HAL
open science

Dynamic causal models of neural system dynamics:current state and future extensions.

Klaas E. Stephan, Lee M. Harrison, Stefan J. Kiebel, Olivier David, Will D. Penny, Karl J. Friston

To cite this version:

Klaas E. Stephan, Lee M. Harrison, Stefan J. Kiebel, Olivier David, Will D. Penny, et al.. Dynamic causal models of neural system dynamics:current state and future extensions.. Journal of Biosciences, 2007, 32 (1), pp.129-44. inserm-00381759

HAL Id: inserm-00381759

https://inserm.hal.science/inserm-00381759

Submitted on 23 Jul 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dynamic causal models of neural system dynamics: current state and future extensions

Klaas E. Stephan[1], Lee M. Harrison[1], Stefan J. Kiebel[1], O. David[2], W.D. Penny[1] & Karl J. Friston[1]

[1] Wellcome Dept. of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK.

[2] INSERM U594 Neuroimagerie Fonctionnelle et Métabolique, Université Joseph Fourier, CHU – Pavillon B – BP 217, 38043 Grenoble Cedex 09, France

Address correspondence to:
Dr Klaas Enno Stephan
Wellcome Dept. of Imaging Neuroscience
Institute of Neurology, University College London
12 Queen Square
London WC1N 3BG, UK
phone +44-207-8337472
fax +44-207-8131420
e-mail: k.stephan@fil.ion.ucl.ac.uk

Keywords: fMRI, EEG, event-related potentials, effective connectivity, neural system, dynamic causal modelling

Word Count: abstract: 198 words
main text: 5,556 words (excl. references and legends)

Abstract

Complex processes resulting from the interaction of multiple elements can rarely be understood by analytical procedures; instead, mathematical models of system dynamics are required. This insight, which disciplines like physics have embraced for a long time already, is gradually gaining importance in the study of cognitive processes by functional neuroimaging. In this field, causal mechanisms in neural systems are described in terms of effective connectivity. Recently, Dynamic Causal Modelling (DCM) was introduced as a generic method to estimate effective connectivity from neuroimaging data in a Bayesian fashion. One of the key advantages of DCM over previous methods is that it distinguishes between neural state equations and modality-specific forward models that translate neural activity into a measured signal. Another strength is its natural relation to Bayesian Model Selection (BMS) procedures. In this article, we review the conceptual and mathematical basis of DCM and its implementation for functional magnetic resonance imaging data and event-related potentials. After introducing the application of BMS in the context of DCM, we conclude with an outlook to future extensions of DCM. These extensions are guided by the long-term goal of using dynamic system models for pharmacological and clinical applications, particularly with regard to synaptic plasticity.

Introduction

Modern cognitive neuroscience uses a variety of non-invasive techniques for measuring brain activity. These techniques include electrophysiological methods, e.g. electroencephalography (EEG) and magnetoencephalograpy (MEG), and functional imaging methods, e.g. positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). Two intertwined concepts, functional specialization and functional integration, have been guiding neuroimaging applications over the last decades (Friston 2002a). Functional specialization assumes a local specialization for certain aspects of information processing but allows for the possibility that this specialization is anatomically segregated across different cortical areas. Most current functional neuroimaging experiments use this perspective and interpret the areas that are activated by a certain task component as the elements of a distributed system. However, this explanation is somewhat speculative and clearly incomplete as long as one does not characterize how the local computations are bound together by context-dependent interactions among these areas. This binding is the functional integration within the system which can be characterized in two ways, namely in terms of functional connectivity and effective connectivity. While functional connectivity describes statistical dependencies between data, effective connectivity rests on a mechanistic model of the causal effects that generated the data (Friston 1994). This article focuses exclusively on a recently established technique for determining the effective connectivity in neural systems of interest on the basis of measured fMRI and EEG/MEG data: Dynamic Causal Modelling (DCM; Friston et al. 2003). We review the conceptual and mathematical basis of DCM and Bayesian Model Selection (BMS; Penny et al. 2004a)

and demonstrate some applications, using empirical and simulated data. We also touch on some future extensions of DCM that are driven by the long-term goal of using DCM for pharmacological and clinical applications, particularly with regard to questions about synaptic plasticity.

Effective connectivity and a general state equation for neural systems

The term *effective connectivity* has been defined by various authors in convergent ways. A general definition is that effective connectivity describes the causal influences that neural units exert over another (Friston 1994). More specifically, other authors have proposed that "effective connectivity should be understood as the experiment- and time-dependent, simplest possible circuit diagram that would replicate the observed timing relationships between the recorded neurons" (Aertsen & Preißl 1991). Both definitions emphasize that determining effective connectivity requires a causal model of the interactions between the elements of the neural system of interest. Before we describe the specifics of the model on which DCM rests, let us derive a general mathematical form of models of effective connectivity. For this purpose, we choose deterministic differential equations with time-invariant parameters as a mathematical framework. Note that these are not the only possible mathematical representation of systems; in fact, many alternatives exist, e.g. state space models or iterative maps. The underlying concept, however, is quite universal: a *system* is defined by a set of elements with n time-variant properties that interact with each other. Each time-variant property x_i ($1 \leq i \leq n$) is called a *state variable*, and the n-vector $x(t)$ of all state variables in the system is called the *state vector* (or simply *state*) of the system at time t:

