J. Anat. (2004) **205**, pp443–470

REVIEW

On the role of general system theory for functional neuroimaging

Klaas Enno Stephan[1,2]

[1]*The Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, UK*
[2]*Department of Psychology, School of Biology, Henry Wellcome Building, University of Newcastle upon Tyne, UK*

Abstract

One of the most important goals of neuroscience is to establish precise structure–function relationships in the brain. Since the 19th century, a major scientific endeavour has been to associate structurally distinct cortical regions with specific cognitive functions. This was traditionally accomplished by correlating microstructurally defined areas with lesion sites found in patients with specific neuropsychological symptoms. Modern neuroimaging techniques with high spatial resolution have promised an alternative approach, enabling non-invasive measurements of regionally specific changes of brain activity that are correlated with certain components of a cognitive process. Reviewing classic approaches towards brain structure–function relationships that are based on correlational approaches, this article argues that these approaches are not sufficient to provide an understanding of the operational principles of a dynamic system such as the brain but must be complemented by models based on general system theory. These models reflect the connectional structure of the system under investigation and emphasize context-dependent couplings between the system elements in terms of effective connectivity. The usefulness of system models whose parameters are fitted to measured functional imaging data for testing hypotheses about structure–function relationships in the brain and their potential for clinical applications is demonstrated by several empirical examples.
Key words connectivity; fMRI; structure–function relationships; systems theory.

1. Introduction

This review article is an attempt to discuss a traditional goal of neuroscience, the characterization of the relation between structure and function in the brain, from the perspective of general system theory (von Bertalanffy, 1969). The article starts with an overview of causal and correlative approaches in neuroscience towards the investigation of structure–function relationships (SFRs) in neural systems. Introducing a few simple concepts from general system theory, some formal implications for the investigation of SFRs in neural systems are derived. These implications are then evaluated in the context of functional neuroimaging. I will argue that classic applications of functional neuroimaging are insufficient to provide insights into SFRs and need to be complemented by principled models of neural systems that properly reflect the connectional structure of the system as well as the bridging principles from structure to function. One of the most useful ways of expressing these bridging principles is in terms of effective connectivity. Several models of effective connectivity are introduced and their strengths and limitations are discussed.

Many of the ideas expressed in this article are not novel and have been expressed in similar ways before (e.g. Horwitz et al. 1999; McIntosh, 2000; Friston, 2002). What this article hopes to contribute, however, is a generic perspective on models of SFRs in neural systems that is derived from basic principles of general system theory. A further aim of this article is to lend support to the current transformation of neuroimaging from a field using exploratory analyses and data-driven interpretations of the results to a hypothesis-led, model-based discipline that gradually merges with computational neuroscience in order to provide mathematical descriptions of SFRs in the brain.

Correspondence
Dr Klaas Enno Stephan, The Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK. E: k.stephan@fil.ion.ucl.ac.uk

Accepted for publication *11 October 2004*

Although I believe that neural systems cannot be understood without formal mathematical models, I have tried to keep the mathematical descriptions simple, in the hope that those neuroscientists who have not had much exposure to mathematical models of neural systems will find the material accessible. All models discussed here are essentially linear models at the level of larger brain regions (e.g. cortical areas) and do not require a sophisticated knowledge of mathematics to understand them. Furthermore, to present general concepts in a tutorial style, I have expanded on some issues that may appear unnecessarily detailed for readers with experience in system analysis. The latter readers are referred to mathematically more advanced texts on neural system modelling as found, for example, in Friston (2003), Jirsa (2004) or Dayan & Abott (2001).

2. Causal and correlative SFRs

One of the classic goals of neuroscience is to describe SFRs. There is a wide range of well-known examples from different organizational levels of the brain that can be found in any standard textbook on neuroscience. For example, some general functional properties of neurons can be directly explained from the molecular structure of certain ion channels, e.g. the absolute refractory period is a direct consequence of the molecular structure of the sodium channel. The functional differences between magno- and parvocellular neurons throughout the visual system are partially dependent on the different geometry of their dendritic trees, and the topology of fibre systems is sufficient to explain some basic neurophysiological findings such as the contralateral cortical representation of a peripherally presented visual stimulus as well as some clinically rather complex syndromes like the Brown–Séquard syndrome.

These examples are chosen more or less arbitrarily and could be replaced by many others. What is common to all of them is that the formulated SFR expresses a direct causal role of structure for function: 'The brain component C has the functional property F because of its structural property S.' However, many questions in neuroscience are not easily addressed in this fashion. For example, at the level of cortical areas analogous causal definitions of SFRs have proven to be much more difficult. This is not simply due to a lack of knowledge: for many cortical areas, we have an exquisite understanding of the anatomical microstructure and have observed its functional responses under many different combinations of sensory stimulation and cognitive context. Yet, there does not seem to be a single area for which we are able to deduce its functional properties in a direct and causal fashion from its microstructural properties. One obvious explanation for this is the increase in complexity. First, many areas appear to be involved in more than one cognitive function (at least at the level of psychological nomenclature). This has been observed for low-level areas such as V1, which takes part in very different aspects of visual information processing (Lee, 2003), as well as for high-level areas such as Broca's area in the left inferior frontal gyrus (IFG), which has been shown to be involved in functional contexts as diverse as language processing, action observation and local visual search (Hamzei et al. 2003; Manjaly et al. 2003, 2004). Second, in order to explain the observed functional range, we may need to take into account several microstructural variables at once (e.g. neuronal cell types, cyto- and myeloarchitecture, and receptor distributions). Additionally, we may need to consider potential interactions between these structural variables, e.g. the way in which neuronal cell type determines areal function is likely to depend on the intrinsic microcircuitry of the area (Lund, 2002). In other words, determining causal SFRs for cortical areas is a multivariate problem that requires a model of the interactions between the structural variables. Third, functional responses in cortical areas are highly context-sensitive, e.g. they depend on the previous processing history as well as on the nature of the inputs provided by other brain regions (Passingham et al. 2002). For example, the responses of neurons in many visual areas can be drastically altered by changes in cognitive set or attention (Luck et al. 1997; Li et al. 2004). Therefore, any attempt to explain SFRs in cortical areas must be able to account for such context dependencies, which are observed ubiquitously (Albright & Stoner, 2002). Finally, and most importantly, because no cortical area operates in isolation but is connected to a large number of other areas by anatomical long-range connections (so-called association fibres), the functional behaviour of a particular area cannot be explained by its local microstructure only. Indeed, strong changes of the neural responses in various areas have been reported after a particular input from a remote area was experimentally abolished (Hupe et al. 1998) or enhanced (Moore & Armstrong, 2003). Therefore, the structure of the connectional pattern with other areas has to be taken into account when formulating a hypothesis on the SFR of a given area.

In general terms, for any given component of the brain the definition of causal SFRs becomes more difficult (1) the more complex the structure of this component, (2) the more complex its functional range and, most importantly, (3) the less isolatable and context-independent it is (i.e. the more interactions it has with other components). In other words, how easily causal SFRs can be established depends on whether one needs to adopt an explicit systems perspective. This issue is one of the core problems of the general scientific inquiry (von Bertalanffy, 1969) and will be addressed in more detail below. In the case of cortical areas, as demonstrated above, a systems perspective appears mandatory for unravelling causal SFRs because cortical areas not only have a complex internal structure and subserve complex functions that are highly context-dependent, but are also densely connected among each other (and with subcortical structures) through association fibres.

Historically, the difficulties in establishing causal SFRs for cortical areas have had considerable consequences. In cognitive neuroscience, the mechanistic view that underlies SFRs in the strict sense has largely been exchanged for a black box perspective where the aim is merely to state which areas (defined by intrinsic structural homogeneity in terms of neuron types, microcircuitry and external connections) are consistently observed to be involved in a certain functional context. In other words, major parts of neuroscience have been aiming at the more modest goal of merely establishing correlations between structure and function. Since the 19th century, much interdisciplinary work has been devoted to establishing such structure–function correlations (SFCs) for cortical areas. This required (1) a parcellation of the cortex into distinct areas and (2) methods for measuring the involvement of these entities in a given function. The structural basis of this endeavour was (and still is) delivered by neuroanatomy in the form of cortical parcellation schemes that are based on microstructural critera, using cyto-, myelo- and/or receptor-architectonics (e.g. Brodmann, 1909; Vogt & Vogt, 1919; von Bonin & Bailey, 1947; Zilles et al. 2002). Modern atlases provide probabilistic information about the spatial location of cortical areas in reference to a population of parcellated brains (Amunts et al. 2000). The methods for establishing the involvement of a given area in a certain cognitive function have traditionally been provided both by neurophysiology (e.g. using invasive recordings from animals) and by neuropsychology (which explores cognitive deficits after lesions to one or several areas).

Both neurophysiological and neuropsychological techniques for exploring the functional role of a given area do, however, have severe limitations. For example, invasive recordings, with the exception of a very special and small population of patients, are ethically restricted to animals. Furthermore, they are methodologically constrained in that they usually only allow one to assess a small patch of cortex, and usually only test for very few functions. Neuropsychological studies of brain lesions also suffer from major problems of interpretation. First, brain lesions are rarely confined to a single area but often spread across large parts of the cortex and can also affect fibre tracts in the white matter. Second, the brain is extraordinarily plastic, and the occurrence of compensatory mechanisms can render the relation between a spatially specific lesion and loss of function opaque. Third, given that cortical areas are densely interconnected with each other, lesioning of areas can lead to widespread and complex effects in the cortical network. A striking example is given by paradoxical lesion effects in which a cognitive function that was compromised after a first lesion is largely restored after a second lesion (Sprague, 1966; Lomber et al. 2002). Experimental lesion studies in animals and theoretical models have demonstrated that a correct interpretation of the functional consequences of lesions requires knowledge about the connectivity of the lesioned area (Payne et al. 1996; Young et al. 2000).

About 20 years ago, positron emission tomography (PET) became available as a new method to determine SFCs, followed by functional magnetic resonance imaging (fMRI) in the early 1990s. By measuring changes of regional cerebral blood flow (rCBF) and blood oxygen-level-dependent (BOLD) signals, respectively, PET and fMRI offer non-invasive, whole-brain, high-resolution measurements of regionally specific changes of brain activity that are correlated with certain components of a cognitive task. Therefore, these techniques promised to revolutionize the search for SFCs as they overcome many of the problems associated with invasive recordings and lesion studies discussed above. Indeed, since their introduction the number of SFCs described for cortical areas has exploded. Today, at least one functional label seems to have been proposed for each cortical region. There is no doubt that the use of the correlative approach has generated a lot of useful information about which areas are potential elements of the neural systems for implementing particular cognitive processes. Currently, however, this approach

appears to have reached saturation point. There are two main reasons for this. The first is an increasing tension between the implicit tendency towards localizationist interpretations of neuroimaging results, and the diversity of findings that appear to contradict the idea of one-to-one relations between specific cortical areas and specific cognitive functions. Secondly, and more importantly, although ever longer lists of observed correlations between structural entities and cognitive processes are being produced, there is only very modest, if any, progress in our understanding of the causal mechanisms that underlie these correlations.

In this article I argue that, in order to provide us with a deeper understanding of SFRs in the brain, functional neuroimaging will need to adopt an explicit systems perspective, using causal models of brain function that are based on neuroanatomical information about the structure of the investigated system, particularly with regard to the connectivity between areas. First I briefly review general system theory and its importance for biological questions, focusing on how its principles can be applied to neuroscientific questions. Then I summarize the current conceptual and methodological foundations of neuroimaging and explain their relation to systems theory. I distinguish 'functional specialization' approaches from those that emphasize the role of causal interactions between separate areas, i.e. models of effective connectivity. Finally I discuss several neuroimaging studies, where the usefulness of system models based on effective connectivity becomes particularly evident.

3. General system theory

3.1. The significance of general system theory for scientific investigations

The central goal of most scientific disciplines is to understand systems, i.e. ensembles of interacting elements. Today, this statement sounds almost trivial, yet the scientific focus on the systems concept has been established only relatively recently. Ludwig von Bertalanffy, a German–Canadian biologist and philosopher, wrote some seminal papers in the 1920s in which he argued that most complex scientific phenomena could only be understood properly if one found a mathematical description of how their behaviour as a whole emerged from the interactions of their parts. He suggested a very general framework for describing and analysing systems and demonstrated the existence of system isomorphisms, i.e. the existence of general mathematical descriptions that explained the dynamic behaviour of very different kinds of systems at different scales and across fields as diverse as physics, biology, economy and sociology. Although this work formed the foundation of what became eventually known as general system theory (see the collection of his early essays in von Bertalanffy, 1969), it remained unpublished for almost two decades. After the first papers had appeared in the 1940s, the systems concept experienced a scientific breakthrough, supported by the rise of cybernetics, 'the science of control and communication in the animal and the machine' (Wiener, 1948), which was introduced by Norbert Wiener (1948) and advanced by Ross Ashby (1956).

Today, biology uses the systems concept to address questions at all levels of resolution: molecular (e.g. the interactions between different genes mediated by the proteins they encode), cellular (e.g. the functional integration of different populations of neurons), within a given organ (e.g. the instantiation of cognitive functions by the interaction of different cortical areas), between different organs (e.g. endocrine mechanisms of regulation between hypothalamus, hypophyseal gland and peripheral glands) and between entire organisms (e.g. in ecology or population biology). The omnipresence of the systems concept in biology and most other sciences is so strong that a recent special issue of the journal *Science* on 'Systems Biology' confirmed von Bertalanffy's (1969) previous diagnosis: 'The [systems] concept has pervaded all fields of science and penetrated into popular thinking, jargon, and mass media' (Chong & Ray, 2002).

But what exactly is needed to speak of a 'system' and why is the systems concept so useful for framing scientific questions? A general, yet informal, definition is that a system is a set of elements that interact with each other in a spatially and temporally specific fashion. Before we attempt a more formal definition, let us remind ourselves that one of the classic principles of scientific inference is to 'analyse' a given phenomenon, i.e. to break it down into atomic units and processes that can be investigated independently of each other. This approach is appealing because it reduces a complex problem to a set of simpler problems, each of which can be addressed under conditions where it is easier to control for potentially confounding influences. For example, the kinetics of a biochemical process mediated by a certain enzyme can be studied in

isolation, measuring the rate of transformation under different conditions. Having studied a set of different biochemical processes separately in this fashion, however, one would still not be able to predict quantitatively what their collective dynamics is like when these processes happen simultaneously in a shared environment, e.g. within a living organism.

This uncertainty is due to the fact that the different processes may interact, e.g. one process may change the substrate/product ratio of another process, or the efficacy of an enzyme that is relevant for a particular process may change due to the presence of allosteric (in)activators that are produced by a second process or due to dynamic changes in gene expression mediated by a third process. In a similar fashion, isolating a neuron from the system in which it naturally participates (e.g. by growing it in a dish) allows one to measure its response spectrum to experimentally controlled inputs. However, this response spectrum may look very different if observed in the real system depending on the temporal structure of inputs from other neurons, presence of modulatory transmitters, metabolic interactions with glial cells, etc. As a third and different example, but which converges onto the same kind of problem, many principles of thermodynamics in physics are explicitly restricted to an isolated (autonomous) system, i.e. a closed ensemble of elements that is not perturbed by any kind of structured input from its outside. For an isolated system, the second principle of thermodynamics states that over time entropy will increase to a maximum. This precludes the existence of ordered structure in the system (e.g. with regard to the spatial distribution of the system elements). However, most natural phenomena show a remarkable degree of order and organization. This is because they result from open (non-autonomous) systems that receive temporally structured inputs from their environment. For example, the spatio-temporal structure of brain activity is partially dependent on inputs from the external world that enter the brain through sensory interfaces.

In summary, the general problem of analytical procedures in science is that they usually do not allow one to reconstruct the behaviour of the whole system because, on their own, they are blind to predicting the consequences arising from interactions between the atomic elements and processes studied in isolation. As a consequence, analytical procedures need to be complemented with a theoretical framework that can be used to understand and predict the dynamics of the system as a whole. This framework is provided by general system theory.

3.2. A formal perspective on systems and their SFRs

As explained above, statements on causal SFRs are usually quite difficult to derive if the phenomenon of interest must be investigated from a systems perspective. In practice, the necessity to adopt a formal systems perspective is often ignored by neuroimaging studies and replaced by hand-waving statements on SFRs. Most commonly, SFRs are simply replaced by SFCs, for example in morphometric studies that correlate the volume of a certain brain region with behavioural indices or, even more commonly, in functional neuroimaging by correlating a specific cognitive process with the co-activation of a 'network' of areas. At best, the observed co-activation pattern is interpreted informally by referring to the structural connections between the regions as inferred from tract tracing experiments in the monkey. As demonstrated by the seminal analyses of Felleman & Van Essen (1991) and Young (1992), however, individual connections are not sufficient to understand structural or functional properties of a given system without formally analysing its entire connectivity pattern.

If one does not try to ignore the necessity for formal system analyses in neuroscience, but embraces this perspective, powerful new insights on SFRs become possible. First of all, as will be demonstrated below, a formal definition of a system allows one to pinpoint, in conceptual and mathematical terms, what is meant precisely by structure, function and SFR. Second, it is the only way to express the SFR in quantitative terms such that predictions become possible for situations in which the system has not been observed before. Third, it is the only way to understand fully how a system works; this is a necessity to investigate in an informed manner how system function could be restored if some of its components are rendered dysfunctional, e.g. by disease (Payne & Lomber, 2001).

Informally, as mentioned above, a system is generally defined as a set of elements that interact with each other in a spatially and temporally specific fashion. Structure refers to all static, i.e. time-invariant, components and relations of a given system. In analogy, function refers to all those dynamic, i.e. time-variant, components and relations of the system that are conditional on structure. From that it follows that the SFR is defined by the nature of this conditionality.

In the remainder of this section, I describe how these informal definitions can be given a mathematical form. Readers lacking a mathematical background should move to section 4.2.

As a mathematical framework, a set of differential equations with time-invariant parameters is chosen; this formulation follows the early proposal by von Bertalanffy (1950) regarding how systems could generally be formalized, and can be easily extended to cover a whole range of special cases. However, differential equations are not the only possible mathematical representation of dynamic systems though. There are multiple alternatives, including iterative maps and cellular automata to name just two options (see Bar-Yam, 1997). The underlying concept, however, is always the same: a system can be defined by a set of n elements that have 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:

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

If we take a neural system in the brain as an example, say an ensemble of interacting neurons, 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 a variety of neurophysiologically meaningful indices, e.g. the membrane potential in different compartments of the neuron or the status of ion channels at its synaptic sites. This touches on an important distinction: in system construction (e.g. in engineering), the state variables and their mutual dependencies are usually known; in system identification (e.g. when trying to understand a biological system), however, they are not known. This means that we always require a model of the system that represents our current hypothesis of system structure and function. This point will become important later on when we address applications of system-based approaches to functional neuroimaging.

As mentioned above, the crucial point is that the state variables interact with each other, i.e. the change of any state variable depends on the value of at least one other state variable. This mutual functional dependence between the properties of the elements in the system is expressed in a very natural fashion by a set of ordinary differential equations:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, \ldots, x_n) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n(x_1, x_2, \ldots, x_n). \end{aligned} \tag{2}$$

Rewriting Eq. (2) as a function of the state vector leads to the compact statement that the change in the system's state depends on its current state:

$$\frac{dx}{dt} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = F(x). \tag{3}$$

However, this description is not yet sufficient. First of all, the specific form of the dependencies f_i needs to be specified, which requires a set of parameters θ and, second, in the case of no-nautonomous systems (which are those of interest to biology and neuroscience) 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 $u(t)$. Altogether, this leads to a general state equation for non-autonomous systems:

$$\frac{dx}{dt} = \begin{bmatrix} f_1(x_1, \ldots, x_n, u, \theta_1) \\ \vdots \\ f_n(x_1, \ldots, x_n, u, \theta_n) \end{bmatrix} = F(x, u, \theta). \tag{4}$$

where $\theta_1, \ldots, \theta_n$ are the parameter vectors of the individual dependencies f_i, and θ is the overall (concatenated) parameter vector of the system. Such a model provides a causal description of how system dynamics results from system structure, because (1) it describes when and where external inputs enter the system, and (2) how the state changes induced by these inputs evolve in time depending on the system's structure, i.e. its connectivity pattern and any time-invariant property of the system elements and the connections between them (e.g. time constants).

It is important to note that I have made several assumptions to simplify the exposition. First, the description above assumes that all processes in the system are deterministic, i.e. the equations do not account for random processes (noise). Second, we assume that we know the inputs that enter the system. In neuroimaging, this is a tenable assumption because the inputs are experimentally controlled variables such as changes in stimuli or instructions. Third, the inputs to the system

are assumed to be independent and not to interact. In the case of interacting inputs, $u(t)$ itself could be expressed as a set of differential equations in analogy to Eq. (2). Fourth, we have neglected the possibility that changes in system state may depend on its recent history; see Friston (2000a) for an elegant model of general brain function that incorporates this mechanism in the form of 'neuronal transients'. Fifth, and most importantly, we assume that both the mathematical form of the dependencies f_i and the parameters θ are time-invariant. This assumption is valid for systems whose structure does not change during the time of observation.

On the basis of the general system description provided by Eq. (4) we are now in a position to state more accurately what we mean by structure, function and SFRs in a system, or more precisely, in a model of a system:

- Structure is defined by the time-invariant components of the system model, i.e. θ and the mathematical form of the state variable dependencies f_i.
- Function refers to those time-variant components of the system model that are conditional on its structure, i.e. $x(t)$, but not $u(t)$.
- The SFR is represented by F: its integration describes how system dynamics results from system structure. More specifically, integrating F in time determines the temporal evolution of the system state x from the onset of an input $u(0)$ (i.e. at time $t = 0$) up to a time point τ, given a known or assumed initial state $x(0)$ (see Bossel, 1992, pp. 95, 397):

$$x(\tau) = x(0) + \int_0^\tau F(x, u, \theta)dt. \tag{5}$$

In other words, once the system structure (i.e. θ and the form of f_i) is specified and a particular temporal sequence of inputs $u(t)$ is chosen, Eq. (5) provides a complete description of how the functional behaviour of the system (i.e. its dynamics, the trajectory of the state vector x in time) results from its structure and initial state. Notably, the system structure determines both intrinsically sustained dynamics in the absence of inputs and dynamics enforced by external inputs. Without going into details, it should be mentioned that there exists an approximation to Eq. (5) by means of Volterra series that has proven very useful for practical applications to neural systems (Rieke et al. 1997; Friston & Büchel, 2000; Friston et al. 2000, 2003).

All the equations presented so far are extremely general, and F, representing the SFR of the system, could be an arbitrarily complex non-linear function. To illustrate the definitions of structure, function and SFR in more detail, we discuss the case of a system with a linear SFR. Although most natural phenomena are of a non-linear nature, linear system models play an outstanding role in systems science because (1) they are analytically tractable, and (2) given sufficiently long observation periods and non-negligible external input, their dynamics is largely independent of the initial state (Bossel, 1992, p. 386). Therefore, non-linear systems are often investigated in restricted subspaces of interest, using linear models as local approximations. The following model is a prototypical description of a non-autonomous system in which the dynamics can be separated into a linear intrinsic component (the interactions between its n elements) and a linear extrinsic component (m external inputs):

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \vdots \\ \frac{dx_n}{dt} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nm} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}. \tag{6}$$

As Eq. (6) shows, in this system model the change of any given element depends on the state of all other elements in the system and on external inputs that affect it directly or indirectly through connected elements. The SFR of this system can be written in compact matrix form as

$$F(x) = \frac{dx}{dt} = Ax + Cu \tag{7}$$

where the non-zero values of A and C represent the parameters of the system (i.e. θ in Eq. 4) and the functional behaviour of the system at time point τ can be obtained by integration (compare Eq. 5):

$$x(\tau) = e^{A\tau}x(0) + \int_0^\tau e^{A(\tau - t)}Cu(t)dt \tag{8}$$

where e^{At} is the matrix exponential (see Bossel, 1992, pp. 364, 377).

In this model, the system's behaviour has two separable components: intrinsically sustained dynamics (parameter matrix A) and dynamics enforced by external inputs (parameter matrix C). In terms of the general system equation (Eq. 4), this corresponds to $\theta = \{A, C\}$. The first term of Eq. (7) says that the change of the state variable x_i is a linear mixture of all state variables in the system, weighted by the parameters a_{ij}. Importantly,

by defining a particular parameter a_{ij} to be zero, we disallow for any direct effect of x_j on x_i. Conversely, any non-zero parameter a_{ij} represents a causal influence of the dynamics of x_j on that of x_i. The binarized parameter matrix A

$$\tilde{A} = \chi(A) = \begin{bmatrix} \chi(a_{11}) & \cdots & \chi(a_{1n}) \\ \vdots & \ddots & \vdots \\ \chi(a_{n1}) & \cdots & \chi(a_{nn}) \end{bmatrix}, \quad \chi(a) = \begin{cases} 1 \text{ if } a \neq 0 \\ 0 \text{ if } a = 0 \end{cases} \qquad (9)$$

represents the structural connectivity of the system model. The values of A itself correspond to the effective connectivity within the system, i.e. the influence that the system elements exert over another (Friston, 1995). Finally, the values of the matrix C in the second term of Eq. (7) represent the magnitude of the direct effects that external inputs (e.g. sensory information) have on particular system elements. In particular, by setting a particular parameter c_{ij} to be zero, we disallow for any direct effect of the external input u_j on x_i (see Fig. 1 for a concrete example).

This classical model of a linear non-autonomous system with time-invariant parameters has found widespread application in various scientific disciplines (von Bertalanffy, 1969). Natural phenomena that can be described by this kind of system include, for example, fill and depletion processes of biological storages, exponential growth and decay, and oscillatory processes (Bossel, 1992). In section 6.4, we will see that dynamic causal modelling (DCM, Friston et al. 2003) extends the above formulation by bilinear terms that model context-dependencies of intrinsic connection strengths. In this paragraph, the variable names have deliberately been kept similar to those in DCM in order to facilitate the comparison (see Eq. 17). Finally, it should be noted that the framework outlined here is concerned with dynamic systems in continuous time and thus uses differential equations. The same basic ideas, i.e. that the evolution of a system's state is shaped by intrinsic interactions between system elements and external input, can also be applied to dynamic systems in discrete time (using difference equations), as well as to 'static' systems in which the system is at equilibirum at each point of observation. The latter perspective, which is based on regression-like equations, is used by classic system models for functional neuroimaging data, e.g. psychophysiological interactions (PPI; Friston et al. 1997), structural equation modelling (SEM; McIntosh et al. 1994; Büchel & Friston, 1997) or multivariate autoregressive models (MAR; Harrison et al. 2003; Goebel et al. 2003). These will be described in section 6 and juxtaposed to DCM.

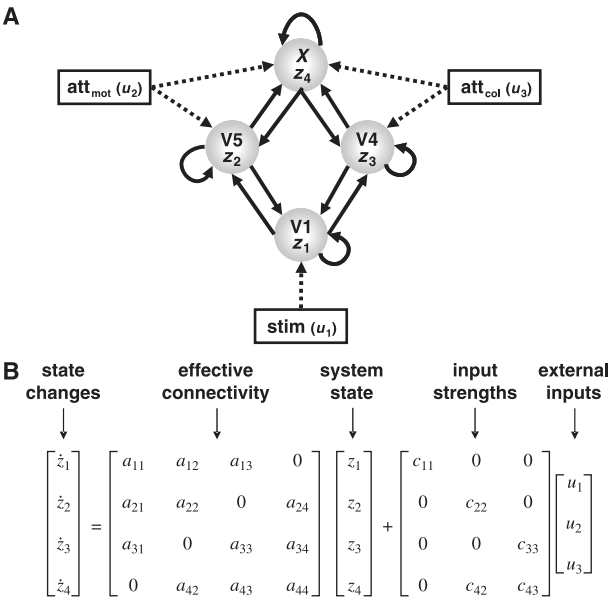


Fig. 1 (A) Concrete example of a dynamic linear model of a non-autonomous system. This model is inspired by the work of Chawla et al. (1999) in whose fMRI study volunteers attended selectively to either motion or colour of a visual stimulus. Chawla et al. (1999) found evidence for a modality-specific gain control effect: attention to motion increased the amplitude of V5 BOLD responses to stimuli whereas attention to colour did the same for V4 responses. This figure shows a dynamic linear model of the neural system underlying the attentional effects observed by Chawla et al. (1999). External inputs are represented by dotted arrows and structural connections are represented by solid arrows. Visual stimuli enter the system through primary visual cortex (V1) which is connected to both V4 and V5. Attention to colour (u_3) and attention to motion (u_2) are modelled to have direct effects on V4 and V5, respectively, as well as on an additional 'higher' area X (e.g. in parietal or prefrontal cortex) that is reciprocally connected with V4 and V5. Note that this model could replicate attention-induced signal increases in V4 and V5, both through direct and indirect (via the backward connections from X) effects. It could not, however, distinguish between gain control effects (increased responses to stimuli) and baseline shifts (increased signal during expectation of stimuli that have not yet appeared). (B) The complete state equation of the model (compare Eqs 6 and 7 in the main text). In order to save space, $\frac{dz_i}{dt}$ has been written as $\dot{z}_i$. Note that self-connections have been modelled for each area (diagonal entries in matrix A). In the absence of negative inputs, this allows the system to model the decay of induced activity.

3.3. Practical implications for neuroimaging

These general concepts have practical implications for neuroimaging because they imply what methodological

steps are required to characterize and understand SFRs in a given neural system. At the very minimum, identification of a neural system consists of at least the following steps.

1. Identification of candidate elements of the system. The choice of necessary system elements is usually based on previous results from analytical procedures. In neuroscience, potential system elements were traditionally identified by means of lesion studies or invasive recordings in animals, combined with microstructural investigations. With the availability of fMRI, conventional analyses using a General Linear Model (GLM) are ideal to inform this choice (see below).

2. Choice of the state variables. The second step is to determine the minimal set of state variables per system element that is needed to model the overall function of the system properly. For example, if one wants to model the dynamics in an ensemble of cortical areas, a choice has to be made regarding how each individual area is represented: in some cases it might be sufficient to model each area by a single state variable representing the mean activity of its entire neuronal population (e.g. Friston et al. 2003), whereas in other cases it might be necessary to use multiple state variables per area, which represent, for example, different layers, columns and neuron types (e.g. excitatory pyramidal cells and inhibitory interneurons; see David & Friston, 2003, for an example). Implicitly, this decision thus concerns the resolution at which the system is investigated. Together with the identification of system elements in step one, the choice of state variables determines the size and semantics of the state vector x.

3. Definition of a structural model and the assumed SFR. This requires us to define the assumed connectional structure of the system (see the example in Fig. 1) and the mathematical form of the interelement dependencies f_i. This step is crucial as it represents the hypothesis of how the functional behaviour of the system depends on its structure. It is obvious that the quality of the structural model depends on how well the structural connectivity is known for the particular neural system of interest.

4. Choice of priors on the parameters. System models differ with regard to how much the parameters are constrained by prior knowledge. At one end of the spectrum, one can sometimes use *a priori* knowledge about the value of specific parameters in the modelled system. For example, biophysical models of neurons, e.g. the Hodgkin–Huxley models, typically use a range of fixed parameter values for ion channel gating probabilities, conductances and reversal potentials that are based on experimental measurements (see Dayan & Abott, 2001). From a Bayesian perspective, this corresponds to priors with infinite precision, and the goal of this kind of model is not to estimate the model parameters given some data, but to show that the system model, given its structure and some realistic inputs, can reproduce some empirically observed functional behaviour. By contrast, models such as Structural Equation Models of neuroimaging data (see below) are usually interested in finding those SFR parameters that best explain how some observed data could have been generated from the system with its assumed structure (McIntosh et al. 1994; Büchel & Friston, 1997). Therefore, these types of model do not usually constrain the parameter values; this corresponds to flat priors with zero precision. An intermediate approach is to constrain parameter values by priors with empirically motivated variance (i.e. non-zero, non-infinite precision). Such priors can either be constructed from basic principles (e.g. the parameter of a decay term could be constrained to be negative) or based on empirically measured distributions of values. A representative of this intermediate approach is DCM, which is described below.

5. Setting criteria of inference. It is crucial to state precisely the actual hypothesis that one intends to test using a system model. For example, if one wishes to establish that the overall SFR (as embodied by F in Eq. 4 and thus by the joint choice of elements, connectional structure and functional form of f_i) is a plausible mechanism underlying a certain functional behaviour, one is primarily interested in how well the model fits observed data. This question is usually addressed in a model comparison context in which different models, representing competing hypotheses, are compared against each other with regard to model fit and model complexity (see Pitt & Myung, 2002; Penny et al. 2004, for details). On the other hand, one may be interested in a particular component of the model, e.g. whether a given connection strength is modulated by context (Büchel & Friston, 1997; Friston et al. 2003). This hypothesis can then be addressed by means of a statistical test on those parameter estimates that represent the modulatory mechanism.

Functional neuroimaging is generally considered to be part of 'systems neuroscience'. If one accepts the above list of necessary steps for system identification, one may ask to what extent common approaches in functional neuroimaging actually provide insights into the SFRs of neural systems. In order to answer this question, we first need to review the conceptual and methodological basis on which most fMRI experiments rest.

4. Methodological and conceptual foundations of fMRI analyses

4.1. Standard convolution models for fMRI analysis

Standard analyses of fMRI data rely on mass-univariate statistical tests: for each volume element (voxel) in the brain, they compute the correlation with some experimentally controlled variable that describes an aspect of function, e.g. a stimulus function or a task sequence. Because we usually deal with more than one experimental condition, the analysis is performed as a multiple linear regression, or equivalently, as an analysis of variance with indicator variables. These are all special cases of the GLM:

$$y = X\beta + e \tag{10}$$

which models voxel-specific BOLD responses y in terms of a linear combination of explanatory variables (columns of the design matrix X) whose contributions are weighted by the parameter vector β, plus an independently and identically distributed Gaussian error term e. The design matrix includes all known variables that may explain the evoked neural responses. Importantly, we can observe neural responses only indirectly in terms of their haemodynamic effects, i.e. evoked BOLD signals, and we need to take this into account when constructing the design matrix X. One way of doing this is to use a canonical haemodynamic impulse response function (HRF), which describes the characteristic haemodynamic response to a brief neural event and thus characterizes the input–output behaviour of a given voxel. In the standard convolution model for fMRI analysis the stimulus functions are convolved with an HRF to give predicted haemodynamic responses that enter as regressors in the design matrix (Friston et al. 1994). To account for variability in the HRF from voxel to voxel and subject to subject (Handwerker et al. 2004), temporal basis functions can be used to express the predicted BOLD response as the linear combination of several functions of peristimulus time (Henson, 2004), or the HRF can be estimated directly from the data (Marrelec et al. 2003).

The goal of this approach is to test where in the brain (i.e. in which voxels) changes in the BOLD signal can be modelled as a function of experimentally controlled changes in cognitive function. Technically, this is usually done in the form of a contrast $c^T\beta$ where β is the vector of parameter estimates from Eq. (10), T is the transpose operator, and c is a weighting vector that expresses the hypothesis to be tested. For example, $c^T\beta = [1 \quad -1]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$ is a contrast that tests the null hypothesis $\beta_1 - \beta_2 = 0$ where β_1 and β_2 are the parameter estimates of two different experimental conditions. Put simply, by representing a linear combination of the experimental conditions, a contrast represents a particular task component.

Following voxel-wise hypothesis testing on the basis of a chosen contrast, the final step of standard fMRI analyses is to create a statistical parametric map (SPM) to visualize the spatial distribution of significant effects. Using a GLM in this fashion is equivalent to asking: what are the brain voxels whose time series are correlated to a certain task component? In other words, the standard convolution model for fMRI is a tool to search for SFCs. This is not only true at a conceptual level, but also in a strict technical sense: whatever the specific statistical question asked by means of a contrast within the context of a GLM, it can be reformulated in terms of testing for partial correlations. This is because for any design matrix X with p columns and for a chosen contrast weight c, one can find a $p \times p$ matrix D such that

$$\begin{aligned} y &= X\beta + e \\ &= XDD^{-1}\beta + e \\ &= \tilde{X}\tilde{\beta} + e \\ \tilde{X} &= XD,\ \tilde{\beta} = D^{-1}\beta \\ X_0 &= [q_1 \quad \cdots \quad q_{p-1}] \\ D &= [c \quad X_0] \\ \tilde{c} &= [1 \quad 0 \quad \cdots \quad 0]. \end{aligned} \tag{11}$$

Given such a matrix D, testing for the contrast $c^T\beta$ is identical to testing for $\tilde{c}^T\tilde{\beta}$; the latter corresponds to determining the partial regression of the voxel time series onto the task component of interest, represented by the product of the design matrix and the contrast weights (i.e. Xc).[1] Partial regression, however,

[1] D corresponds to a transformation matrix of the bases of the design space and can be constructed from a given contrast weight c by standard procedures such as Gram–Schmidt orthogonalization.

can be directly converted to partial correlation in various ways, for example by

$$r^2_{y,Xc|XX_0} = \frac{\text{extra SS due to adding } Xc \text{ to a regression model containing } X_0}{\text{residual SS using only } XX_0 \text{ in the model}} \quad (12)$$

where $r^2_{y,Xc|XX_0}$ is the partical correlation between y and Xc after all other effects XX_0 have been accounted for and SS denotes 'sum of squares' (Kleinbaum et al. 1988, p. 154).

This short technical aside simply serves to demonstrate that the standard convolution model for fMRI data, based on the GLM, is in direct continuity with the classical search for SFCs. The important question is how such correlations, represented spatially by an SPM, can be interpreted. It is instructive to consider what type of system model is implicitly represented by this approach; it is a model in which all system elements are disconnected from each other and the experimental variables act as external inputs that affect each element of the system directly (see Eq. 10). In the context of neuroscience, this model would represent a brain in which the individual processing units (e.g. neurons or cortical areas) are disconnected and receive external inputs in a direct and instantaneous fashion magically through the skull, as it were. Figure 2 illustrates this concept, using the same set of elements as in Fig. 1 to highlight the differences between a system model in analogy to a GLM and one that specifies the interactions between elements as well as the sites where external inputs enter the system. It is obvious that the system model in Fig. 1 has a much higher biological plausibility for describing a neural system in the brain than the model in Fig. 2. This comparison serves to remind us that GLM-based approaches cannot deliver any mechanistic insights into systems as they are blind to both functional interactions and the spatial specificity of external inputs. They are, however, very useful to find candidate elements of a system that one wishes to characterize, particularly in cases where little *a priori* knowledge exists for that system.

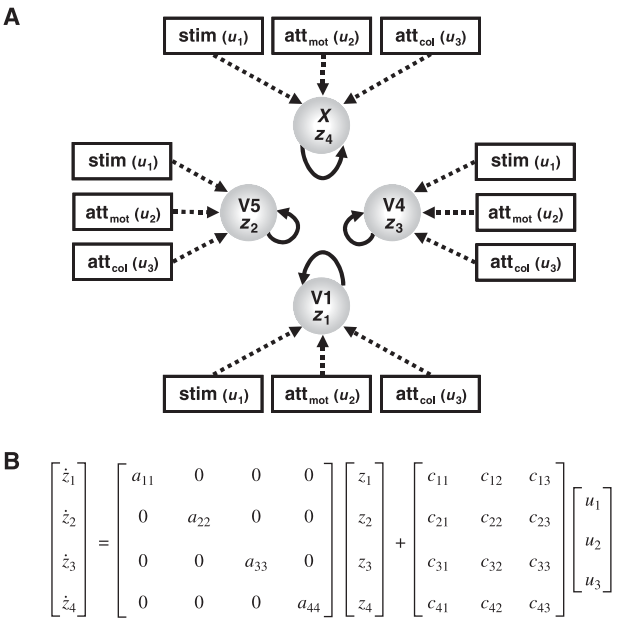


Fig. 2 Reformulation of the system model in Fig. 1 to make it equivalent to a GLM. Here, the areas in the system are completely disconnected (all off-diagonals in matrix *A* are zero) but are directly affected by all inputs (all entries in matrix *C* are non-zero).

4.2. Localizationism, functional specialization and functional integration

The simplest approach to interpreting SPMs of fMRI data is to take the perspective of localizationism. This approach assumes a one-to-one mapping between cortical areas and cognitive functions, a view that historically can be traced back to phrenology and has long been an important theme in neuropsychology (Phillips et al. 1984). In the context of neuroimaging, localizationism predicts a one-to-one SFC, i.e. that there should be significant voxel-wise correlations between a BOLD time series and the cognitive function of interest within a single area only, and that this area should not show analogous correlations with any other cognitive function. This constellation is rarely, if ever, observed. On the contrary, the general finding is that there exists a wealth of one-to-many and many-to-one SFCs across all cognitive domains (see Price & Friston, 2002; and Friston, 2003, for reviews on this topic). One could argue that this is simply due to the coarse resolution of current psychological concepts and the ensuing constraints on experimental designs. Given sufficient progress in psychological theory, it might therefore eventually be possible to demonstrate that, at a very fine-grained conceptual level, each cortical area computes a unique function. An interesting idea in this context is to use the output from a computational model of a specific cognitive function as a regressor in a GLM (O'Doherty et al. 2003; Seymour et al. 2004). However, even these

sophisticated models, which give a more precise account of what a given area may compute, do not change the fundamental limitation of correlative approaches: even a perfect correlation between a local neurophysiological signal and the prediction from a computational model of a cognitive function does not explain in any way how this function is neurally implemented.

As there are no disconnected neural units in the brain, any mechanistic explanation of local brain function in neurophysiological terms must be based on a system model that takes into account the interactions between elements. This notion is backed by much experimental evidence. For example, throughout the whole visual system with its highly specialized areas, local information processing is strongly modulated by a wide range of contextual information, a process that has been demonstrated to depend on backward connections from hierarchically higher areas (Hupe et al. 2001; Moore & Armstrong, 2003). Even at the level of basic visual feature processing in area V1, strong contextual effects have been observed in the absence of any stimulus changes, e.g. modulation of neuronal responses by implicit memory (Olson et al. 2001), spatial attention (Motter, 1993) or feature-based attention (Mehta et al. 2000; Murray & Wojciulik, 2004). Another piece of evidence against localizationism is given by disconnection syndromes in which local information processing in an intact area is altered when its input from remote areas is changed because of lesions in grey or white matter (Absher & Benson, 1993).

For all these reasons, localizationist ideas no longer play an important role in most theories of brain function (as a possible exception, some theories of visual perception still have a strongly modular character, e.g. Grill-Spector et al. 2004). Instead, current cognitive neuroscience takes an explicitly system-based perspective. A common view is that the areas that constitute a given system are functionally specialized, but the exact nature of their individual computations depends on context, e.g. time effects and the nature of their inputs from other areas. The cognitive function is implemented by the aggregate behaviour of the system depending on the neural context, i.e. the context-dependent interactions between the system components (McIntosh, 2000). This perspective is also reflected in the well-known concepts of functional specialization and functional integration (Friston, 1995, 2002). The functional specialization concept 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.

The great majority of current functional neuroimaging experiments have adopted this view and interpret the areas that are jointly correlated to a certain task component as the elements of a distributed system that represents the neural basis of the cognitive task. However, this explanation is incomplete as long as no insight is provided into how the locally specialized computations are bound together by context-dependent interactions between these areas; this is the functional integration within the system. Methodologically, statements on functional specialization require voxel-wise statistical tests for the correlation between regional time series and task components; this is provided by GLM analyses. In contrast, functional integration within distributed neural systems is usually best understood in terms of effective connectivity. As described in section 3.2, effective connectivity is the influence that the system elements exert over another (Friston, 1995). It has been 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). This definition emphasizes that effective connectivity is context-dependent and rests on a causal model of the interactions. Importantly, functional specialization, assessed by GLM analyses, and functional integration, characterized in terms of effective connectivity, are not contradictory approaches, but complement each other: whereas GLM analyses reveal candidate elements of a given system, models of effective connectivity can test hypotheses about the nature of the interactions between these elements and thus about functional principles of the system. As described in section 3.3, these two steps are essential procedures of neural system identification using neuroimaging.

It should be mentioned that, in addition to effective connectivity, another basic metric of functional integration exists, i.e. functional connectivity, which is defined as the temporal correlation between time series from different brain regions (Friston, 1995). Analyses of functional connectivity do not incorporate any knowledge about the system structure and its hypothetical SFR. In this sense, functional connectivity approaches are model-free. Depending on the amount of knowledge about the system under investigation, this can either be a strength or a weakness. If the system is

largely unknown, functional connectivity approaches are very useful because they can be used in an exploratory fashion, either by computing functional connectivity maps with reference to a particular seed region (Bokde et al. 2001; Stephan et al. 2001a; McIntosh et al. 2003) or using a variety of multivariate techniques that find sets of voxel time series that represent distinct (e.g. orthogonal or independent) components of the covariance structure of the data (McIntosh et al. 1996; Friston & Büchel, 2004). The information from these analyses can then be used to generate hypotheses about the system. On the other hand, if some information is available on the system structure and if there is a specific hypothesis about the SFR of the system, models of effective connectivity are usually more appropriate. This article deals with the question of how system models, based on hypotheses about structure and intrinsic mechanisms of the system, can be used to test hypotheses about SFRs, using neuroimaging data. The following sections therefore neglect functional connectivity approaches and deal with models of effective connectivity only.

5. Are system concepts taken seriously in neuroimaging?

At first sight, the system concept as it is expressed in the ideas of 'neural context' and 'functional specialization/integration' described above seems to have been embraced by the neuroimaging community. From subjective experience, a very large proportion of neuroimaging articles frame the interpretation of their results by the concept of distributed neural systems. This can also be demonstrated by a simple literature search: on 16 May 2004, a query using the public literature database *PubMed* (www.pubmed.org) found 344 articles from cognitive studies using fMRI that referred to 'system(s)', 'circuit(s)' or 'network(s)' in their title or abstract[2] (as opposed to 566 fMRI articles that did not mention any of these terms explicitly in the title or abstract). On closer inspection, however, the necessity of system-based analyses is taken much less seriously. So far, most fMRI studies have only demonstrated a significant BOLD correlation with a task component of interest and have thus restricted themselves to statements on regional functional specialization. In the above literature search, only 27 (7.9%) of the fMRI-related articles that did refer to 'system(s)' or 'network(s)' also explicitly mentioned 'connectivity' in the title or abstract.

It is worth asking why there is widespread support for the notion that cognitive functions are implemented by neural systems, and yet relatively few analyses so far have gone beyond functional specialization approaches and investigated the interactions between candidate elements of a system. At least three potential explanations come to mind. The first is simple: analyses of functional interactions tend to be methodologically more challenging than analyses of functional specialization using the standard GLM-based convolution model. Although a variety of publicly available and convenient software tools for GLM analysis of fMRI data have existed for a long time, tools for connectivity analyses that can be used through graphical user interfaces have been provided only relatively recently (e.g. DCM in SPM2, Granger causality analysis in BrainVoyager – see below). Previously, analyses of connectivity had to be done by means of custom-written software (e.g. Büchel & Friston, 1997; Bodke et al. 2001; Stephan et al. 2001a) or by exporting fMRI data to standard statistical packages (McIntosh et al. 1994; Honey et al. 2003).

The second potential explanation is that there still is a certain lack of understanding as to what is needed to investigate and characterize a system properly. This may be due to the fact that traditional teaching curricula in many disciplines such as biology, medicine and psychology have rarely included a formal introduction to systems theory in the past. Although this is now starting to change (see below), the necessary methodological skills and concepts for exploring systems properly are not as widespread as one would hope. For example, a problem that is commonly encountered in neuroscience is the belief that a mere enumeration of the elements in a system conveys a basic understanding of its nature. One particularly salient example is the longstanding confusion about the definition of the 'limbic system': not only is there a large variety of different anatomical enumerations for this vague concept, but precise system models of how certain brain regions interact to mediate a certain function are almost absent (see LeDoux, 1991; Kötter & Stephan, 1997, for reviews on this topic). This notion that a system is sufficiently described by a list of its constituent

[2]The query syntax used was: 'fmri [tiab] and cognitive and (system [tiab] OR systems [tiab] OR circuit [tiab] OR circuits [tiab] OR network [tiab] OR networks [tiab]).

elements is also encountered in many neuroimaging studies. In these studies, the set of activated areas (the 'network') that is found in a particular analysis is taken as a satisfactory description of the system that mediates the cognitive function. In the discussion of these articles, the observed activation pattern is then often dissected into regional activations that are being assigned some particular subfunctionality within the system; this interpretation of individual network nodes relies on informal comparisons with other imaging studies and sometimes primate connectivity data, but is not based on any formal model. The danger of the widespread acceptance of this approach in neuroimaging is that it encourages experiments being performed in an entirely exploratory fashion without any precise *a priori* hypothesis about the system of interest. The results can be interpreted *post hoc* in the form of a story that explains how the observed activation pattern might have been produced by some underlying neural system.

This criticism has been formulated previously, for instance by Kosslyn (1999), and since then the overall quality of neuroimaging research has certainly improved, with a stronger emphasis on specific hypotheses and wider awareness of the importance of functional integration analyses. Still, as discussed above, only a minority of studies take a system-based perspective seriously. The third and final explanation offered for this is that there seems to be an implicit notion that functional maps from GLM analyses are sufficient to provide at least some general information about the interactions among the activated areas. This notion is revealed by a tendency to interpret 'co-activation' as evidence for some sort of co-operation within the same system: 'Sometimes researchers talk about a set of areas as a circuit, but this is usually misleading: in most studies all that is revealed are a set of activated (and/or deactivated) areas, with no information about the flow of information between the areas' (Kosslyn, 1999). Indeed, even in recent papers the finding that several areas are jointly correlated to some task component is still sometimes interpreted as a reflection of mutual correlation and thus of functional connectivity among the areas (e.g. Gold & Buckner, 2002; Dolcos et al. 2004).

A simple example demonstrates that this assumption about the transitivity of correlations is not always valid. Let us imagine two regional BOLD time series, A_1 and A_2 (red and magenta lines in Fig. 3), which have been acquired during a task that is described by the function T (blue line in Fig. 3). For simplicity, both BOLD time series and the task function are represented as sine waves of identical amplitude that simply differ in phase. If, for example, A_1 is shifted by $-\pi/4$ relative to T, the correlation between them is high: $r(A_1, T) = 0.71$. If A_2 is shifted by the same amount but in the opposite direction (i.e. $+\pi/4$) relative to T, it shows exactly the same correlation with the task: $r(A_2, F) = r(A_1, F) = 0.71$. Therefore, in a GLM-based analysis, both A_1 and A_2 would appear in the same SPM as areas that are highly and identically correlated with the task. However, when testing for the correlation between the two time series, A_1 and A_2 are found to be entirely uncorrelated. This can be easily seen from the fact that the correlation between two vectors is identical to the cosine of their angle, and the angle between two periodic functions of the same frequency corresponds to their phase

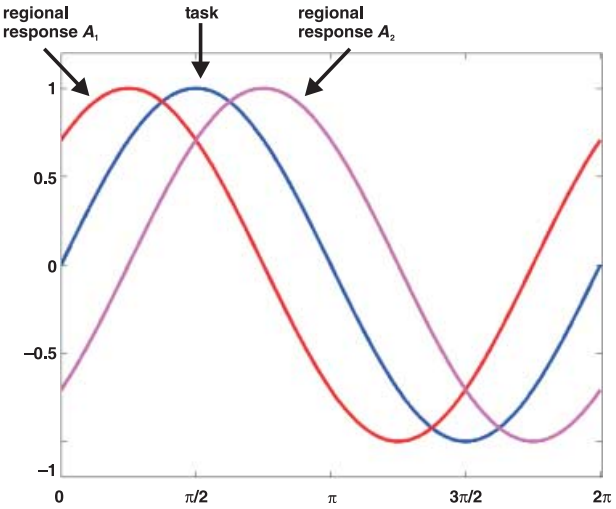


Fig. 3 An example that transitivity of correlation does not generally hold. This example shows two fictitious regional BOLD time series, A_1 and A_2 (red and magenta lines), acquired during a task that is described by the function T (blue line). For simplicity, BOLD time series and the task function are represented as sine waves of identical amplitude that are shifted in phase. The *y*-axis represents BOLD signal amplitude and the *x*-axis represents time in multiples of $\pi/2$ (both axes: arbitrary units). A_1 is shifted by $-\pi/4$ relative to T, whereas A_2 is shifted by the same amount but in the opposite direction (i.e. $+\pi/4$) relative to T. The correlation between the time series and the task function is high and identical for both areas: $r(A_1, T) = r(A_2, T) = 0.71$. Therefore, in a GLM-based analysis, both A_1 and A_2 would appear in the same SPM as areas that are highly and identically correlated with the task. However, when testing for the correlation between the two time series, A_1 and A_2 are found to be entirely uncorrelated: $r(A_1, A_2) = 0$ (see main text for details).

difference ($<A_1, A_2>$ denotes the dot product of the two time series vectors):

$$\begin{aligned} r_{A_1, A_2} &= \frac{<A_1, A_2>}{\sqrt{<A_1, A_1>} \cdot \sqrt{<A_2, A_2>}} \\ &= \cos(\pi/2) \\ &= 0. \end{aligned} \quad (13)$$

In summary, the finding of a set of areas to be jointly correlated to a certain task component (and thus 'co-activated') is not sufficient to demonstrate that these areas are functionally connected to each other nor does it characterize this system in any satisfactory depth: no insights are gained into the mechanisms that underlie the observed correlations between the local time series and the task component. Therefore, after one has identified candidate elements of the neural system by means of a GLM-based analysis, a subsequent analysis of their functional integration is required to provide a model for the SFR of the underlying neural system.

This requirement was recognized very early in the history of neuroimaging, and considerable effort has been invested in establishing techniques that can be used for inferring principles of functional integration from neuroimaging data (e.g. Horwitz et al. 1984, 1998; McIntosh et al. 1994, 1999; Büchel & Friston, 1997; Friston et al. 1997, 2003; Friston & Büchel, 2000). Given the long history of these techniques for assessing connectivity within neural systems and the success of their applications, it is somewhat surprising that they are still playing a subordinate role in current neuroimaging studies. The following section summarizes the conceptual foundations of some of these methods and highlights their strengths and limitations.

6. Models of effective connectivity

As described above, functional integration within distributed neural systems is usually best understood in terms of effective connectivity. Effective connectivity aims to make statements about the influence that neural units exert on another, i.e. statements about causal effects. The fundamental problem is that all we have to infer causality from are observed regional time series and their correlations in time. Inferring causality from correlational data is a longstanding statistical problem: for any given data set, there are multiple ways in which the correlation between two elements A and B might have been produced. For example, (1) A might influence B, (2) B might influence A, (3) A and B might influence each other or (4) A and B might not interact at all but are similarly influenced by a third element. This means that inferences to causal principles must be based on a model of the interactions in the system. This model comprises two components: (1) a structural model that describes which neural units (e.g. cortical areas) are elements of the system and how they are linked by anatomical connections, and (2) a model of the SFR that describes what kind of causal influences shape the dynamics, and how these influences are constrained by the structural model.

Together, the structural and mathematical components represent a model of the overall SFR in the system of interest. If we express this in terms of the general system descriptions presented in section 3.2, the structural component is given by the binary connectivity matrix A (see Eq. 9), and the model of the SFR corresponds to F in Eq. (4).

The choice of the structural model is strongly hypothesis-driven. It is usually based on the results from conventional fMRI analyses to define the nodes of the modelled system and on data from neuroanatomical studies to define the connections. Because of the paucity of connectivity data on the human brain, the latter information usually has to be inferred from tract tracing studies in monkeys, a task that has been facilitated by means of large databases of published connectivity data (Stephan et al. 2001b). In this article, we only deal with system models with very simple structural components, i.e. each element of the system represents the population activity of a whole cortical area; however, several large-scale models have been proposed recently that represent each area by multiple state variables representing, for example, different layers (Kötter et al. 2002) or distinct neuronal populations with different biophysical parameters (Robinson et al. 2001; David & Friston, 2003).

The mathematical models of the assumed SFR reflect different ways of thinking how neural processes take place in the brain, e.g. whether they are linear or nonlinear and whether they are dependent or independent of history, time and context effects. Most of the models that have been proposed in the past are static linear models based on regression and covariance partitioning techniques, e.g. SEM (McIntosh et al. 1994; Büchel & Friston, 1997) or MAR (Harrison et al. 2003; Göbel et al. 2003). We briefly review and juxtapose these methods to the most recent approach, DCM,

which uses a dynamic and bilinear model. Finally, as a special case, we briefly discuss PPIs (Friston et al. 1997). Although PPIs contain elementary components of system descriptions as outlined in section 3.2, they only address pair-wise interactions, which renders them too simple to be a proper system model.

To keep the notation comparable across models, the following convention has been adopted: lower case variables denote column vectors, and upper case variables denote matrices. *y* represents measured data and *z* are hidden states. *u* represents external inputs into the system. *A*, *B* and *C* are parameter matrices with *A* representing context-independent ('intrinsic') connectivity between system components, *B* representing context-dependent modulation of these connections and *C* representing the strengths of external inputs *u*. For PPIs, the parameters are scalars, and are analogously named, i.e. *a*, *b* and *c*.

Non-mathematically inclined readers should go to section 7.

6.1. Structural equation modelling (SEM)

SEM has been an established statistical technique in the social sciences for several decades, but was only introduced to neuroimaging in the early 1990s by McIntosh & Gonzalez-Lima (1991). It is a multivariate, hypothesis-driven technique that is based on a structural model that represents the hypothesis about the causal relations between several variables (see McIntosh et al. 1994; Büchel & Friston, 1997; Bullmore et al. 2000, for methodological details). In the context of fMRI these variables are the measured BOLD time series $y_1, \ldots, y_n$ of *n* brain regions and the hypothetical causal relations are based on anatomically plausible connections between the regions. The strength of each connection $y_i \rightarrow y_j$ is specified by a so-called 'path coefficient', which, analogous to a partial regression coefficient, indicates how the variance of y_i depends on the variance of y_j if all other influences on y_i are held constant.

The statistical model of standard SEM implementations for fMRI data can be summarized by the regression-like equation

$$y = Ay + u \tag{14}$$

where *y* is an $n \times s$ matrix of *n* area-specific BOLD time series with *s* scans each, *A* is an $n \times n$ matrix of path coefficients (with zeros for non-existent connections), and *u* is an $n \times s$ matrix of zero mean Gaussian error terms, which are driving the modelled system ('innovations', see Eq. 15 below). Parameter estimation is achieved by minimizing the difference between the observed and the modelled covariance matrix Σ of the areas (Bollen, 1989). For any given set of parameters, Σ can be computed by transforming Eq. (14):

$$\begin{aligned} y &= (I - A)^{-1}u \\ \Sigma &= yy^T \\ &= (I - A)^{-1}uu^T(I - A)^{-1^T} \end{aligned} \tag{15}$$

where *I* is the identity matrix and *T* denotes the transpose operator. Note that the model on which SEM rests is very similar to the general equation for non-autonomous linear systems (with the exception that SEM is a static model and the inputs to the modelled system are random noise; compare Eqs 14 and 7). The first line of Eq. (15) can be understood as a generative model of how system function results from the system's connectional structure: observed BOLD activity results from filtering the Gaussian innovations *u* by a function of the interregional connectivity matrix, i.e. $(I - A)^{-1}$. This is a concrete example of how models of effective connectivity represent models of SFRs, although, as we will see below, other techniques such as DCM allow for biologically more realistic models.

If an SEM is fitted to the BOLD time series of a given experiment, the resulting path coefficients (i.e. the parameters in *A*) describe the effective connectivity of the modelled system across the entire experimental session. This is usually not very interesting. What one would like to know instead is how the coupling between certain regions changes as a function of experimentally controlled context, e.g. differences in coupling between two different tasks. Notably, SEM does not account for temporal order: if the regional time series were permuted in the same fashion, the estimated parameters would not change. In the case of blocked designs, this makes it possible to partition a time series into condition-specific subseries to which separate SEMs are fitted. These SEMs can then be compared to test for condition-specific differences in effective connectivity (for examples, see Büchel et al. 1999; Honey et al. 2002). An alternative (and arguably more elegant) approach is to incorporate bilinear terms in the model that represent the modulation of a given connection by an experimentally controlled context (e.g. Büchel & Friston, 1997;

Rowe et al. 2002, 2004); in this case, only a single SEM is fitted to the entire time series.

6.2. Multivariate autoregressive models (MAR)

In contrast to SEM, autoregressive models explicitly address the temporal aspect of causality in BOLD time series, focusing on the causal dependence of the present on the past: each data point of a regional time series is explained as a linear combination of past data points from the same region. MAR models extend this approach to n brain regions, modelling the n-vector of regional BOLD signals at time t (y_t) as a linear combination of p past data vectors whose contributions are weighted by the parameter matrices A_i:

$$y_t = \sum_{i=1}^{p} y_{t-i} A_i + u_t. \quad (16)$$

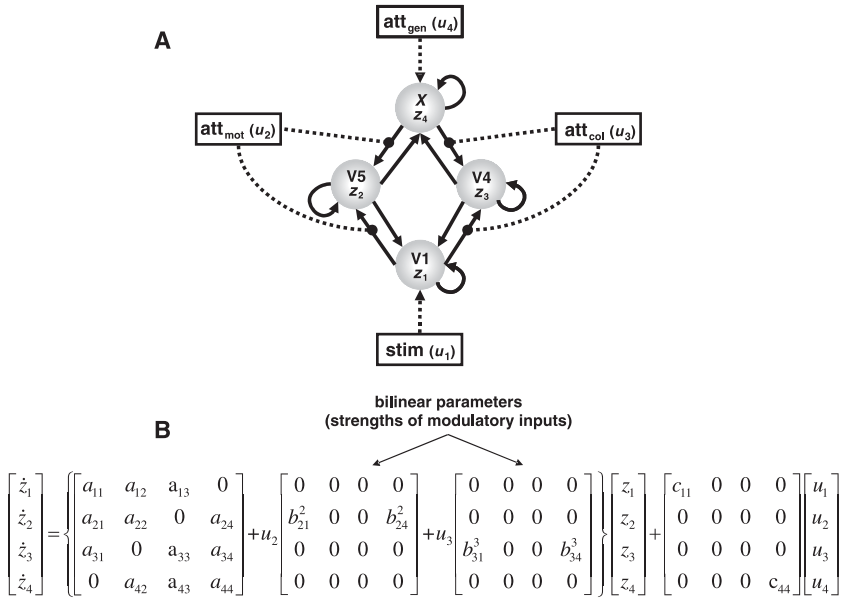
In summary, MAR models directed influences among a set of regions whose causal interactions, expressed at the BOLD level, are inferred via their mutual predictability from past time points. Although MAR is an established statistical technique, specific implementations for fMRI were suggested only recently. Harrison et al. (2003) suggested an MAR implementation that allowed for the inclusion of bilinear variables representing modulatory effects of contextual variables on connections and used a Bayesian parameter estimation scheme (Penny & Roberts, 2002). This Bayesian scheme also determined the optimal model order, i.e. the number of past time points (p in Eq. 16) to be considered by the model. A complementary MAR approach, based on the idea of 'Granger causality' (Granger, 1969), was proposed by Goebel et al. (2003). In this framework, given two time-series y_1 and y_2, y_1 is considered to be caused by y_2 if its dynamics can be predicted better using past values from y_1 and y_2 as opposed to using past values of y_1 alone.

6.3. The need for models of effective connectivity at the neural level

Both SEM and MAR have limitations. One disadvantage of SEM is that one is restricted to use structural models of relatively low complexity: models with reciprocal connections and loops often become non-identifiable or show unstable parameter estimates (see Bollen, 1989, for details). However, there are heuristics for dealing with complex models that use multiple fitting steps in which different parameters are held constant while changing others (see McIntosh et al. 1994, for an example). A second problem, as mentioned above, is that SEM is not a proper time series model. A third complication is shared by SEM and MAR: testing for context-dependent changes in effective connectivity becomes problematic in event-related designs. This is because of the transient nature of the evoked responses, which makes both partitioning of the time series (in SEM) and the use of bilinear modulation terms (in SEM and MAR) difficult (see Gitelman et al. 2003). Finally, the standard formulations of SEM and MAR do not make use of our knowledge when external inputs (e.g. sensory stimulation) entered the system; instead, the driving inputs are random innovations (see Eqs 14–16). This is suboptimal for systems where we know the external inputs: for a causal description of the system dynamics, we need to describe (1) when and where external inputs enter the system and (2) how the initially induced activity then propagates through the rest of the system according to its connectional structure. However, there are ways of adapting both methods such that knowledge about the inputs is incorporated into the models (e.g. Harrison et al. 2003; Mechelli et al. 2002).

Possibly the most important limitation of both methods, however, is a conceptual one. The causal architecture of the system that we would like to unravel is expressed at the level of the neuronal dynamics. However, the parameters in SEM and MAR are fitted to BOLD series, which result from a convolution of the underlying neural activity. Because this transformation of neural activity to BOLD has non-linear components (Friston et al. 2000; Miller et al. 2001), any inference on interregional coupling obtained by SEM or MAR is only an indirect one, and it is not trivial to estimate to what degree the estimated coupling was affected by the transformation from the neural to the BOLD level.

What is needed to enable inferences about neural parameters in the context of fMRI are models that combine two things: (1) a parsimonious but neurobiologically plausible model of neural dynamics, and (2) a biophysically plausible haemodynamic forward model that describes the transformation from neural activity to BOLD. These models make it possible to fit neural and haemodynamic parameters such that the resulting BOLD series, generated by the forward model, are optimally similar to the observed BOLD time series. Of course this general



$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\\ \dot{z}_4\end{bmatrix} = \left\{\begin{bmatrix} a_{11} & a_{12} & a_{13} & 0\\ a_{21} & a_{22} & 0 & a_{24}\\ a_{31} & 0 & a_{33} & a_{34}\\ 0 & a_{42} & a_{43} & a_{44}\end{bmatrix} + u_2\begin{bmatrix} 0 & 0 & 0 & 0\\ b^2_{21} & 0 & 0 & b^2_{24}\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} + u_3\begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ b^3_{31} & 0 & 0 & b^3_{34}\\ 0 & 0 & 0 & 0\end{bmatrix}\right\}\begin{bmatrix} z_1\\ z_2\\ z_3\\ z_4\end{bmatrix} + \begin{bmatrix} c_{11} & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & c_{44}\end{bmatrix}\begin{bmatrix} u_1\\ u_2\\ u_3\\ u_4\end{bmatrix}$$

Fig. 4 (A) Reformulation of the system model in Fig. 1 from the perspective of DCM. Here, attention to colour (u_3) and attention to motion (u_2) no longer have direct effects on V4 and V5, respectively, but modulate the strengths of the afferent connections to V4 and V5, respectively. Modality-specific modulation of the connections from V1 accounts properly for gain control effects, i.e. attention induces signal increases in V4 and V5 only in the presence of visual stimuli. In contrast, modality-specific modulation of the connections from the 'higher' area *X* (e.g. in parietal or prefrontal cortex) whose activity is directly influenced by attention independent of modality (see direct input att$_{gen}$) is a mechanism to represent baseline shifts, i.e. attention-induced signal increases in the absence of stimuli. (B) Mathematically, the state equation of this model differs from the equation in Fig. 1 by the inclusion of bilinear terms (see *B* matrices) that encode context-dependent changes in connections. Compare Eq. (18) in the main text.

type of model is not restricted to fMRI; indeed, models of this kind have been suggested for EEG (Yamashita et al. 2004). For fMRI, DCM (Friston et al. 2003) is the only approach to date that marries models of neural dynamics with biophysical forward models.

6.4. Dynamic causal modelling (DCM)

DCM offers a simple model for the neural dynamics in a system of *n* interacting brain regions. It models the changes of a neural state vector *z* in time, with each region in the system being represented by a single state variable (see Eq. 17). These neural state variables do not map precisely onto some common neurophysiological measurement but represent a summary index of neural population dynamics in the respective regions. The neural dynamics is 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) or they can modulate the coupling among regions (e.g. during learning or attention). The changes of the neural states in time (i.e. the first derivative of the state vector *z* with regard to time *t*) are therefore a function of the states themselves, the inputs *u* and some parameters θ^n that define the functional architecture and interactions among brain regions at a neuronal level (*n* in θ^n is not an exponent but a superscript that denotes 'neural'):

$$\begin{bmatrix}\frac{dz_1}{dt}\\ \vdots\\ \frac{dz_n}{dt}\end{bmatrix} = \frac{dz}{dt} = F(z, u, \theta^n). \quad (17)$$

Note that this equation has exactly the same form as the one that was introduced in the earlier section on general system theory (see Eq. 4) and on which many other system models have been based in the past (von Bertalanffy, 1950; Bossel, 1992). Concerning the specific definition of *F*, the neural state equation in DCM uses a bilinear form:

$$\frac{dz}{dt} = Az + \sum_{j=1}^{m} u_j B^j z + Cu. \quad (18)$$

Equation (18) is an extension of Eq. (7), which was introduced earlier for a general description of linear non-autonomous systems. