population from which the group of subjects is sampled) as well as its exceedance probability (i.e., the posterior probability that its frequency is higher than the frequency of any other model considered). This approach has recently been finessed by Rigoux et al. (2014) who introduced "protected exceedance probabilities" that account for the possibility that observed differences between model evidences may have arisen by chance.

Applications of BMS to questions of diagnosis and pathophysiology

BMS represents a principled approach to deciding which of several hypotheses (models) best explains the observed data of an individual subject. As described above, in principle, it provides an attractive foundation for establishing computational assays to address diagnostic questions. That is, provided one has well-founded theories about alternative pathophysiological or pathocomputational mechanisms underlying a given symptom, and these theories can be cast as competing models, model comparison could provide a direct and formal basis for differential diagnosis in individual patients (Stephan and Mathys, 2014; Fig. 4).

So far, however, subject-by-subject BMS has rarely been applied in clinical studies. A major reason for this gap may be that, at least in psychiatry, we do not always have precise hypotheses about alternative disease mechanisms. Alternatively, when such theories do exist, it may not be straightforward to represent them in concrete models that can be implemented under existing computational frameworks – or which can explain the measurements that can be acquired from the neuronal circuit of interest.

Some notable exceptions exist. These include studies that carefully disambiguated alternative cognitive strategies across patients, as in the study by Schlagenhauf et al. (2014) described below, or studies examining rare patients with neurological conditions. For example, Cooray et al. (2015) used BMS to compare alternative explanations of seizure activity in two patients with anti-NMDAR encephalitis. Applying a DCM to artefact-free EEG data acquired during the occurrence of seizures, the authors compared different theoretical formulations of how alterations of excitatory and/or inhibitory connectivity by NMDAR pathology could lead to seizure activity. In both patients, this model comparison provided a non-trivial explanation of seizure initiation, highlighting a characteristic increase in inhibitory connectivity at seizure onset and a subsequent increase in excitatory connectivity.

While the application of BMS to individual patients is the exception so far, many computational neuroimaging studies conducted in recent years, particularly those using DCM, have adopted random effects model comparison. These studies address a wide range of syndromatically defined disorders; for example, schizophrenia (Dauvermann et al., 2013; Deserno et al., 2012; Schlagenhauf et al., 2014; Schmidt et al., 2013), bipolar disorder (Breakspear et al., 2015), or Parkinson's disease (Rowe et al., 2010). Although most of them do not directly address a diagnostic problem (but see Boly et al., 2011), these studies provide important insights into pathophysiological mechanisms, while respecting heterogeneity across patients.

The remainder of this section discusses three examples from recent work by different groups. These examples illustrate a spectrum of hypothesis testing strategies afforded by BMS, concerning pathophysiological differentiation of patient groups, detection of patient subgroups, and identifying potential mechanisms (and targets) of therapeutic interventions, respectively.

The first example illustrates how model selection can help distinguish between pathophysiological explanations and support diagnostic differentiation of patient groups. (Boly et al., 2011) sought to identify potential mechanisms for diminished consciousness levels in two

[6] For simulations of this case, see Figure 3 in Friston et al. (2016).

[7] Example simulations of this scenario can be found in Figure 2 of Stephan et al. (2009b).

groups of brain-damaged patients (N = 21) with "minimally conscious state" (MCS) and "vegetative state" (VS), respectively. The differential diagnosis of these two disorders of consciousness by clinical means is difficult, and neurophysiological biomarkers for disambiguating these two states would be highly desirable. Comparing these two groups and a healthy control group with EEG, Boly and colleagues found a differential reduction in long-latency components of the mismatch negativity (MMN), an event-related response to surprising (auditory) stimuli. Using DCM and BMS, they asked whether pathological changes in a five-area network (primary and secondary auditory cortex bilaterally, right IFG) – known to generate the MMN in healthy volunteers – could explain this neurophysiological difference across groups. Specifically, they tested whether the hierarchically higher areas, secondary auditory cortex and IFG, might have become disconnected from primary auditory cortex. This was not the case: BMS indicated that the same five-area network most plausibly generated the measured responses in both patient groups and healthy controls. Subsequent statistical analysis of posterior connectivity estimates (under the most plausible model) indicated that the selective impairment of a single connection – from IFG to secondary auditory cortex in the right hemisphere – accounted for the scalp-wide ERP abnormalities in VS, but not MCS, patients. This result suggests that MCS and VS may be differentiated by a reduction of top-down influences from frontal to temporal cortex, and more generally, that a disturbance of recurrent message passing in sensory processing hierarchies might offer a hallmark for diminished consciousness.

A second case study shows how BMS can enable characterisation of subgroups within a heterogeneous sample (van Leeuwen et al., 2011). Strictly speaking, this example does not concern a clinical disorder, but a relatively rare cognitive anomaly: colour-grapheme synaesthesia. Individuals with this type of synaesthesia experience a sensation of colour when reading letters. As shown in previous fMRI studies (Hubbard et al. 2005; Weiss et al. 2009), this experience is paralleled by an activation of the "colour area" V4; a phenomenon for which two competing explanations have been offered. One idea is that V4 might be activated in a bottom-up fashion through input from a grapheme-processing area in the fusiform gyrus. An alternative proposal is that V4 might be activated top-down, via parietal areas that "bind" information on colours and letters. van Leeuwen et al. (2011) cast these two opposing theories into distinct DCMs (see Fig. 7), which they applied to fMRI data from 19 individuals with synaesthesia. Random effects BMS applied to the whole population failed to find a clear difference between the two models. Strikingly, however, when considering individual differences in the phenomenology of the experience, two neurophysiologically distinct subgroups emerged: in individuals who experienced colour as being externally co-localised with letters ("projectors") the DCM implementing the bottom-up hypothesis had considerably higher evidence. By contrast, in subjects experiencing an internal association between colour and letters ("associators"), the DCM representing the top-down hypothesis was clearly superior. While this perfect dichotomy was established by means of a random-effects group BMS procedure, eyeballing the graphical representation of the individual log evidence differences (Fig. 2 in van Leeuwen et al., 2011) allows for an approximate estimate of the diagnostic properties of this computational assay. For example, using a "positive evidence" threshold (Kass and Raftery, 1995), the BMS procedure appears to have a sensitivity of 80% and a specificity of 100% for detecting "associators".

The final and third example demonstrates how potential mechanisms of pharmacological interventions can be identified using model comparison. This is an important issue for pharmacological fMRI studies that contend with individual variability in receptor expression, receptor sensitivity, and/or transmitter metabolism. This variability can introduce considerable heterogeneity in fMRI responses and connectivity estimates under pharmacological manipulation – even when dosage is carefully controlled for – and confounds the detection of any significant effects at the group level (for example, see van Schouwenburg et al., 2013). A recently introduced BMS strategy offered an innovative approach to circumvent these issues (Piray et al., 2015). This study related previous findings from pharmacological studies in rodents to humans in order to clarify how the putative dopamine-dependency of connections intrinsic to the human basal ganglia relates to individual trait impulsivity. The authors acquired task-free fMRI data in healthy volunteers, using a within-subject, cross-over factorial design combining D2 agonists (bromocriptine) and antagonists (sulpiride) with placebo. By modelling the pharmacological intervention as an unsigned session-wise modulatory effect in a DCM of the basal ganglia, Piray et al. were able to identify those connections that were sensitive to dopaminergic manipulations; regardless of heterogeneity of dopamine physiology across subjects and the ensuing variability in signed connectivity estimates. Using this strategy, they were able to show that stimulation and blockade of D2 receptors exerted opposing effects on a specific set of connections – and demonstrate that trait impulsivity was related to the dopaminergic sensitivity of the connection from ventral striatum to the dorsal caudate nucleus.

Disentangling pathophysiology from differences in cognitive strategy

The considerable heterogeneity of patients in psychiatric and neurological spectrum diseases not only poses a challenge for differential diagnosis, but also introduces an important caveat for the investigation of pathophysiology with neuroimaging. Specifically, if the investigator is unaware that patients engage in different cognitive processes, apply diverging strategies or operate under fundamentally different beliefs in solving a given task, the ensuing differences in brain activity can be falsely interpreted as pathophysiological differences. We illustrate this important point with an empirical example from a recent model-based fMRI study:

Schlagenhauf et al. (2014) compared patients with schizophrenia and healthy controls during a probabilistic reversal learning task. In this study, subjects were required to decide between two choices, probabilistically rewarded in 80% and 20% of trials, respectively. Importantly, the preferable choice changed from time to time (reversal), following a probabilistic rule that the participants did not know. In solving this task, different structural aspects of the task could potentially guide an individual's behaviour: (i) sensitivity to reward and punishment, (ii) mirror symmetry in outcome probabilities, and (iii) probabilistic timing of reversals. For each subject, a set of computational models capturing these aspects to various degrees were compared, in order to determine the driving factors behind an individual's choices. The models considered included variants of the Rescorla-Wagner (RW) model, a classical reinforcement learning model in which stimulus-response links are updated through PEs (Rescorla and Wagner, 1972), and variants of Hidden Markov models (HMM) which represent subjective belief trajectories (here, which response option is presently more likely to be rewarded and how likely it is to switch). Unlike the RW model variants, the HMM can also capture the statistical structure of sudden reversals.

Importantly, this study used a formal model comparison procedure (Huys et al., 2011) to evaluate the plausibility of each model in each group; additionally, it tested for each model whether it provided an above-chance explanation of the behaviour of each individual subject. The results indicated that, overall, healthy subjects' behaviour was best described by a HMM with differential reward and punishment sensitivities. By contrast, among patients, two distinct subgroups of patients were revealed by model comparison (Fig. 7). In one schizophrenia subgroup, the HMM was similarly convincing as in controls, indicating that those patients used a similar strategy to solve the task. However, the second group displayed behaviour poorly explained by the HMM. These patients were characterized by higher positive symptoms, showed lower performance and responded more randomly, and their behaviour was best (although still poorly) explained by a simpler RW model which did not incorporate a representation of reversal probability. Notably, the different behaviour of this group was not simply explained by lower premorbid IQ or attentional deficits.

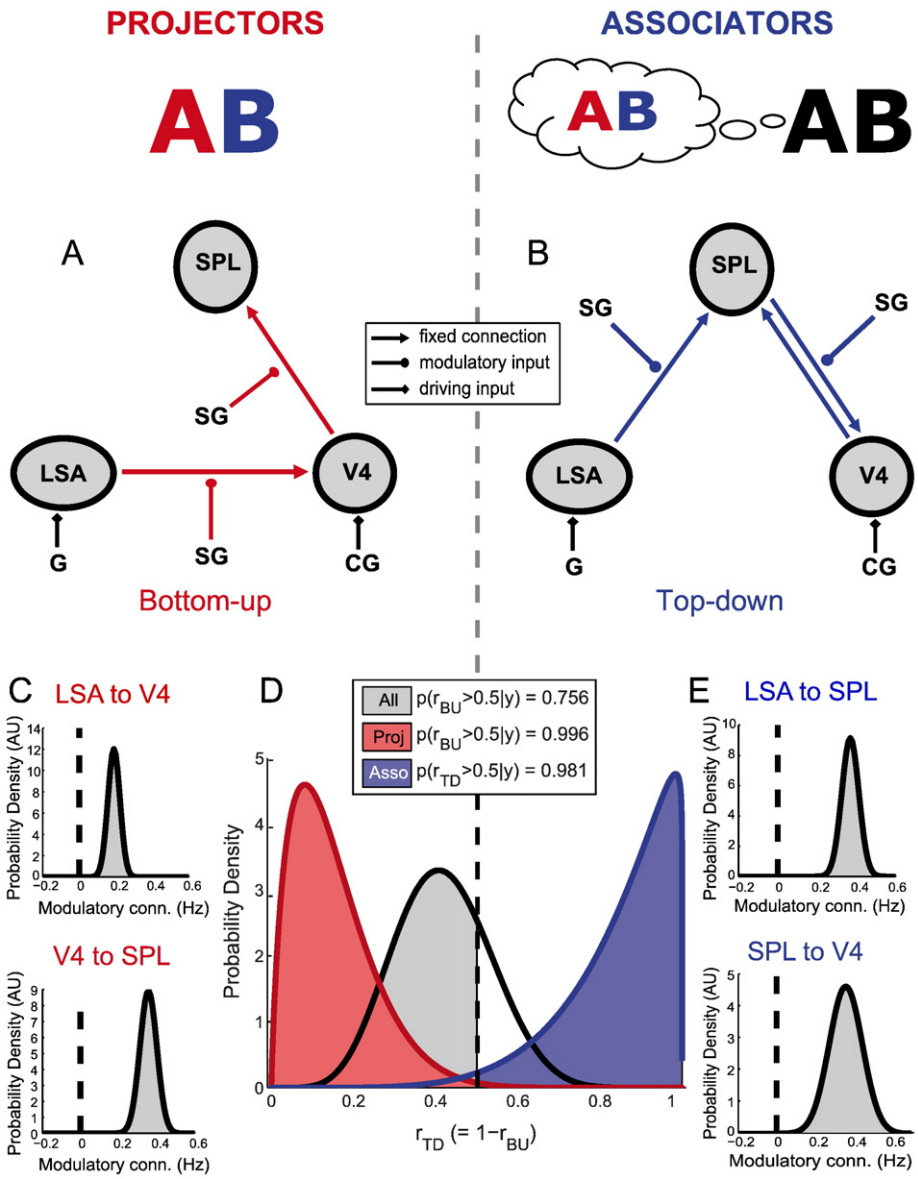

Fig. 7. Bayesian model selection for distinguishing between two subtypes of colour-grapheme synaesthesia. See main text for details. Reproduced, with permission, from (van Leeuwen et al., 2011).

In a subsequent model-based fMRI analysis, using the trajectories from the HMM, both patient groups demonstrated a failure to activate the ventral striatum during informative negative feedback, compared to healthy controls. Thus, the ventral striatal hypoactivation appeared to characterize schizophrenia patients independently of the task demands, suggesting that it was a core characteristic of the disease. However, a more careful interpretation was mandated. As there was no evidence that these subjects engaged in the same computations as controls, the absence of a BOLD signal related to this computation should not, by itself, be interpreted as a biological dysfunction, but more likely simply reflected the fact that an entirely different cognitive process took place. Hence, the fact that the hypoactivation was present in the subjects with good HMM fit was seen as strong evidence for a deficit specific to schizophrenia. The prefrontal cortex, by contrast, was differentially engaged between the patient groups, with a deficit present only in those subjects who did not show behavioural evidence of employing the complex HMM strategy.

This empirical example demonstrates that acknowledging individual cognitive differences can be crucial for interpreting neuroimaging results from heterogeneous patient groups. As illustrated above, between-group differences in brain activity may either be a consequence of underlying neurobiological differences when performing the same cognitive operation or due to differences in task solving strategies (Wilkinson and Halligan, 2004). This issue has often been addressed by matching groups on measures of average task performance. However, indistinguishable average task performance can be found under different individual strategies/models (Price and Friston, 2002). Statistical comparison of alternative computational models of trial-wise decisions is a superior alternative, since the trial-by-trial dynamics of observed responses contains additional information about the underlying computations which is lost by averaging.

It is worth mentioning that "resting state" fMRI studies are not unaffected by the problem of cognitive process heterogeneity. While the name implies an invariant and context-independent state of brain function, the "resting state" simply corresponds to unconstrained cognition in the absence of external sensory stimulation. While this has been recognised since early PET studies in the 1990s (Andreasen et al., 1995) and led to the labelling of "REST" as "random episodic silent thought" (Andreasen, 2011), it is only relatively recently that variability in cognitive processes during the resting state has gained scientific

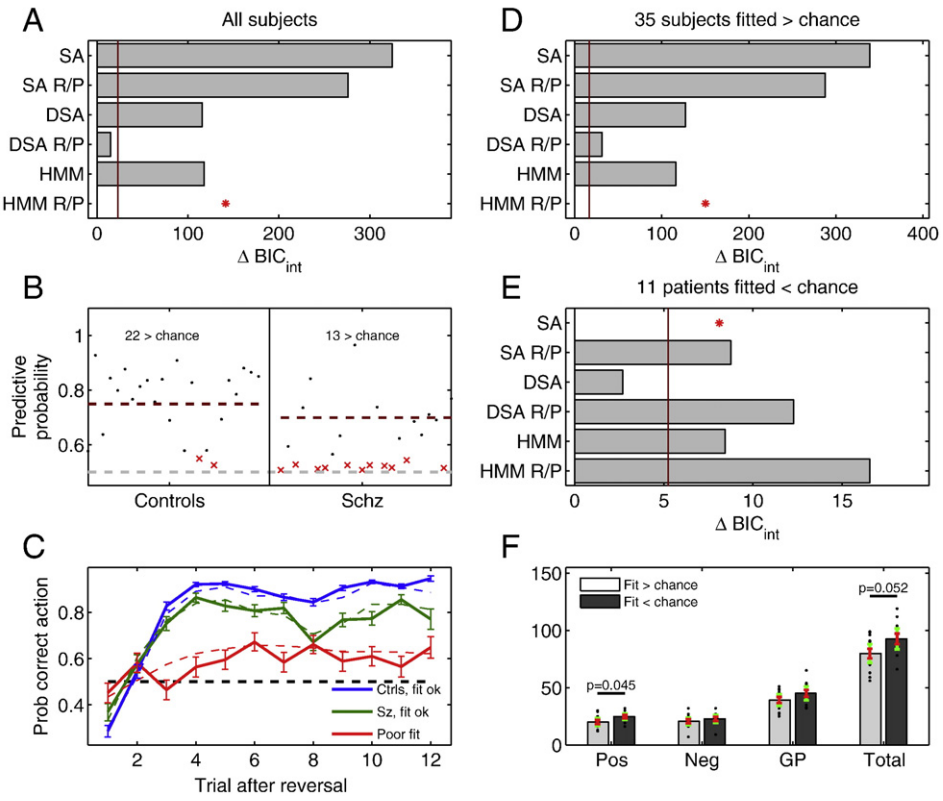

Fig. 8. The figure, reproduced from (Schlagenhauf et al., 2014) with permission, shows an example of the utility of model comparison in determining the most likely cognitive strategy of individuals. This example concerns a reversal learning study of unmedicated schizophrenia patients and healthy controls. A: Model comparison using the Bayesian Information Criterion (ΔBIC_{int}; compared to the best model). The best model has the lowest score ($\Delta BIC_{int} = 0$). SA: stimulus-action standard Rescorla-Wagner model, where the Q-value of only the chosen option was updated by a prediction error. DSA: Double-update model, where the Q values for both actions were updated on every trial. HMM: Hidden Markov Models, which assume that participants choose their action based on their belief about the underlying state of the task. R/P: reward/punishment version of the respective model, in which rewards and punishments had differential effects on learning. B: Model fit for each individual participant using the predictive probability of the HMM (black dots). Red crosses: participants whose data were not fitted better than chance. Red dashed lines: group means for participants with data fitted better than chance. This graph shows that behaviour from a substantial subgroup of schizophrenia patients was not explained above chance by the HMM and hence do not rely on the computations assumed by this model to solve the task. C: Average learning curves after reversals for participants with data fitted worse than chance (red), and for controls (blue) and patients with data fitted better than chance (green) by the HMM. D: Model comparison for participants whose behaviour was explained better than chance by the HMM. E: ΔBIC_{int} scores for patients with data fit poorly by the HMM (worse than chance). Asterisks indicate the best fitting model. F: Differences in clinical symptoms (Positive and Negative Syndrome Scale, PANSS) across patients whose behaviour was best explained by the HMM (as healthy controls), compared to patients with behaviour poorly explained by the HMM. Pos: positive symptom score; Neg: negative symptom score; GP: general psychopathology score; Total: PANSS total score.

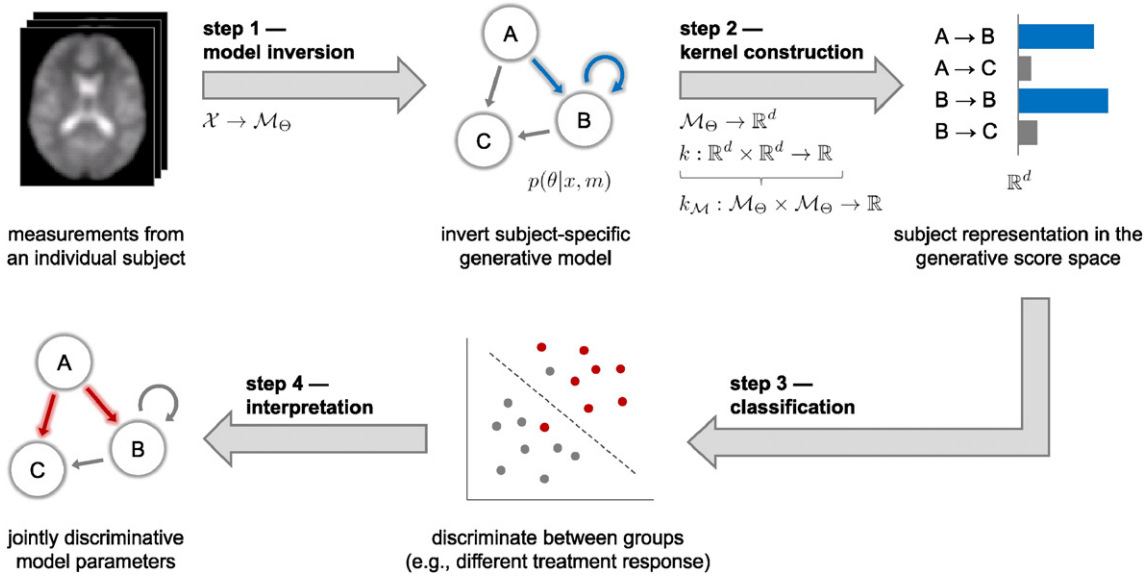

Fig. 9. Graphical summary of the idea behind generative embedding, illustrated for the supervised case (classification). Adapted, with permission, from (Brodersen et al., 2011).

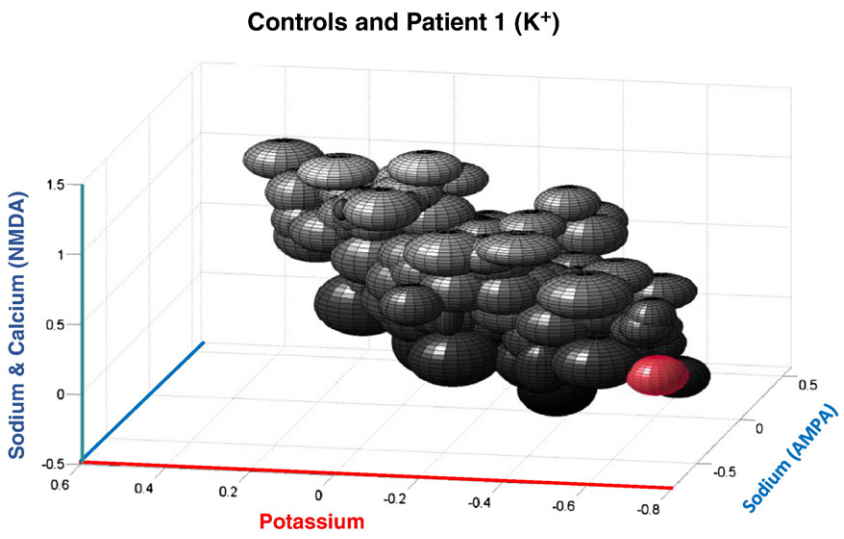

Fig. 10. Illustration of the diagnostic potential of model-based pathophysiological characterisation. This plot shows estimated conductances of two ionotropic receptors (AMPA and NMDA receptors) and of potassium channels in a patient (red ellipsoid) suffering from a known mutation of gene KCNJ2 which encodes a potassium inward-rectifying channel. These estimates are contrasted against those from 94 healthy controls, showing that the patient is located at the edge of the multivariate population distribution defined by the three conductance estimates. Ellipsoids represent 95% Bayesian confidence regions. Adapted, with permission, from (Gilbert et al., 2016).

traction, e.g., individual differences in mental imagery and mind wandering (Kucyi et al., 2013). In contrast to task-based paradigms, however, the absence of behavioural readouts makes it more difficult to establish differences in cognitive processes during "rest" and account for them in the interpretation of neurophysiological activity.

From model structure to parameter estimates: generative embedding and Bayesian model averaging

The prospect of using model selection for diagnostic or prognostic decisions about individual patients – as illustrated in Fig. 4 – represents an elegant strategy as it directly maps onto the decision-making process of the clinician who evaluates the differential diagnosis (plausibility of competing hypotheses). There are at least two scenaria, however, in which it is more straightforward to address diagnostic questions at the level of model parameters. First, as already mentioned above, in many psychiatric conditions, we do not yet have sufficiently precise pathophysiological theories that we can articulate competing models with sufficient precision. An alternative is to formulate a general model and search for subsets of informative parameters (e.g., finding diagnostically relevant connections within a fully connected network model). Second, even when a set of alternative models can be formulated, diagnostic inference may depend on the actual value of a parameter that is common to all models. In other words, instead of the binary absence or presence of a parameter (connection or computational quantity), it may be the degree to which it is expressed that is diagnostically informative. For example, as in the example by Boly et al. (2011), a particular connection may always be active during a particular cognitive operation, regardless of disease state; however, its strength may differ between healthy and pathological conditions.

In both cases, we need to shift the focus from model structure to parameter values. The first case motivates a generative embedding approach, where the posterior densities of model parameters define a feature space for subsequent (un)supervised learning; an approach we discuss in detail below. The second case suggests averaging parameter values across models. This can be achieved by Bayesian model averaging (BMA), which we turn to in the next section.

Bayesian model averaging (BMA)

When clinical questions are not straightforwardly addressed by model comparison alone but require examining the quantitative value of parameters encoding specific mechanisms, one could simply perform model selection first, identify the optimal model and then interrogate the posterior densities of the parameters of interest. While this is perfectly possible, a fully Bayesian perspective would account for uncertainty about the model itself. This is referred to as Bayesian model averaging (BMA) (Hoeting and Madigan, 1999; Penny et al., 2010). BMA computes an average posterior across all models considered, weighting the contribution from each model by its posterior probability (see Eq. 11):

$$p(\theta|y) = \sum_{j=1}^{|M|} p(\theta|y, m_j)\, p(m_j|y) \tag{16}$$

Here, M denotes the space of all models considered (cf. Eq. 8). Eq. 16 produces a single posterior density of the parameters which properly combines all sources of uncertainty – about the parameter values and about the models themselves.[8]

BMA can be applied to each subject separately (as in Eq. 16); alternatively, it can proceed on the basis of a random effects BMS procedure in order to exploit knowledge about model heterogeneity in the group as a whole (Penny et al., 2010). Notably, as long as the model space is identical, BMA makes it possible to test for differences in specific mechanisms (as represented by particular parameters), even when the optimal model differs across subjects or groups.

So far, BMA has mainly found applications in group-level studies of patients and controls in whom the optimal model differed (e.g., Dauvermann et al., 2013; Schmidt et al., 2013; Sladky et al., 2015). For example, Schmidt et al. (2013) applied DCM to fMRI data from 4 groups – healthy controls, at risk mental state (ARMS) subjects, medicated and non-medicated first episode psychosis (FEP) patients – performing an N-back working memory task. Importantly, the fMRI data from these groups were optimally explained by different fronto-parietal network models, prohibiting a straightforward comparison of

[8] An illustrative application to empirical data can be found in Figures 4–8 of Penny et al. (2010).

effective connectivity estimates across groups and requiring BMA to average parameter estimates across models. The comparison of averaged posterior estimates across groups indicated that right fronto-parietal coupling in controls was significantly higher than in non-medicated FEP patients; with ARMS individuals taking intermediate values. Although the cross-sectional nature of this study does not allow for causal conclusions, it is interesting to note that coupling strength in medicated FEP patients did not differ significantly from healthy controls. The implication that antipsychotic medication may be restoring frontal-parietal coupling would need to be tested in future studies with a prospective design.

An approach not unrelated to BMA has been pursued by recent studies examining the interplay between distinct decision-making mechanisms that operate with or without reference to an explicit model of the environment or task structure (Daw et al., 2011). Here, both mechanisms are represented as components of a larger model, where trial-wise decisions are modelled as a linear mixture of predictions from both sub-models.

Generative embedding

As mentioned above, in many psychiatric and some neurological diseases, we lack precise ideas how prominent symptoms and signs are generated from underlying (hidden) mechanisms. This void of concrete pathophysiological hypotheses can render the formulation of concrete models encoding alternative disease mechanisms difficult and thus aggravate differential diagnosis based on model selection or model averaging. In this case, generative embedding constitutes a powerful alternative.

Generative embedding was introduced to neuroimaging by Brodersen et al. (2011) and rests on a simple but powerful idea: the embedding of (un)supervised learning, such as classification or clustering, into the parameter space of a generative model (for a summary, see Fig. 9). This effectively uses a generative model as a theory-led feature selection device which creates a low-dimensional and mechanistically interpretable set of features to which machine learning techniques can be applied. This addresses the two key challenges for applications of conventional machine learning approaches to neuroimaging data, which we briefly alluded to in the Introduction. The first challenge is a mismatch between the small number of subjects and the very high dimensionality of the data. For example, an fMRI dataset typically offers on the order of 10^8 data features (several 100,000 voxels, each with a time series of 10^3 signal samples). This renders feature selection a key problem: there is a huge number of alternatives how one could, for example, construct a classifier based on specific aspects of the measured data in order to predict an independent variable (e.g., clinical outcome) in individual subjects. An exhaustive search for the most informative features becomes prohibitively expensive in this scenario. As a consequence, machine learning analyses of neuroimaging data often resort to *ad hoc* feature selection, such as using timeseries from anatomically predefined regions of interest. Alternatively, they frequently adopt a piecemeal strategy by analysing a subset of the data at a time, such as widely used "searchlight" procedures (Kriegeskorte et al., 2006).

The second challenge is that machine learning approaches often operate directly on features of the observed data. This has two disadvantages. One problem is that "raw" data do not only reflect (neuronal) processes of interest, but may contain confounding (e.g., vascular) influences that can vary across the brain and individuals. Similarly, machine learning techniques essentially represent "black box" procedures that establish a purely statistical relationship between a set of predictors (features of measured data) and a target variable of interest (e.g., diagnostic label or clinical outcome). By contrast, they do not clarify which particular neurobiological processes underlying the observed data could be driving the predictive relationship. This may limit the long-term utility of a purely ML-based strategy. For example, a successful prediction of clinical outcome cannot be understood in terms of the biological processes that determine this outcome and would therefore represent potential targets for treatment development.

Generative embedding proposes that, instead of extracting aspects of the measured data as features for (un)supervised learning, it is the posterior parameter densities obtained by inversion of a computational model that should inform classification or clustering. This simple idea addresses the two challenges outlined above. First, a generative model represents a low-dimensional description of the mechanisms by which measured data are generated; these mechanisms are enshrined into the structural form of the likelihood function and represented by the values of its parameters. By inverting the generative model, the data are decomposed into a predicted component, which can be summarised by a few numbers (parameter estimates) plus noise (cf. Eq. 1). A model can thus be seen as a theory-led dimensionality reduction device that projects high-dimensional and noisy data onto a subspace of much lower dimensionality, maintaining only those aspects of the data which are deemed informative (from the perspective of the model).

Second, as its dimensions are provided by the model parameters, this subspace has a mechanistic interpretation. That is, the position of each of the data points (subjects) in parameter space can be interpreted with regard to the neurophysiological or computational mechanisms that are specified by the model. This property is visualised by Fig. 10, using the results by Gilbert et al. (2016) as an example.

At this point, one might argue that model-based feature selection has been used in machine learning for a long time; that is, the use of voxel-wise regression weights, from a conventional GLM, as inputs for multivariate classification (Pereira et al., 2009). This is not entirely incorrect; however, the GLM-based approach does not address the above problems in the same way as a generative embedding strategy. For example, it can only account for hemodynamic confounds to a limited degree; it does not rest on a generative model and hence only provides point estimates of parameters, not their uncertainty, and, most importantly, it does not convey a truly mechanistic understanding, in terms of a biological process, but only provides a spatial mosaic of point-wise correlates of an experimental condition.

The advantages of a model-based approach to (un)supervised learning and single-subject predictions have been highlighted by several recent papers (Doyle et al., 2013; Huys et al., 2016; Wiecki et al., 2015)). In the following, we summarise a few examples of how it has found application in recent neuroimaging studies of patients.

The initial article on generative embedding by Brodersen et al. (2011) used data from stroke patients with aphasia resulting from a lesion in left inferior frontal and/or temporal cortex. This proof of concept study focused on modelling activity in parts of the auditory cortex that were unaffected by the lesion during passive listening to speech, asking whether the model parameter estimates would predict the absence or presence of the "hidden" lesion (i.e., out of the field of view). The rationale behind this approach is that auditory cortex activity is in receipt of backward projections from more anterior temporal and prefrontal regions, and lesions of the latter induce a functional reorganisation of lower auditory areas (Schofield et al., 2012). Brodersen et al. (2011) constructed a six-region DCM of the auditory system (comprising the auditory thalamus, Heschl's gyrus, and planum temporale), which allowed for reducing the high-dimensional fMRI data to 20 connectivity parameters. The posterior mean parameter estimates were then used as features for a discriminant classifier, a support vector machine (SVM), in order to predict, subject by subject, the presence or absence of a lesion in IFG. Using this approach, patients could be differentiated from matched controls with much higher accuracy (98%, leave-one-out cross-validation) than with conventional classification approaches operating either on regional activity measures or estimates of functional connectivity (e.g., correlation or PCA). The latter achieved cross-validated accuracy levels that ranged from chance performance to approximately 80% accuracy and thus stayed significantly below the performance of the DCM-based classification. Notably, the deliberate

removal of neurobiologically plausible connections also drastically diminished the predictive power of the model-based classification, demonstrating that informed theories are crucial for the success of a model-based approach.

Equally, if not more importantly than the increase in performance, this study provided an example of how generative embedding can convey an understanding of the biological mechanisms which allow for differentiation or prediction. Specifically, Brodersen et al. (2011) found that connection strengths from the right to the left auditory areas during passive speech perception were particularly informative for predicting absence or presence of a lesion in left IFG. This highlighted that the removal of top-down influences due to the remote lesion induced plasticity in early auditory areas leading to characteristic (and possibly compensatory) alterations of interhemispheric transfer of speech inputs, from the non-dominant right hemisphere to the language-dominant left hemisphere.

Two further examples demonstrate that the supervised application of generative embedding can markedly enhance classification accuracy. Brodersen et al. (2014) used a three-region DCM of fMRI data from a working memory task (Deserno et al., 2012) for classification (SVM) of patients with schizophrenia and healthy controls. They found that connectivity estimates provided by DCM enabled a significantly higher accuracy (78%) than classification based on either functional connectivity (62%) or regional activity levels (55%). Wiecki et al. (2015) applied a drift-diffusion model (DDM) to behavioural data from patients with Parkinson's disease with a deep brain stimulator targeting the subthalamic nucleus. They were able to predict the state of the stimulator (on/off) from the model parameter estimates with significantly higher accuracy (81%) than from the data alone (67%).

Applications of generative embedding to clinical questions

The studies discussed so far do not address real clinical problems but only served to evaluate the potential utility of a model-based classification approach. By contrast, recent work by Gilbert et al. (2016) illustrates how model-based classification might contribute diagnostic relevant information. This study concerned genetically determined alterations of ion channel conductance (channelopathies) that play a possible role in numerous neuropsychiatric disorders, ranging from epilepsy to schizophrenia (Klassen et al., 2011). Gilbert et al. (2016) used a DCM representing a cortical microcircuit consisting of 3 different types of neurons (pyramidal cells, excitatory and inhibitory interneurons) equipped with 3 different ionotropic receptors (AMPA, NMDA and $GABA_A$) and associated sodium, potassium and calcium channels. Applying this generative model to MEG data from a large group of controls ($N = 94$), the authors established a reference distribution against which patients could be compared. Specifically, they examined two patients with known monogenic channelopathies concerning specific potassium and calcium channels, respectively, and showed that the ensuing parameter estimates of the respective ion channel conductances placed the patients at the edges of the multivariate population distribution (compare Fig. 10). The implication that identification of functionally relevant channelopathies might be feasible from non-invasively recorded M/EEG data, is of clinical relevance because the genetic characterisation of channelopathies is not sufficient to predict their neurophysiological consequences. For example, individual clinical risk depends on the genetic and functional status of other ion channels; this is reflected by the conclusion of a recent genetic study that "in silico modelling of channel variation in realistic cell and network models will be crucial to future strategies assessing mutation profile pathogenicity and drug response in individuals with a broad spectrum of excitability disorders" (Klassen et al., 2011).

Moving from neurology to psychiatry, the clinically perhaps most relevant application of a supervised generative embedding approach is to establish model-based predictors about individual treatment responses and clinical trajectories. This requires prospective studies where initial neuroimaging measurements are combined with clinical follow-up assessments. While these studies are rare due to their laborious and time-consuming nature, a recent paper by Harle et al. (2015) provided a compelling demonstration for the power of model-based predictions. In brief, these authors acquired fMRI data from occasional stimulant users during a stop-signal task and demonstrated that a Bayesian model of this task was able to predict clinically relevant abuse and dependence symptoms 3 years later. Importantly, this model-based prediction, based on prediction error induced activity in several brain regions including prefrontal, insular and cingulate cortex, significantly outperformed predictions based on clinical variables and conventional fMRI analyses.

A second central challenge concerns the heterogeneous nature of spectrum disorders defined by phenomenological classification schemes. A recent study used an unsupervised generative embedding approach to demonstrate how spectrum diseases could be dissected into mechanistically separate subgroups (Brodersen et al., 2014). The study used a simple three-region DCM for inferring connectivity between visual, parietal, and prefrontal regions from fMRI data of an N-back working memory task (Deserno et al., 2012). The ensuing posterior estimates of effective connection strengths entered a variational Gaussian mixture model, which served to identify the most likely partitioning of schizophrenic patients into connectivity-subgroups. The results pointed to three distinct patient subgroups which were distinguished by different changes in visual-parietal-prefrontal coupling under working memory load. Importantly, this clustering was validated by relating it to independent clinical symptoms (which the model did not have access to), showing that the three physiologically defined subgroups differed significantly with respect to negative symptoms.

At this point, one might question whether the attempt to dissect spectrum disorders into discrete sub-entities is the most promising approach, or whether a dimensional perspective would not be more appropriate than a categorical disease concept. It is worth pointing out that, under a generative modelling approach, categorical and dimensional perspectives coexist and can be reconciled naturally (see also the discussion in Stephan et al., 2016). That is, in computational models, most parameters are of a continuous nature, and the disease-relevant mechanisms they encode would naturally underpin a dimensional description. On the other hand, variations of certain model parameters can induce abrupt qualitative shifts in system behaviour (i.e., bifurcations); this in turn, speaks to the plausibility of categorical classifications.

It is, of course, straightforward to extend unsupervised generative embedding approaches to computational models of behaviour. This might identify patient subgroups in spectrum diseases that are characterized by different information processing styles or task solving strategies (Huys et al., 2016; Wiecki et al., 2015). A recent empirical demonstration was provided by (Zhang et al., 2016) who showed that parameter estimates from a generative model (hierarchical drift diffusion model) applied to saccadic responses could differentiate between patients with movement disorders (progressive supranuclear palsy and Parkinson's disease) with significantly higher accuracy than using the "raw" behavioural data.

Regardless of whether a physiological and/or computational perspective is adopted, however, a key challenge for the future will be to validate any putative subgroups or procedures for differential diagnosis in prospective studies that use clinically relevant outcomes such as treatment response as real-world benchmarks (Stephan et al., 2015). Establishing the clinical utility of single-subject computational assays is by no means a trivial endeavour. Due to the necessity of clinical follow-up, these studies typically take a long time and are resource-demanding. One might also be concerned that it could take a long time until computational assays for psychiatry reach a practically acceptable level of sensitivity and specificity for single-patient decisions. Here, it is worth noting that there is no universally accepted threshold, and routine tests in other medical disciplines vary greatly in terms of sensitivity and specificity, depending on the urgency of the problem,

availability of alternatives, and benefit-cost considerations. For example, fecal occult blood test screening for colorectal cancer has a sensitivity of only 60–80% at best (Burch et al., 2007), compared to a sensitivity of 99.7% for HIV tests (Chou et al., 2005). In psychiatry, given the almost complete lack of predictive clinical tests, even moderately accurate computational assay could be extremely useful, provided they address key clinical problems – such as predicting individual treatment response in schizophrenia or depression – and provided the necessary data can be acquired in a cost-efficient manner.

Caveats and application domains

As explained in the introduction, this paper focuses on two of the three most commonly used computational approaches to neuroimaging data: generative models and model-based fMRI. Compared to more complex biophysical network models, these approaches utilise sufficiently simplified formulations that model inversion and parameter estimation becomes feasible. Nevertheless, certain caveats exist which have been reviewed in previous papers (e.g., Daunizeau et al., 2011; Stephan et al., 2010, 2015) and which we briefly summarise here.

Generative models attempt to solve the "inverse problem" of identifying hidden variables from measurements. The feasibility of this endeavour can be jeopardised by various challenges, such as identifiability problems or overfitting. Overfitting can be avoided by the regularisation afforded by priors (Bishop, 2006) and can be detected by model comparison, as discussed above. Additionally, in biological systems, parameters often show strong interdependencies (Gutenkunst et al., 2007), leading to potential identifiability problems (i.e., several equally good solutions exist). In generative models, this problem can be addressed in at least two ways. First, many inversion schemes, such as the VB scheme of DCM, provide an estimate of posterior parameter covariances; in turn, these are related to classical sensitivity indices (Deneux and Faugeras, 2006) and can be used as a diagnostic for identifying potentially problematic cases that may need follow-up investigations with simulations. Second, some implementations of BMS penalise models with identifiability problems automatically. For example, this is readily visible in the case of the negative free energy approximation to the log evidence under Gaussian assumptions; compare Eq. 15 (second term).

In addition to relative statements about model quality, as achieved by straightforward application of BMS described above, one would sometimes also like to specify model goodness in absolute terms. The latter can be achieved in two distinct ways. One possibility is to compute the posterior predictive density; this specifies how well future data can be predicted given the parameter estimates obtained from currently available data (for an example, see Huys et al. (2011). An alternative is to examine how well the model allows for solving a concrete problem, e.g. predicting a clinical outcome or treatment response (Fig. 1). This can either be addressed with generative embedding (using a cross-validation scheme with held out data) or by applying BMS to hierarchical generative models whose parameters not only generate subject-wise neuroimaging or behavioural measurements but also clinical variables (see Friston et al., 2016).

Moving on to application domains, the physician or biomedical scientist with little background in computational modelling may wonder which model is most appropriate for which application. This paper is not an ideal place for attempting an answer – not least because it focuses on general principles of generative modelling without explaining specific models in depth – and we point the reader to previous overviews of relevance for this issue (Gershman, 2015; Huys et al., 2016; Kahan and Foltynie, 2013; Stephan et al., 2015; Stephan et al., 2010). Generally the choice of an adequate model for a particular problem depends on at least three factors: the specific clinical question of interest, the amount of prior knowledge available, and practical constraints on data acquisition and analysis (e.g., time and equipment available, patient compliance, benefit/cost ratio); see the discussion in (Stephan et al., 2015). In all brevity, a generative modelling approach requires a priori hypotheses about disease mechanisms which can be represented by an existing model class; if this is not the case, more exploratory and descriptive approaches are preferable, such as the direct application of machine learning methods to behavioural and/or neuroimaging data (Klöppel et al., 2012; Orru et al., 2012; Wolfers et al., 2015). Furthermore, the stronger the practical constraints regarding time, costs and necessary equipment, the more attractive models of behaviour become, given the relative ease and cheapness of data acquisition, including the possibility of acquiring data online (Gillan et al. 2016; Moran et al., 2008) or via mobile phones (Rutledge et al., 2014). On the other hand, generative models of fMRI data are indispensable for characterising the connectivity of networks involving deep sources (e.g., brain stem) while estimates of processes at synaptic and ion channel levels critically require generative models of M/EEG data (for examples, see Cooray et al., 2015; Gilbert et al., 2016; Moran et al. 2008, 2011).

Hierarchical models

Computational models are opening up tantalising possibilities to infer, from non-invasively measured brain activity and behaviour, on core mechanisms of disease. In this paper, we have provided an overview of generative modelling and considered two main approaches – model selection and generative embedding – which can link the inferential power of computational models to clinical predictions in individual patients. We have tried to present the key ideas behind these concepts without going too deeply into mathematical details, in the hope that this will help biomedically and clinically trained colleagues to access the rapidly growing literature on computational modelling in psychiatry and neurology.

This paper is concerned with individual subject predictions and has therefore focused on approaches that rest on inverting the generative model(s) for each subject separately. In other words, these approaches separate the problem of inference (on model structure and parameters) from the problem of prediction (of a clinically relevant variable) or population structure learning (subgroup detection). While this two-step procedure is presently the most widely used strategy and is also more easily accessible from a didactic point of view, complementary and more sophisticated approaches are appearing on the horizon. This involves hierarchical models which allow for single-subject inference while exploiting information about the population as a whole. Generally, this hierarchical strategy comes under the rubric of "empirical Bayes" where the estimation of individual subject's parameters proceeds under priors that are informed by the variability across subjects (Friston et al., 2002; Huys et al., 2011). However, a critical conceptual advance concerns the development of hierarchical generative models which, in addition to subject-wise neuroimaging or behavioural data, include the clinical variable of interest (e.g., distribution of treatment outcomes in the population or structure of subgroups) as an explanandum.

First examples of such unified hierarchical models are presently appearing in the literature. For example, Friston et al. (2016) have proposed a hierarchical generative model of fMRI group data which uses a nonlinear first-level model (DCM) to explain individual subjects' fMRI data, and a linear second-level model to explain the distribution of connectivity parameter estimates at the group level. The latter can be informed by any clinical variable of interest, such as diagnostic labels, treatment responses, or clinical scores, and can thus be used for model-based predictions in terms of either classification or regression. Notably, this model is inverted using extremely efficient model reduction techniques based on VB.

An alternative approach by (Raman et al., 2016) unifies DCM (of individual subject's fMRI data) with mixture models (of population structure) within a single hierarchical generative model which is inverted using MCMC techniques. This approach allows for simultaneously inferring connectivity in individual subjects and for detecting subgroups defined by model parameters. The inversion of subject-specific DCMs is

governed by subgroup-specific prior distributions that are determined in an empirical Bayesian fashion. These new developments open up exciting possibilities for exploiting generative models for clinical diagnosis and prognosis.

Summary and outlook

This paper has provided an overview of the emerging use of computational models for clinically relevant single-subject predictions. Our particular focus has been on generative models which enable inference on (patho)physiological and (patho)computational mechanisms from individual behavioural and neuroimaging measurements. These models may prove useful for supporting clinical decision-making on their own (e.g., differential diagnosis through Bayesian model selection) or in conjunction with machine learning techniques that use parameter estimates as features (generative embedding). This combination of generative modelling and machine learning has great potential for tackling key clinical problems in psychiatry and neurology that arise from the heterogeneity of current disease constructs, such as outcome prediction and individual treatment allocation. The success of this endeavour will depend on carefully designed prospective validation studies and close collaborations between clinically and computationally trained scientists.

We hope that this paper makes a useful contribution to this necessary interdisciplinary exchange and provides inspiration for the development and deployment of computational neuroimaging approaches to future diagnostic applications.

Acknowledgements

We would like to thank our colleagues at TNU and FIL for helpful comments and feedback. We acknowledge support by the René and Susanne Braginsky Foundation (KES), the University of Zurich (KES), the UZH Clinical Research Priority Programs (CRPP) “Molecular Imaging” (KES) and “Multiple Sclerosis” (KES, SR), the Deutsche Forschungsgemeinschaft (TR-SFB 134) (KES), the Swiss National Science Foundation (320030L_153449/1) (QJMH), and the Wellcome Trust (KJF, RJD).

References

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The computational anatomy of psychosis. Front. Psychiatry 4, 47.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.
Andreasen, N.C., 2011. A journey into chaos: creativity and the unconscious. Mens Sana Monogr. 9, 42–53.
Andreasen, N.C., O'Leary, D.S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G.L., Ponto, L.L., Hichwa, R.D., 1995. Remembering the past: two facets of episodic memory explored with positron emission tomography. Am. J. Psychiatry 152, 1576–1585.
Anticevic, A., Hu, X., Xiao, Y., Hu, J., Li, F., Bi, F., Cole, M.W., Savic, A., Yang, G.J., Repovs, G., Murray, J.D., Wang, X.J., Huang, X., Lui, S., Krystal, J.H., Gong, Q., 2015. Early-course unmedicated schizophrenia patients exhibit elevated prefrontal connectivity associated with longitudinal change. J. Neurosci. 35, 267–286.
Aponte, E.A., Raman, S., Sengupta, B., Penny, W.D., Stephan, K.E., Heinzle, J., 2016. Mpdcm: a toolbox for massively parallel dynamic causal modeling. J. Neurosci. Methods 257, 7–16.
Arbabshirani, M.R., Kiehl, K.A., Pearlson, G.D., Calhoun, V.D., 2013. Classification of schizophrenia patients based on resting-state functional network connectivity. Front. Neurosci. 7, 133.
Beal, M., Ghahramani, Z., 2003. The Variational Bayesian EM algorithms for incomplete data: with application to scoring graphical model structures. In: Bernardo, J., Bayarri, M., Berger, J., Dawid, A. (Eds.), Bayesian Statistics. Cambridge University Press, Cambridge.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Berlin.
Boly, M., Garrido, M.I., Gosseries, O., Bruno, M.A., Boveroux, P., Schnakers, C., Massimini, M., Litvak, V., Laureys, S., Friston, K., 2011. Preserved feedforward but impaired top-down processes in the vegetative state. Science 332, 858–862.
Breakspear, M., Roberts, G., Green, M.J., Nguyen, V.T., Frankland, A., Levy, F., Lenroot, R., Mitchell, P.B., 2015. Network dysfunction of emotional and cognitive processes in those at genetic risk of bipolar disorder. Brain 138, 3427–3439.
Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7, e1002079.
Brodersen, K.H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W.D., Buhmann, J.M., Stephan, K.E., 2014. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage Clin. 4, 98–111.
Brown, M.R., Sidhu, G.S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P.H., Greenshaw, A.J., Dursun, S.M., 2012. ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Front. Syst. Neurosci. 6, 69.
Burch, J.A., Soares-Weiser, K., St John, D.J., Duffy, S., Smith, S., Kleijnen, J., Westwood, M., 2007. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. J. Med. Screen. 14, 132–137.
Calderhead, B., Girolami, M., 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data Anal. 53, 4028–4045.
Casey, B.J., Craddock, N., Cuthbert, B.N., Hyman, S.E., Lee, F.S., Ressler, K.J., 2013. DSM-5 and RDoC: progress in psychiatry research? Nat. Rev. Neurosci. 14, 810–814.
Chen, C.C., Kiebel, S.J., Friston, K.J., 2008. Dynamic causal modelling of induced responses. NeuroImage 41, 1293–1312.
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.
Chou, R., Huffman, L.H., Fu, R., Smits, A.K., Korthuis, P.T., Force, U.S.P.S.T., 2005. Screening for HIV: a review of the evidence for the U.S. Preventive Services Task Force. Ann. Intern. Med. 143, 55–73.
Cooray, G.K., Sengupta, B., Douglas, P., Englund, M., Wickstrom, R., Friston, K., 2015. Characterising seizures in anti-NMDA-receptor encephalitis with dynamic causal modelling. NeuroImage 118, 508–519.
Craddock, R.C., Holtzheimer 3rd, P.E., Hu, X.P., Mayberg, H.S., 2009. Disease state prediction from resting state functional connectivity. Magn. Reson. Med. 62, 1619–1628.
Cuthbert, B.N., Insel, T.R., 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 11, 126.
D'Ardenne, K., Lohrenz, T., Bartley, K.A., Montague, P.R., 2013. Computational heterogeneity in the human mesencephalic dopamine system. Cogn. Affect Behav. Neurosci. 13, 747–756.
Daunizeau, J., Friston, K.J., Kiebel, S.J., 2009. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Phys. D 238, 2089–2118.
Daunizeau, J., David, O., Stephan, K.E., 2011. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. NeuroImage 58, 312–322.
Daunizeau, J., Stephan, K.E., Friston, K.J., 2012. Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? NeuroImage 62, 464–481.
Dauvermann, M.R., Whalley, H.C., Romaniuk, L., Valton, V., Owens, D.G., Johnstone, E.C., Lawrie, S.M., Moorhead, T.W., 2013. The application of nonlinear dynamic causal modelling for fMRI in subjects at high genetic risk of schizophrenia. NeuroImage 73, 16–29.
David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., Friston, K.J., 2006. Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage 30, 1255–1272.
David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. 6, 2683–2697.
Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876–879.
Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.
DeBattista, C., Kinrys, G., Hoffman, D., Goldstein, C., Zajecka, J., Kocsis, J., Teicher, M., Potkin, S., Preda, A., Multani, G., Brandt, L., Schiller, M., Iosifescu, D., Fava, M., 2011. The use of referenced-EEG (rEEG) in assisting medication selection for the treatment of depression. J. Psychiatr. Res. 45, 64–75.
Deco, G., Kringelbach, M.L., 2014. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. Neuron 84, 892–905.
Deco, G., Jirsa, V.K., McIntosh, A.R., 2013a. Resting brains never rest: computational insights into potential cognitive architectures. Trends Neurosci. 36, 268–274.
Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G.L., Hagmann, P., Corbetta, M., 2013b. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. J. Neurosci. 33, 11239–11252.
den Ouden, H.E., Friston, K.J., Daw, N.D., McIntosh, A.R., Stephan, K.E., 2009. A dual role for prediction error in associative learning. Cereb. Cortex 19, 1175–1185.
den Ouden, H.E., Daunizeau, J., Roiser, J., Friston, K.J., Stephan, K.E., 2010. Striatal prediction error modulates cortical coupling. J. Neurosci. 30, 3210–3219.
Deneux, T., Faugeras, O., 2006. Using nonlinear models in fMRI data analysis: model selection and activation detection. NeuroImage 32, 1669–1689.
Deserno, L., Sterzer, P., Wustenberg, T., Heinz, A., Schlagenhauf, F., 2012. Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. J. Neurosci. 32, 12–20.
Deserno, L., Huys, Q.J., Boehme, R., Buchert, R., Heinze, H.J., Grace, A.A., Dolan, R.J., Heinz, A., Schlagenhauf, F., 2015. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc. Natl. Acad. Sci. U. S. A. 112, 1595–1600.
Dima, D., Roiser, J.P., Dietrich, D.E., Bonnemann, C., Lanfermann, H., Emrich, H.M., Dillo, W., 2009. Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. NeuroImage 46, 1180–1186.
Dima, D., Dietrich, D.E., Dillo, W., Emrich, H.M., 2010. Impaired top-down processes in schizophrenia: a DCM study of ERPs. NeuroImage 52, 824–832.
Dombrovski, A.Y., Szanto, K., Clark, L., Reynolds, C.F., Siegle, G.J., 2013. Reward Signals, Attempted Suicide, and Impulsivity in Late-Life Depression (JAMA Psychiatry).
Doyle, O.M., Tsaneva-Atansaova, K., Harte, J., Tiffin, P.A., Tino, P., Diaz-Zuccarini, V., 2013. Bridging paradigms: hybrid mechanistic-discriminative predictive models. IEEE Trans. Biomed. Eng. 60, 735–742.

Du, Y., Pearlson, G.D., Liu, J., Sui, J., Yu, Q., He, H., Castro, E., Calhoun, V.D., 2015. A group ICA based framework for evaluating resting fMRI markers when disease categories are unclear: application to schizophrenia, bipolar, and schizoaffective disorders. NeuroImage 122, 272–280.
Duff, E.P., Vennart, W., Wise, R.G., Howard, M.A., Harris, R.E., Lee, M., Wartolowska, K., Wanigasekera, V., Wilson, F.J., Whitlock, M., Tracey, I., Woolrich, M.W., Smith, S.M., 2015. Learning to identify CNS drug action and efficacy using multistudy fMRI data. Sci. Transl. Med. 7 (274ra216).
Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. Neuroimage. 19 (3), 1240–1249.
Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466–477.
Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16, 484–512.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34, 220–234.
Friston, K.J., Stephan, K.E., Montague, R., Dolan, R.J., 2014. Computational psychiatry: the brain as a phantastic organ. The Lancet Psychiatry 1, 148–158.
Friston, K.J., Litvak, V., Oswal, A., Razi, A., Stephan, K.E., van Wijk, B.C., Ziegler, G., Zeidman, P., 2016. Bayesian model reduction and empirical Bayes for group (DCM) studies. NeuroImage 128, 413–431.
Garrido, M.I., Friston, K.J., Kiebel, S.J., Stephan, K.E., Baldeweg, T., Kilner, J.M., 2008. The functional anatomy of the MMN: a DCM study of the roving paradigm. NeuroImage 42, 936–944.
Gershman, S.J., 2015. A unifying probabilistic view of associative learning. PLoS Comput. Biol. 11, e1004567.
Gilbert, J.R., Symmonds, M., Hanna, M.G., Dolan, R.J., Friston, K.J., Moran, R.J., 2016. Profiling neuronal ion channelopathies with non-invasive brain imaging and dynamic causal models: case studies of single gene mutations. NeuroImage 124, 43–53.
Gillan, C.M., Kosinski, M., Whelan, R., Phelps, E.A., Daw, N.D., 2016. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. http://dx.doi.org/10.7554/eLife.11305.
Glascher, J.P., O'Doherty, J.P., 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscip. Rev. Cogn. Sci. 1, 501–510.
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595.
Gradin, V.B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., Steele, J.D., 2011. Expected value and prediction error abnormalities in depression and schizophrenia. Brain 134, 1751–1764.
Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P., 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. 3, 1871–1878.
Hadley, J.A., Nenert, R., Kraguljac, N.V., Bolding, M.S., White, D.M., Skidmore, F.M., Visscher, K.M., Lahti, A.C., 2014. Ventral tegmental area/midbrain functional connectivity and response to antipsychotic medication in schizophrenia. Neuropsychopharmacology 39, 1020–1030.
Harle, K.M., Stewart, J.L., Zhang, S., Tapert, S.F., Yu, A.J., Paulus, M.P., 2015. Bayesian neural adjustment of inhibitory control predicts emergence of problem stimulant use. Brain 138, 3413–3426.
Heinz, A., 2002. Dopaminergic dysfunction in alcoholism and schizophrenia–psychopathological and behavioral correlates. Eur. Psychiatry 17, 9–16.
Hjelm, R.D., Calhoun, V.D., Salakhutdinov, R., Allen, E.A., Adali, T., Plis, S.M., 2014. Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. NeuroImage 96, 245–260.
Hoeting, J.A., Madigan, D.E.A., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–401.
Hubbard, E.M., Arman, A.C., Ramachandran, V.S., Boynton, G.M., 2005. Individual differences among grapheme-color synesthetes: brain-behavior correlations. Neuron 45, 975–985.
Huys, Q.J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R.J., Dayan, P., 2011. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput. Biol. 7, e1002028.
Huys, Q.J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., Roiser, J.P., 2012. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput. Biol. 8, e1002410.
Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge between neuroscience and clinical applications. Nat. Neurosci. (in press).
Iglesias, S., Mathys, C., Brodersen, K.H., Kasper, L., Piccirelli, M., den Ouden, H.E., Stephan, K.E., 2013. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron 80, 519–530.
Ingalhalikar, M., Parker, W.A., Bloy, L., Roberts, T.P., Verma, R., 2014. Creating multimodal predictors using missing data: classifying and subtyping autism spectrum disorder. J Neurosci Methods 235, 1–9.
Kahan, J., Foltynie, T., 2013. Understanding DCM: ten simple rules for the clinician. NeuroImage 83, 542–549.
Kapur, S., 2003. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. Am. J. Psychiatry 160, 13–23.
Kapur, S., Phillips, A.G., Insel, T.R., 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol. Psychiatry 17, 1174–1179.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
King, R., Barchas, J.D., Huberman, B.A., 1984. Chaotic behavior in dopamine neurodynamics. Proc. Natl. Acad. Sci. U. S. A. 81, 1244–1247.
Klassen, T., Davis, C., Goldman, A., Burgess, D., Chen, T., Wheeler, D., McPherson, J., Bourquin, T., Lewis, L., Villasana, D., Morgan, M., Muzny, D., Gibbs, R., Noebels, J., 2011. Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell 145, 1036–1048.
Klein-Flügge, M.C., Hunt, L.T., Bach, D.R., Dolan, R.J., Behrens, T.E., 2011. Dissociable reward and timing signals in human midbrain and ventral striatum. Neuron 72, 654–664.
Klöppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourao-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. NeuroImage 61, 457–463.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.
Krystal, J.H., State, M.W., 2014. Psychiatric disorders: diagnosis to therapy. Cell 157, 201–214.
Kucyi, A., Salomons, T.V., Davis, K.D., 2013. Mind wandering away from pain dynamically engages antinociceptive and default mode brain networks. Proc. Natl. Acad. Sci. U. S. A. 110, 18692–18697.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.
Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.
Li, B., Daunizeau, J., Stephan, K.E., Penny, W., Hu, D., Friston, K., 2011. Generalised filtering and stochastic DCM for fMRI. NeuroImage 58, 442–457.
Lieder, F., Daunizeau, J., Garrido, M.I., Friston, K.J., Stephan, K.E., 2013. Modelling trial-by-trial changes in the mismatch negativity. PLoS Comput. Biol. 9, e1002911.
Lythe, K.E., Moll, J., Gethin, J.A., Workman, C.I., Green, S., Lambon Ralph, M.A., Deakin, J.F., Zahn, R., 2015. Self-blame-selective hyperconnectivity between anterior temporal and subgenual cortices and prediction of recurrent depressive episodes. JAMA Psychiatry 72, 1119–1126.
MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge.
Maia, T.V., Frank, M.J., 2011. From reinforcement learning models to psychiatric and neurological disorders. Nat. Neurosci. 14, 154–162.
Mansson, K.N., Frick, A., Boraxbekk, C.J., Marquand, A.F., Williams, S.C., Carlbring, P., Andersson, G., Furmark, T., 2015. Predicting long-term outcome of internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl. Psychiatry 5, e530.
Mathys, C., Daunizeau, J., Friston, K.J., Stephan, K.E., 2011. A Bayesian foundation for individual learning under uncertainty. Front. Hum. Neurosci. 5, 39.
McGrath, C.L., Kelley, M.E., Holtzheimer, P.E., Dunlop, B.W., Craighead, W.E., Franco, A.R., Craddock, R.C., Mayberg, H.S., 2013. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA Psychiatry 70, 821–829.
Miller, J.M., Schneck, N., Siegle, G.J., Chen, Y., Ogden, R.T., Kikuchi, T., Oquendo, M.A., Mann, J.J., Parsey, R.V., 2013. fMRI response to negative words and SSRI treatment outcome in major depressive disorder: a preliminary study. Psychiatry Res. 214, 296–305.
Montague, P.R., Hyman, S.E., Cohen, J.D., 2004. Computational roles for dopamine in behavioural control. Nature 431, 760–767.
Montague, P.R., Dolan, R.J., Friston, K.J., Dayan, P., 2012. Computational psychiatry. Trends Cogn. Sci. 16, 72–80.
Moran, R.J., Stephan, K.E., Kiebel, S.J., Rombach, N., O'Connor, W.T., Murphy, K.J., Reilly, R.B., Friston, K.J., 2008. Bayesian estimation of synaptic physiology from the spectral responses of neural masses. Neuroimage. 42 (1), 272–284.
Moran, R.J., Stephan, K.E., Seidenbecher, T., Pape, H.C., Dolan, R.J., Friston, K.J., 2009. Dynamic causal models of steady-state responses. NeuroImage 44, 796–811.
Moran, R.J., Symmonds, M., Stephan, K.E., Friston, K.J., Dolan, R.J., 2011. An in vivo assay of synaptic function mediating human cognition. Curr. Biol. 21, 1320–1325.
Moran, R.J., Campo, P., Symmonds, M., Stephan, K.E., Dolan, R.J., Friston, K.J., 2013. Free energy, precision and learning: the role of cholinergic neuromodulation. J. Neurosci. 33, 8227–8236.
Murray, G.K., Corlett, P.R., Clark, L., Pessiglione, M., Blackwell, A.D., Honey, G., Jones, P.B., Bullmore, E.T., Robbins, T.W., Fletcher, P.C., 2008. Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. Mol. Psychiatry 13 (239), 267–276.
Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental sparse and other variants. In: Jordan, M.I. (Ed.), Learning in Graphical Models. Kluwer Academic Publishers, Dordrecht.
Nejad, A.B., Madsen, K.H., Ebdrup, B.H., Siebner, H.R., Rasmussen, H., Aggernaes, B., Glenthoj, B.Y., Baare, W.F., 2013. Neural markers of negative symptom outcomes in distributed working memory brain activity of antipsychotic-naive schizophrenia patients. Int. J. Neuropsychopharmacol. 16, 1195–1204.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337.
Orru, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36, 1140–1152.
Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T.T., Kiebel, S.J., Blankenburg, F., 2012. Evidence for neural encoding of Bayesian surprise in human somatosensation. NeuroImage 62, 177–188.
Owen, M.J., 2014. New approaches to psychiatric diagnostic classification. Neuron 84, 564–571.
Payzan-LeNestour, E., Dunne, S., Bossaerts, P., O'Doherty, J.P., 2013. The neural representation of unexpected uncertainty during value-based decision making. Neuron 79, 191–201.
Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. NeuroImage 59, 319–330.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004a. Comparing dynamic causal models. NeuroImage 22, 1157–1172.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004b. Modelling functional integration: a comparison of structural equation and dynamic causal models. NeuroImage 23 (Suppl. 1), S264–S274.

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, e1000709.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209.
Piray, P., den Ouden, H.E., van der Schaaf, M.E., Toni, I., Cools, R., 2015. Dopaminergic modulation of the functional ventrodorsal architecture of the human striatum. Cereb. Cortex.
Pitt, M.A., Myung, I.J., 2002. When a good fit can be bad. Trends Cogn. Sci. 6, 421–425.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 229.
Price, C.J., Friston, K.J., 2002. Degeneracy and cognitive anatomy. Trends Cogn. Sci. 6, 416–421.
Raman, S., Deserno, L., Schlagenhauf, F., Stephan, K.E., 2016. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. J. Neurosci. Methods (in press).
Ranlund, S., Adams, R.A., Diez, A., Constante, M., Dutt, A., Hall, M.H., Maestro Carbayo, A., McDonald, C., Petrella, S., Schulze, K., Shaikh, M., Walshe, M., Friston, K., Pinotsis, D., Bramon, E., 2015. Impaired prefrontal synaptic gain in people with psychosis and their relatives during the mismatch negativity. Hum. Brain Mapp.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton Century Crofts, New York, pp. 64–99.
Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D., 2011. Decoding brain states from fMRI connectivity graphs. NeuroImage 56, 616–626.
Rigoux, L., Daunizeau, J., 2015. Dynamic causal modelling of brain-behaviour relationships. NeuroImage 117, 202–221.
Rigoux, L., Stephan, K.E., Friston, K.J., Daunizeau, J., 2014. Bayesian model selection for group studies - revisited. NeuroImage 84, 971–985.
Romaniuk, L., Honey, G.D., King, J.R., Whalley, H.C., McIntosh, A.M., Levita, L., Hughes, M., Johnstone, E.C., Day, M., Lawrie, S.M., Hall, J., 2010. Midbrain activation during Pavlovian conditioning and delusional symptoms in schizophrenia. Arch. Gen. Psychiatry 67, 1246–1254.
Rosa, M.J., Portugal, L., Hahn, T., Fallgatter, A.J., Garrido, M.I., Shawe-Taylor, J., Mourao-Miranda, J., 2015. Sparse network-based models for patient classification using fMRI. NeuroImage 105, 493–506.
Rowe, J.B., Hughes, L.E., Barker, R.A., Owen, A.M., 2010. Dynamic causal modelling of effective connectivity from fMRI: are results reproducible and sensitive to Parkinson's disease and its treatment? NeuroImage 52, 1015–1026.
Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., Wager, T.D., 2014. Representation of aversive prediction errors in the human periaqueductal gray. Nat. Neurosci. 17, 1607–1612.
Rutledge, R.B., Skandali, N., Dayan, P., Dolan, R.J., 2014. A computational and neural model of momentary subjective well-being. Proc. Natl. Acad. Sci. U. S. A. 111, 12252–12257.
Schlagenhauf, F., Huys, Q.J., Deserno, L., Rapp, M.A., Beck, A., Heinze, H.J., Dolan, R., Heinz, A., 2014. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. NeuroImage 89, 171–180.
Schmaal, L., Marquand, A.F., Rhebergen, D., van Tol, M.J., Ruhe, H.G., van der Wee, N.J., Veltman, D.J., Penninx, B.W., 2015. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. Biol. Psychiatry 78, 278–286.
Schmidt, A., Smieskova, R., Aston, J., Simon, A., Allen, P., Fusar-Poli, P., McGuire, P.K., Riecher-Rossler, A., Stephan, K.E., Borgwardt, S., 2013. Brain connectivity abnormalities predating the onset of psychosis: correlation with the effect of medication. JAMA Psychiatry 70, 903–912.
Schofield, T.M., Penny, W.D., Stephan, K.E., Crinion, J.T., Thompson, A.J., Price, C.J., Leff, A.P., 2012. Changes in auditory feedback connections determine the severity of speech processing deficits after stroke. J. Neurosci. 32, 4260–4270.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Schwartenbeck, P., FitzGerald, T.H., Mathys, C., Dolan, R., Friston, K., 2015. The dopaminergic midbrain encodes the expected certainty about desired outcomes. Cereb. Cortex 25, 3434–3445.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., Frackowiak, R.S., 2004. Temporal difference models describe higher-order learning in humans. Nature 429, 664–667.
Silva, R.F., Castro, E., Gupta, C.N., Cetin, M., Arbabshirani, M., Potluru, V.K., Plis, S.M., Calhoun, V.D., 2014. The tenth annual MLSP competition: Schizophrenia classification challenge. IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2014.
Sladky, R., Hoflich, A., Kublbock, M., Kraus, C., Baldinger, P., Moser, E., Lanzenberger, R., Windischberger, C., 2015. 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.
Stephan, K.E., Mathys, C., 2014. Computational approaches to psychiatry. Curr. Opin. Neurobiol. 25, 85–92.
Stephan, K.E., Marshall, J.C., Friston, K.J., Rowe, J.B., Ritzl, A., Zilles, K., Fink, G.R., 2003. Lateralized cognitive processes and lateralized task control in the human brain. Science 301, 384–386.
Stephan, K.E., Penny, W.D., Marshall, J.C., Fink, G.R., Friston, K.J., 2005. Investigating the functional role of callosal connections with dynamic causal models. Ann. N. Y. Acad. Sci. 1064, 16–36.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. NeuroImage 38, 387–401.
Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fMRI. NeuroImage 42, 649–662.
Stephan, K.E., Friston, K.J., Frith, C.D., 2009a. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. Schizophr. Bull. 35, 509–527.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009b. Bayesian model selection for group studies. NeuroImage 46, 1004–1017.
Stephan, K.E., Tittgemeyer, M., Knosche, T.R., Moran, R.J., Friston, K.J., 2009c. Tractography-based priors for dynamic causal models. NeuroImage 47, 1628–1638.
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.
Stephan, K.E., Iglesias, S., Heinzle, J., Diaconescu, A.O., 2015. Translational perspectives for computational neuroimaging. Neuron 87, 716–732.
Stephan, K.E., Bach, D.R., Fletcher, P.C., Flint, J., Frank, M.J., Friston, K.J., Heinz, A., Huys, Q.J., Owen, M.J., Binder, E.B., Dayan, P., Johnstone, E.C., Meyer-Lindenberg, A., Montague, P.R., Schnyder, U., Wang, X.J., Breakspear, M., 2016. Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. Lancet Psychiatry 3, 77–83.
Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., Hirsch, J., 2006. Predictive codes for forthcoming perception in the frontal cortex. Science 314, 1311–1314.
Tanabe, J., Reynolds, J., Krmpotich, T., Claus, E., Thompson, L.L., Du, Y.P., Banich, M.T., 2013. Reduced neural tracking of prediction error in substance-dependent individuals. Am. J. Psychiatry 170, 1356–1363.
van Leeuwen, T.M., den Ouden, H.E., Hagoort, P., 2011. Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. J. Neurosci. 31, 9879–9884.
van Schouwenburg, M.R., Zwiers, M.P., van der Schaaf, M.E., Geurts, D.E., Schellekens, A.F., Buitelaar, J.K., Verkes, R.J., Cools, R., 2013. Anatomical connection strength predicts dopaminergic drug effects on fronto-striatal function. Psychopharmacology 227, 521–531.
Vossel, S., Mathys, C., Stephan, K.E., Friston, K.J., 2015. Cortical coupling reflects Bayesian belief updating in the deployment of spatial attention. J. Neurosci. 35, 11532–11542.
Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. N. Engl. J. Med. 368, 1388–1397.
Wang, X.J., Krystal, J.H., 2014. Computational psychiatry. Neuron 84, 638–654.
Weiss, P.H., Fink, G.R., 2009. Grapheme-colour synaesthetes show increased grey matter volumes of parietal and fusiform cortex. Brain 132, 65–70.
Wiecki, T.V., Poland, J.S., Frank, M.J., 2015. Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. Clin. Psychol. Sci. (in press).
Wilkinson, D., Halligan, P., 2004. The relevance of behavioural measures for functional-imaging studies of cognition. Nat. Rev. Neurosci. 5, 67–73.
Winton-Brown, T.T., Fusar-Poli, P., Ungless, M.A., Howes, O.D., 2014. Dopaminergic basis of salience dysregulation in psychosis. Trends Neurosci. 37, 85–94.
Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349.
Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron 46, 681–692.
Zhang, J., Rittman, T., Nombela, C., Fois, A., Coyle-Gilchrist, I., Barker, R.A., Hughes, L.E., Rowe, J.B., 2016. Different decision deficits impair response inhibition in progressive supranuclear palsy and Parkinson's disease. Brain 139, 161–173.