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \qquad (1)$$

Taking an ensemble of interacting neurons as an example, the system elements would correspond to the individual neurons, each of which is represented by one or several state variables. These state variables could refer to various neurophysiological properties, e.g. postsynaptic potentials, status of ion channels, etc. Critically, the state variables interact with each other, i.e. the evolution of each state variable depends on at least one other state variable. For example, the postsynaptic membrane potential depends on which and how many ion channels are open; vice versa, the probability of voltage-dependent ion channels opening depends on the membrane potential. Such mutual functional dependencies between the state variables of the system can be expressed quite naturally by a set of ordinary differential equations that operate on the state vector:

$$\frac{dx}{dt} = \begin{bmatrix} f_1(x_1,\ldots,x_n) \\ \vdots \\ f_n(x_1,\ldots,x_n) \end{bmatrix} = F(x) \qquad (2)$$

However, this description is not yet sufficient. First of all, the specific form of the dependencies f_i needs to be specified, i.e. the nature of the causal relations between state variables. This requires a set of parameters θ which determine the form and strength of influences between state variables. In neural systems, these parameters usually correspond to time constants or synaptic strengths of the connections between the system elements. The Boolean nature of θ, i.e. the pattern of absent and present connections, and the mathematical form of the dependencies f_i represent the *structure* of the system. And second, in the case of non-autonomous systems (and these are the ones of interest to

biology) we need to consider the input into the system, e.g. sensory information entering the brain. We represent the set of all m known inputs by the m-vector function $u(t)$. Extending Eq. 2 accordingly leads to a general state equation for non-autonomous deterministic systems

$$\frac{dx}{dt} = F(x, u, \theta) \qquad (3)$$

A model whose form follows this general state equation provides a causal description of how system dynamics results from system structure, because it describes (i) when and where external inputs enter the system and (ii) how the state changes induced by these inputs evolve in time depending on the system's structure. Given a particular temporal sequence of inputs $u(t)$ and an initial state $x(0)$, one obtains a complete description of how the dynamics of the system (i.e. the trajectory of its state vector x in time) results from its structure by integration of Eq. 3:

$$x(\tau) = x(0) + \int_0^{\tau} F(x, u, \theta) dt \qquad (4)$$

Equation 3 therefore provides a general form for models of effective connectivity in neural systems. As described elsewhere (Friston et al. 2003; Stephan 2004), all established models of effective connectivity can be related to this general equation or, in the case of regression-like models, its static equivalent $x = F(x, u, \theta)$. In the next sections, we show how DCM models neural population dynamics using a bilinear implementation of this general form. This is combined with a forward model that translates neural activity into a measured signal.

Before we proceed to DCM, it is worth pointing out that we have made two main assumptions in this section to simplify the exposition to the general state equation. First, it is assumed that all processes in the system are deterministic and occur instantaneously. Random components (noise) and delays could be accounted for by using stochastic differential equations and delay differential equations, respectively. Second, we assume that we know the inputs that enter the system. This is a tenable assumption in neuroimaging because the inputs are experimentally controlled variables, e.g. changes in stimuli or instructions. It may also be helpful to point out that using time-invariant dependencies f_i and parameters θ is not a restriction. Although the mathematical form of f_i per se is static, the use of time-varying inputs u allows for dynamic changes in what components of f_i are "activated". For example, input functions that can only take values of one or zero and that are multiplied with the different terms of a polynomial function can be used to induce time-dependent changes from nonlinear to linear behaviour (e.g. by "switching off" all higher order terms in the polynomial) or vice versa. Also, there is no principled distinction between states and time-invariant parameters. Therefore, estimating time-varying parameters can be treated as a state estimation problem.

Principles of Dynamic Causal Modelling (DCM)

An important limitation of previous methods for determining effective connectivity from functional imaging data, e.g. structural equation modelling (Büchel & Friston 1997; McIntosh & Gonzalez-Lima 1994) or multivariate autoregressive models (Goebel et al. 2003; Harrison et al. 2003), is that they operate at the level of the measured signals. This is a serious problem because the causal architecture of the system that we would like to identify is expressed at the level of neuronal dynamics which is not directly observed

using non-invasive techniques. In the case of fMRI data, for example, previous models of effective connectivity were fitted to the measured time series which result from a haemodynamic convolution of the underlying neural activity. Since classical models do not include the forward model linking neuronal activity to the measured haemodynamic data, analyses of inter-regional connectivity performed at the level of haemodynamic responses are problematic. For example, different brain regions can exhibit marked differences in neurovascular coupling, and these differences, expressed in different latencies, undershoots, etc., may lead to false inference about connectivity. A similar situation is seen with EEG data where there is a big difference between signals measured at each electrode and the underlying neuronal activity: changes in neural activity in different brain regions lead to changes in electric potentials that superimpose linearly. The scalp electrodes therefore record a mixture, with unknown weightings, of potentials generated by a number of different sources.

Therefore, to enable inferences about connectivity between neural units we need models that combine two things: (i) a parsimonious but neurobiologically plausible model of neural population dynamics, and (ii) a biophysically plausible forward model that describes the transformation from neural activity to the measured signal. Such models make it possible to fit jointly the parameters of the neural and of the forward model such that the predicted time series are optimally similar to the observed time series. This combination of a model of neural dynamics with a biophysical forward model is a core feature of DCM. Currently, DCM implementations exist both for fMRI data and event-related potentials (ERPs) as measured by EEG/MEG. These modality-specific implementations are briefly summarized in the next sections.

DCM for fMRI

DCM for fMRI uses a simple model of neural dynamics in a system of n interacting brain regions. It models the change of a neural state vector x in time, with each region in the system being represented by a single state variable, using the following bilinear differential equation:

$$\begin{aligned} \frac{dx}{dt} &= F\left(x, u, \theta^{n}\right) \\ &= \left(A + \sum_{j=1}^{m} u_j B^{(j)} \right) x + Cu \end{aligned} \tag{5}$$

Note that this neural state equation follows the general form for deterministic system models introduced by Equation 3, i.e. the modelled state changes are a function of the system state itself, the inputs u and some parameters $\theta^{(n)}$ that define the functional architecture and interactions among brain regions at a neuronal level. The neural state variables represent a summary index of neural population dynamics in the respective regions. The neural dynamics are driven by experimentally controlled external inputs that can enter the model in two different ways: they can elicit responses through direct influences on specific regions (e.g. evoked responses in early sensory cortices; the C matrix) or they can modulate the coupling among regions (e.g. during learning or attention; the B matrices). Given this bilinear state equation, the neural parameters $\theta^{(n)} = \{A, B, C\}$ can be expressed as partial derivatives of F:

$$
\begin{aligned}
A &= \left.\frac{\partial F}{\partial x}\right|_{u=0} \\
B^{(j)} &= \frac{\partial^2 F}{\partial x \partial u_j} \\
C &= \left.\frac{\partial F}{\partial u}\right|_{x=0}
\end{aligned}
\tag{6}
$$

As can be seen from these equations, the matrix A represents the effective connectivity among the regions in the absence of input, the matrices $B^{(j)}$ encode the change in effective connectivity induced by the jth input u_j, and C embodies the strength of direct influences of inputs on neuronal activity. Figure 1 summarises this bilinear state equation and shows a specific example model.

[Figure 1 near here]

DCM for fMRI combines this model of neural dynamics with an experimentally validated haemodynamic model that describes the transformation of neuronal activity into a BOLD response. This so-called "Balloon model" was initially formulated by Buxton et al. (1998) and later extended by Friston et al. (2000). Briefly, it consists of a set of differential equations that describe the relations between four haemodynamic state variables, using five parameters $\theta^{(h)}$. More specifically, changes in neural activity elicit a vasodilatory signal that leads to increases in blood flow and subsequently to changes in blood volume and deoxyhaemoglobin content. The predicted BOLD signal is a non-linear function of blood volume and deoxyhaemoglobin content. This haemodynamic model is summarised by Figure 2 and described in detail by Friston et al. (2000).

[Figure 2 near here]

The combined neural and haemodynamic parameter set $\theta = \{\theta^{(n)}, \theta^{(h)}\}$ is estimated from the measured BOLD data, using a fully Bayesian approach with empirical priors for the haemodynamic parameters and conservative shrinkage priors for the coupling parameters. Details of the parameter estimation scheme, which rests on a gradient ascent procedure embedded into an expectation maximization (EM) algorithm and uses a Laplace (i.e. Gaussian) approximation to the true posterior, can be found in Friston (2002b).

Once the parameters of a DCM have been estimated from measured BOLD data, the posterior distributions of the parameter estimates can be used to test hypotheses about connection strengths. Due to the Laplace approximation, the posterior distributions are defined by their posterior mode or maximum a posteriori (MAP) estimate and their posterior covariance. Usually, the hypotheses to be tested concern context-dependent changes in coupling. A classical example is given by Figure 3. Here, DCM was applied to fMRI data from a single subject, testing the hypothesis that in a hierarchical system of visual areas (c.f. Figure 1) attention to motion enhanced the backward connections from the inferior frontal gyrus (IFG) onto superior parietal cortex (SPC) and from SPC onto V5, respectively. Other examples of single-subject analyses can be found in Mechelli et al. (2003), Penny et al. (2004b) and Stephan et al. (2005). For statistical inference at the group level, various options exist. The simplest approach is to enter the conditional estimates of interest into a classical second-level analysis; for examples see Bitan et al. (2005) and Smith et al. (2006). A more coherent approach may be to use Bayesian analyses at the group level as well (e.g. Garrido et al. 2006).

[Figure 3 near here]

Fitted to regional fMRI time series, a given DCM explains how local brain responses were generated from the interplay of the three mechanisms described by the state equation (Eq. 5): inter-regional connections, their contextual modulation and driving inputs. Figure 4 provides a simple fictitious example that is based on simulated data. In this example we fix the parameters and use DCM as a model to generate synthetic data, as opposed to its usual use, i.e. estimating parameter values from empirical data. Let us imagine we are dealing with a 2x2 factorial experiment (Fig. 4A) where one experimental factor controls sensory stimulation (stimulus S_1 vs. stimulus S_2) and a second factor controls task requirements (task T_1 vs. task T_2). Let us further imagine that, using conventional statistical parametric mapping, we had found a main effect of sensory stimulation in a particular brain area x_1 (with observed time series y_1; see Fig. 4B, upper panel) and a stimulus-by-task interaction in area x_2 (with observed time series y_2). This interaction means that the difference between stimulus S_1 and stimulus S_2 is larger during task T_1 than during task T_2 (see Fig. 4B, lower panel). We can generate the (noise-free) data shown in Fig. 4B using the DCM displayed by Fig. 4C. The stimulus main effect in area x_1 results from the driving inputs to x_1 being much stronger for stimulus S_1 than for stimulus S_2. This differential effect is then conveyed onto area x_2 by the connection from x_1 to x_2. Critically, the strength of this connection is strongly enhanced during task T_1, but only marginally influenced during task T_2. This difference in modulation causes the interaction in area x_2 (note that this model would have produced an interaction in area x_1 as well if we had chosen a stronger back-connection from x_2 to x_1).

Usually, of course, DCM is applied in the reverse fashion, i.e. to estimate the parameters $\theta^{(n)} = \{A, B, C\}$ from measured fMRI data as in Fig. 4B. The goal is to infer

the neuronal mechanisms that have shaped local brain responses, e.g. the presence of main effects or interactions. Simulations like the one described above can also be used to explore the robustness of parameter estimation in DCM. For example, one can generate data multiple times, adding observation noise (see Fig. 4D), and then trying to re-estimate the parameters from the noisy data.

[Figure 4 near here]

We are currently working on various extensions to DCM for fMRI. For example, one extension is to represent each region in the model by multiple state variables, e.g. populations of excitatory and inhibitory neurons (Marreiros et al., in preparation; see also Harrison et al. 2005). A similar approach has already been implemented in DCM for ERPs which is described in the following section. Another extension for DCM for fMRI is to augment the state equation by additional non-linear, e.g. quadratic, terms (Stephan et al., in preparation):

$$\frac{dx}{dt} = \left(A + \sum_{j=1}^{m} u_j B^{(j)} \right) x + Cu + \begin{bmatrix} x^T D^{(1)} x \\ \vdots \\ x^T D^{(n)} x \end{bmatrix} \tag{7}$$

This extension enhances the kind of dynamics that DCM can capture and enables the user to implement additional types of models. For example, beyond modelling how connection strengths are modulated by external inputs, one can now model how connection strengths change as a function of the output from areas. This ability is critical for various applications, e.g. for marrying reinforcement learning models with DCM (c.f. Stephan 2004). In a neural system model of descriptive learning theories like temporal difference learning, the prediction error, encoded by the activity of a particular neural

unit, determines the change of connection strength between other neural units that encode properties of conditional and unconditional stimuli (see Schultz & Dickinson 2000). Figure 5A shows a simulation example where the connection from an area x_1 to another area x_2 is enhanced multiplicatively by the output from a third region x_3, i.e. $\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2 + d_{13}^{(2)}x_1x_3$. Critically, x_3 is not only driven by external inputs, but also receives an input from x_2. This means that for an excitatory connection from x_2 to x_3, a positive reinforcement effect results: the higher activity in x_3, the more strongly inputs from x_1 to x_2 will be enhanced, leading to higher activity in x_2, which, in turn, drives x_3 even further. Figure 5B shows an example of this effect, using simulated data. Such a model, of course, lives on the brink of stability and is prone to runaway excitation. In contrast, an inhibitory connection from x_2 to x_3 makes the model extremely stable because the higher the activity in x_3, the higher the response in x_2 to x_1 inputs and thus the stronger the inhibitory feedback onto x_3 (not shown here).

[Figure 5 near here]

DCM for ERPs

ERPs as measured with EEG or MEG have been used for decades to study electrophysiological correlates of cognitive operations. Nevertheless, the neurobiological mechanisms that underlie their generation are still largely unknown. DCM for ERPs was developed as a biologically plausible model to understand how event-related responses result from the dynamics in coupled neural ensembles. It rests on a neural mass model which uses established connectivity rules in hierarchical sensory systems to assemble a

network of coupled cortical sources (David & Friston 2003; David et al. 2005). These rules characterise connections with respect to their laminar patterns of origin and termination and distinguish between (i) forward (or bottom-up) connections originating in agranular layers and terminating in layer 4, (ii) backward (or top-down) connections originating and terminating in agranular layers, and (iii) lateral connections originating in agranular layers and targeting all layers. These inter-areal cortico-cortical connections are excitatory, using glutamate as neurotransmitter, and arise from pyramidal cells.

[Figure 6 near here]

Each region or source is modelled as a microcircuit in which three neuronal subpopulations are combined and assigned to granular and supra-/infragranular layers. A population of excitatory pyramidal (output) cells receives inputs from inhibitory and excitatory populations of interneurons via intrinsic (intra-areal) connections. Within this model, excitatory interneurons can be regarded as spiny stellate cells found predominantly in layer 4 and in receipt of forward connections. Excitatory pyramidal cells and inhibitory interneurons are considered to occupy infra- and supragranular layers and receive backward and lateral inputs.

The neural state equations are summarized in Figure 7. To perturb the system and model event-related responses, the network receives inputs via input connections. These connections are exactly the same as forward connections and deliver input u to the spiny stellate cells in layer 4. Input u represents afferent activity relayed by subcortical structures and are modelled as two parameterized components, a gamma density function (representing an event-related burst of input that is delayed and dispersed by subcortical

synapses and axonal conduction) and a discrete cosine set (representing fluctuations in input over peristimulus time). The influence of this input on each source is controlled by a parameter vector C. Overall, the DCM is specified in terms of the state equations shown in Figure 7 and a linear forward model

$$\begin{aligned} \frac{dx}{dt} &= f(x,u,\theta) \\ y &= Lx_0 + \varepsilon \end{aligned} \tag{8}$$

where x_0 represents the transmembrane potential of pyramidal cells and L is a lead field matrix coupling electrical sources to the EEG channels. In comparison to DCM for fMRI, the forward model is a simple linearity as opposed to the nonlinear haemodynamic model in DCM for fMRI. In contrast, the state equations of DCM for ERPs are much more complex and realistic (compare Figures 1 and 7). As an example, the state equation for the inhibitory subpopulation is

$$\begin{aligned} \dot{x}_7 &= x_8 \\ \dot{x}_8 &= \frac{H_e}{\tau_e}((C^B + C^L + \gamma_3 I)S(x_0)) - \frac{2x_8}{\tau_e} - \frac{x_7}{\tau_e^2} \end{aligned} \tag{9}$$

The parameter matrices C^F, C^B, C^L encode forward, backward and lateral connections respectively. Within each subpopulation, the dynamics of neural states are determined by two operators. The first transforms the average density of presynaptic inputs into the average postsynaptic membrane potential. This is modelled by a linear transformation with excitatory (e) and inhibitory (i) kernels parameterized by $H_{e,i}$ and $\tau_{e,i}$. $H_{e,i}$ control the maximum postsynaptic potential and $\tau_{e,i}$ represent lumped rate constants (i.e. lumped across dendritic spines and the dendritic tree). The second operator S transforms the

average potential of each subpopulation into an average firing rate. This is assumed to be instantaneous and is a sigmoid function. Intra-areal interactions among the subpopulations depend on constants $\gamma_{1...4}$ which control the strength of intrinsic connections and reflect the total number of synapses expressed by each subpopulation. In Equation 9, the top line expresses the rate of change of voltage as a function of current. The second line specifies how current changes as a function of voltage, current and presynaptic input from extrinsic and intrinsic sources. For simplification, our description here has omitted the fact that in DCM for ERPs all intra- and inter-areal connections have conduction delays. This is implemented by delay differential equations.

[Figure 7 near here]

Just as with DCM for fMRI, the DCM for ERPs is usually used to investigate whether coupling strengths change as a function of experimental context. Figure 8 shows an example of a DCM applied to EEG data from a single subject performing an auditory oddball task (David et al. 2006): forward and backward connections between primary auditory and orbitofrontal cortex are stronger during processing of oddball stimuli compared to standard stimuli.

[Figure 8 near here]

Several extensions of DCM for ERPs are planned or already under way. For example, Kiebel et al. (2006) demonstrated that one does not necessarily have to assume known lead field parameters (L in Eq. 8) for the forward model. Instead, it is possible to estimate lead-field and coupling parameters simultaneously and thus use DCM for ERPs as a source reconstruction approach with physiologically informed constraints. Future

efforts will concentrate on further enhancing the biological realism of the model. One approach may be to free the coupling parameters between the neuronal populations *within* regions, thus enabling a characterisation of local interactions between excitatory and inhibitory neuronal populations, for example. Another and more long-term goal will be to represent the effects of particular neurotransmitters, like dopamine or acetylcholine, on neuronal population activity. This will be particularly important for potential clinical applications of DCM (see below).

Bayesian Model Selection (BMS)

A generic problem encountered by any kind of modelling approach is the question of model selection: given some observed data, which of several alternative models is the optimal one? This problem is not trivial because the decision cannot be made solely by comparing the relative fit of the competing models. One also needs to take into account the relative complexity of the models as expressed, for example, by the number of free parameters in each model. Model complexity is important to consider because there is a trade-off between model fit and generalisability (i.e. how well the model explains different data sets that were all generated from the same underlying process). As the number of free parameters is increased, model fit increases monotonically whereas beyond a certain point model generalisability decreases. The reason for this is "overfitting": an increasingly complex model will, at some point, start to fit noise that is specific to one data set and thus become less generalisable across multiple realizations of the same underlying generative process. (Generally, in addition to the number of free parameters, the complexity of a model also depends on its functional form; see Pitt &

Myung 2002. This is not an issue for DCM, however, because here competing models usually have the same functional form.)

Therefore, the question "What is the optimal model?" can be reformulated more precisely as "What is the model that represents the best balance between fit and complexity?" In a Bayesian context, the latter question can be addressed by comparing the evidence, $p(y \mid m)$, of different models. According to Bayes theorem

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta \mid m)}{p(y \mid m)} \tag{10}$$

the model evidence can be considered as a normalization constant for the product of the likelihood of the data and the prior probability of the parameters, therefore

$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta \mid m)\, d\theta \tag{11}$$

Here, the number of free parameters (as well as the functional form) are considered by the integration. Unfortunately, this integral cannot usually be solved analytically, therefore an approximation to the model evidence is needed.

In the context of DCM, one potential solution could be to make use of the Laplace approximation, i.e. to approximate the model evidence by a Gaussian that is centered on its mode. As shown by Penny et al. (2004a), this yields the following expression for the natural logarithm (*ln*) of the model evidence ($\eta_{\theta|y}$ denotes the MAP estimate, $C_{\theta|y}$ is the posterior covariance of the parameters, C_{ε} is the error covariance, θ_p is the prior mean of the parameters, and C_p is the prior covariance):

$$
\begin{aligned}
\ln p(y \mid m) &= accuracy(m) - complexity(m) \\
&= \left[-\frac{1}{2}\ln\left|C_\varepsilon\right| - \frac{1}{2}\left(y - h(u,\eta_{\theta|y}\right)^T C_\varepsilon^{-1}\left(y - h(u,\eta_{\theta|y}\right) \right] \\
&- \left[\frac{1}{2}\ln\left|C_p\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\eta_{\theta|y} - \theta_p\right)^T C_p^{-1}\left(\eta_{\theta|y} - \theta_p\right) \right]
\end{aligned}
\tag{12}
$$

This expression properly reflects the requirement, as discussed above, that the optimal model should represent the best compromise between model fit (accuracy) and model complexity. We use it routinely in the context of DCM for ERPs (compare David et al. 2006).

In the case of DCM for fMRI, a complication arises. This is due to the complexity term which depends on the prior density, for example, the prior covariance of the intrinsic connections. This is problematic in the context of DCM for fMRI because the prior covariance is defined in a model-specific fashion to ensure that the probability of obtaining an unstable system is very small. (Specifically, this is achieved by choosing the prior covariance of the intrinsic coupling matrix A such that the probability of obtaining a positive Lyapunov exponent of A is $p < 0.001$; see Friston et al. 2003 for details.) Consequently, one cannot easily compare models with different numbers of connections. Therefore, alternative approximations to the model evidence are useful for DCMs of this sort.

Suitable approximations, which do not depend on the prior density, are afforded by the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), respectively. As shown by Penny et al. (2004a), for DCM these approximations are given by

$$
\begin{aligned}
BIC &= accuracy(m) - \frac{d_\theta}{2}\ln N \\
AIC &= accuracy(m) - d_\theta
\end{aligned}
\tag{13}
$$

where d_θ is the number of parameters and N is the number of data points (scans). If one compares the complexity terms of BIC and AIC, it becomes obvious that BIC pays a heavier penalty than AIC as soon as one deals with 8 or more scans (which is virtually always the case for fMRI data):

$$
\begin{aligned}
&\frac{d_\theta}{2}\ln N > d_\theta \\
&\Rightarrow N > e^2 \approx 7.39
\end{aligned}
\tag{14}
$$

Therefore, BIC will be biased towards simpler models whereas AIC will be biased towards more complex models. This can lead to disagreement between the two approximations about which model should be favoured. In DCM for fMRI, we have therefore adopted the convention that, for any pairs of models m_i and m_j to be compared, a decision is only made if AIC and BIC concur (see below); the decision is then based on that approximation which gives the smaller *Bayes factor (BF)*:

$$
BF_{ij} = \frac{p(y \mid m_i)}{p(y \mid m_j)} \tag{15}
$$

Just as conventions have developed for using p-values in frequentist statistics, there are conventions for the use of Bayes factors. For example, Raftery (1995) suggests interpretation of Bayes factors as providing weak (BF < 3), positive (3 ≤ BF < 20), strong (20 ≤ BF < 150) or very strong (BF ≥ 150) evidence for preferring one model over another.

BMS plays a central role in the application of DCM. The search for the best model, amongst several competing ones, precedes (and is often equally important to) the question which parameters of the model represent significant effects. Several studies have used BMS successfully to address complex questions about the architecture of neural systems. For example, Penny et al. (2004) investigated which connections in a system of hierarchically connected visual areas were most likely to underlie the modulatory effects of attention to motion that were observed in the BOLD responses of area V5. They found, using data from a single subject, that the best model was one in which attention enhanced V5 responses to V1 inputs. In another single-subject study, Stephan et al. (2005) systematically derived 16 different models that could have explained BOLD activity in visual areas during lateralised presentation of visual word stimuli. They found evidence that, in this subject, inter-hemispheric connections served task-dependent information transfer from the non-dominant to the dominant hemisphere – but only when the stimulus was initially received by the non-dominant hemisphere. Finally, Garrido et al. (2006) extended the previous work by David et al. (2006) and applied BMS in the context of an auditory oddball study, measured with EEG, to find the most likely explanation, in terms of coupling changes, for the well-known mismatch negativity potential. They found that their group of healthy controls was divided into two subgroups characterised by different optimal models. Re-examining the ERPs of these subgroups separately revealed a significant difference in the expression of mismatch-related responses that would have been missed in conventional ERP analyses. This example highlights that BMS may also be of considerable interest for defining clinical

populations for whom biological markers are presently lacking. This issue is taken up in the next and final section.

Outlook to future applications of DCM

DCM is currently the most advanced framework for inferring the effective connectivity in neural systems from measured functional neuroimaging data. Our hope is that over the next years, the generic framework of DCM and the ongoing developments, some of which were briefly described in this article, will contribute to a more mechanistic understanding of brain function. Of particular interest will be the use of neural system models like DCM (i) to understand the mechanisms of drugs and (ii) to develop models that can serve as diagnostic tools for diseases linked to abnormalities of connectivity and synaptic plasticity, e.g. schizophrenia.

Concerning pharmacology, many drugs used in psychiatry and neurology change synaptic transmission and thus functional coupling between neurons. Therefore, their therapeutic effects cannot be fully understood without models of drug-induced connectivity changes in particular neural systems. So far, only relatively few studies have studied pharmacologically induced changes in connectivity (e.g. Honey et al. 2003). As highlighted in a recent review by Honey & Bullmore (2004), an exciting possibility for the future is to use system models at the early stage of drug development to screen for substances that induce desired changes of connectivity in neural systems of interest with a reasonably well understood physiology. The success of this approach will partially depend on developing models that include additional levels of biological detail (e.g. effects of different neurotransmitters, see above) while being parsimonious enough to ensure mathematical identifiability and physiological interpretability; see Breakspear et

al. (2003), Harrison et al. (2005), Jirsa (2004) and Robinson et al. (2001) for examples that move in this direction.

Another important goal is to explore the utility of models of effective connectivity as diagnostic tools (Stephan 2004). This seems particularly attractive for psychiatric diseases whose phenotypes are often very heterogeneous and where a lack of focal brain pathologies points to abnormal connectivity (dysconnectivity) as the cause of the illness. Given a pathophysiological theory of a specific disease, connectivity models might allow one to define an *endophenotype* of that disease, i.e. a biological marker at intermediate levels between genome and behaviour, which enables a more precise and physiologically motivated categorization of patients (Gottesman & Gould 2003). Such an approach has received particular attention in the field of schizophrenia research where a recent focus has been on abnormal synaptic plasticity leading to dysconnectivity in neural systems concerned with emotional and perceptual learning (Friston 1998; Stephan et al. 2006). A major challenge will be to establish neural systems models which are sensitive enough that their connectivity parameters can be used reliably for diagnostic classification and treatment response prediction of individual patients. Ideally, such models should be used in conjunction with paradigms that are minimally dependent on patient compliance and are not confounded by factors like attention or performance. Given established validity and sufficient sensitivity and specificity of such a model, one could use it in analogy to biochemical tests in internal medicine, i.e. to compare a particular model parameter (or combinations thereof) against a reference distribution derived from a healthy population (Stephan 2004). Alternatively, if different clinical subgroups exhibit different "fingerprints" of dysconnectivity, each represented by a particular DCM, model selection

could provide a powerful approach to classify patients (compare the study by Garrido et al. 2006 on healthy subjects). Such procedures could help to decompose current psychiatric entities like schizophrenia into more well-defined subgroups characterized by common pathophysiological mechanisms and may facilitate the search for genetic underpinnings.

Acknowledgments

This work was supported by the Wellcome Trust.

References

Aertsen A, Preißl H (1991) Dynamics of activity and connectivity in physiological neuronal Networks. In: Schuster HG (ed.) Non Linear Dynamics and Neuronal Networks, VCH Publishers, New York, pp. 281-302

Bitan T, Booth JR, Choy J, Burman DD, Gitelman DR, Mesulam MM (2005) Shifts of effective connectivity within a language network during rhyming and spelling. Journal of Neuroscience 25, 5397-5403

Breakspear M, Terry JR, Friston KJ (2003) Modulation of excitatory synaptic coupling facilitates synchronization and complex dynamics in a biophysical model of neuronal dynamics. Network: Computation in Neural Systems 14, 703-732

Büchel C, Friston KJ (1997) Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. Cerebral Cortex 7, 768-778

David O, Friston KJ (2003) A neural mass model for MEG/EEG: coupling and neuronal dynamics. NeuroImage 20, 1743-1755

David O, Harrison LM, Friston KJ (2005) Modelling event-related responses in the brain. NeuroImage 25, 756-770

David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ (2006) Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage, in press

Friston KJ (1994) Functional and effective connectivity in neuroimaging: a synthesis. Human Brain Mapping 2, 56-78

Friston KJ (1998) The disconnection hypothesis. Schizophrenia Research 30, 115-125

Friston KJ, Mechelli A, Turner R, Price CJ (2000) Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466-477

Friston KJ (2002) Bayesian estimation of dynamical systems: An application to fMRI. NeuroImage 16, 513-530

Friston KJ (2002) Beyond phrenology: What can neuroimaging tell us abut distributed circuitry? Annual Reviews in Neuroscience 25, 221-250

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. NeuroImage 19, 1273-1302

Garrido M, Kilner JM, Kiebel SJ, Stephan KE, Friston KJ (2006) Dynamic causal modelling of the mismatch negativity. Under review.

Goebel R, Roebroeck A, Kim DS, Formisano E (2003) Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magnetic Resonance Imaging 21, 1251-1261

Harrison LM, Penny W, Friston KJ (2003) Multivariate autoregressive modeling of fMRI time series. NeuroImage 19, 1477-1491

Harrison LM, David O, Friston KJ (2005) Stochastic models of neuronal dynamics. Philosophical Transactions of the Royal Society London B Biological Sciences 360, 1075-1091

Honey GD, Suckling J, Zelaya F, Long C, Routledge C, Jackson S, Ng V, Fletcher PC, Williams SCR, BrownJ, Bullmore ET (2003) Dopaminergic drug effects on physiological connectivity in a human cortico-striato-thalamic system. Brain 126, 1767-1281

Honey G, Bullmore E (2004) Human pharmacological MRI. Trends in Pharmacological Sciences 25, 366-374

Jirsa VK (2004) Connectivity and dynamics of neural information processing. Neuroinformatics 2, 183-204

Kiebel SJ, David O, Friston KJ (2006) Dynamic causal modelling of evoked responses in EEG/MEG with lead-field parameterization. NeuroImage, in press

Mechelli A, Price CJ, Noppeney U, Friston KJ (2003) A dynamic causal modeling study on category effects: bottom–up or top–down mediation? Journal of Cognitive Neuroscience 15, 925–934

McIntosh AR, Gonzalez-Lima F (1994) Structural equation modeling and its application to network analysis in functional brain imaging. Human Brain Mapping 2, 2-22

Penny WD, Stephan KE, Mechelli A, Friston KJ (2004a) Comparing dynamic causal models. NeuroImage 22, 1157-1172

Penny WD, Stephan KE, Mechelli A, Friston KJ (2004b) Modeling functional integration: a comparison of structural equation and dynamic causal models. NeuroImage 23, S264-274

Pitt MA, Myung IJ (2002) When a good fit can be bad. Trends in Cognitive Neuroscience 6, 421-425

Robinson PA, Rennie CJ, Wright JJ, Bahramali H, Gordon E, Rowe DL (2001) Prediction of electroencephalographic spectra from neurophysiology. Physical Reviews E 63, 021903

Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annual Reviews of Neuroscience 23, 473-500.

Smith APR, Stephan KE, Rugg MD & Dolan RJ (2006) Task and content modulate amygdala-hippocampal connectivity in emotional retrieval. Neuron 49: 631-638

Stephan KE (2004) On the role of general system theory for functional neuroimaging. Journal of Anatomy 205, 443-470

Stephan KE, Penny WD, Marshall JC, Fink GR, Friston KJ (2005) Investigating the functional role of callosal connections with dynamic causal models. Annals of the New York Academy of Sciences 1064:16-36

Stephan KE, Baldeweg T, Friston KJ (2006) Synaptic plasticity and dysconnection in schizophrenia. Biological Psychiatry, in press.

Figure legends

Figure 1
A. The bilinear state equation of DCM for fMRI. B. An example of a DCM describing the dynamics in a hierarchical system of visual areas. This system consists of areas V1 and V5 and the superior parietal cortex (SPC). Each area is represented by a single state variable ($x_1 \ldots x_3$). Black arrows represent connections, grey arrows represent external inputs into the system and thin dotted arrows indicate the transformation from neural states into haemodynamic observations (thin boxes; see Fig. 2 for the haemodynamic forward model). In this example, visual stimuli (photic) drive activity in V1 which is propagated to V5 and SPC through the connections between the areas. The V1→V5 connection is allowed to change whenever the visual stimuli are moving, and the SPC→V5 connection is modulated whenever attention is directed to motion. The state equation for this particular example is shown on the right.

Figure 2
Summary of the haemodynamic model used by DCM for fMRI. Neuronal activity induces a vasodilatory and activity-dependent signal s that increases blood flow f. Blood flow causes changes in volume and deoxyhaemoglobin (v and q). These two haemodynamic states enter the output nonlinearity which results in a predicted BOLD response y. Adapted, with permission by Elsevier Ltd., from Friston et al. (2003).

Figure 3
A. DCM applied to data from a study on attention to visual motion by Büchel & Friston (1997). The model is similar to the one shown in Fig. 1, except for the addition of another area, the inferior frontal gyrus (IFG). The most interesting aspect of this model

concerns the role of motion and attention in exerting bilinear effects on connections in the model. The presence of motion in the visual stimulation enhances the connection from area V1 to the motion sensitive area V5. The influence of attention is to enable backward connections from the IFG to the superior parietal cortex (SPC) from SPC to V5. Dotted arrows connecting regions represent significant bilinear affects in the absence of a significant intrinsic coupling. Inhibitory self-connections are not displayed for clarity. B. Fitted responses based upon the conditional estimates and the adjusted data. The insert shows the approximate location of the regions. Adapted, with permission by Elsevier Ltd., from Friston et al. (2003).

Figure 4

A. Summary of a fictitious 2x2 factorial experimental design, comprising task and stimulus factors. B. Simulated BOLD responses of two areas, y_1 and y_2 (without observation noise). The first area shows a main effect of stimulus and the second area additionally shows a stimulus-by-task interaction. The red and green bars denote when task 1 and task 2 are performed, respectively. C. The DCM which was used to generate the noise-free responses shown in panel B. As shown schematically, all inputs were box-car functions. +++ denotes strongly positive and + denotes weakly positive inputs and connection strengths, – denotes negative connection strengths. The different strengths of the driving inputs induce a main effect of stimulus in the first area, x_1. This effect is conveyed onto the second area, x_2, by means of the $x_1 \rightarrow x_2$ connection. Critically, the strength of this connection varies as a function of which task is performed. This bilinear modulation induces a stimulus-by-task interaction in x_2 (c.f. panel B). D. Data generated

from the model shown in panel C but with additional observation noise (signal-to-noise ratio of unity).

Figure 5

A. Example of a DCM with second-order terms in the state equation. In this example, the third area modulates the connection from the first to the second area. The first area is driven by two different stimuli (stim1, stim2; randomly mixed events, represented as delta functions, 4 seconds apart) and the third area is driven by some inputs representing cognitive set (mod1, mod2; alternating blocks of 15 s duration, shown as grey boxes in Fig. 5B). Note that the third area is not only driven by external input but also receives an input from the second area. +++ denotes strongly positive and + denotes weakly positive inputs and connection strengths, – denotes negative connection strengths. B. Simulated responses of this system (note that all inputs and connections were given positive weights in this simulation). From top to bottom, the plots show the neural (x) and haemodynamic (y) responses in alternating fashion. The x-axis denotes time (for haemodynamic responses in seconds, for neural responses in time bins of 4 ms), the y-axis denotes arbitrary units. It can be seen easily that evoked activity in the first area only causes a significant response in the second area if the third area shows a high level of activity and thus enables the $x_1 \rightarrow x_2$ connection. Furthermore, due to the excitatory $x_2 \rightarrow x_3$ connection, a positive reinforcement effect results. Both mechanisms lead to obvious nonlinearities in the generated data (see thick arrows for an example). Note that this model, similar the one in Fig. 4, also generates a "stim × mod" interaction in the second area. This is harder to see by eye than in Fig. 4 because here the driving inputs are randomly mixed events and additionally, strong non-linear effects occur.

Figure 6
A schema of the neural populations which are modelled separately for each region in DCM for ERPs. Different regions are coupled by forward, backward and lateral connections, all of which originate from excitatory pyramidal cells but target specific populations. The figure shows a typical hierarchical network composed of three cortical areas. Extrinsic inputs evoke transient perturbations around the resting state by acting on a subset of sources, usually the lowest in the hierarchy. Reproduced with permission by Elsevier Ltd. from David et al. (2006).

Figure 7
Schematic of the neural model in DCM for ERPs. This schema shows the state equations describing the dynamics of a microcircuit representing an individual region (source). Each region contains three subpopulations (pyramidal, spiny stellate and inhibitory interneurons) that are linked by intrinsic connections and have been assigned to supragranular, granular and infragranular cortical layers. Different regions are coupled.through extrinsic (long-range) excitatory connections as described in Fig. 7. Reproduced with permission by Elsevier Ltd. from David et al. (2006).

Figure 8
DCM for ERPs measured during an auditory oddball paradigm. Left: Predicted (thick) and observed (thin) responses in measurement space. These are a projection of the scalp or channel data onto the first three spatial modes or eigenvectors of the channel data. The predicted responses are based on the conditional expectations of the DCM parameters and show very good agreement with the measured data. Right: Graph depicting the sources and connections of a DCM in which both forward and backward connections were allowed to change between odddball and standard trials. The relative strength of

coupling strengths for oddball relative to standard stimuli are shown alongside the connections. The percent conditional confidence that this difference is greater than zero is shown in brackets. Only changes with 90% confidence or more (solid lines) are shown numerically. In all connections the coupling was stronger during oddball processing, relative to standards. A1: primary auditory cortex, OF: orbitofrontal cortex, PC: posterior cingulate cortex, STG: superior temporal gyrus. Reproduced with permission by Elsevier Ltd. from David et al. (2006).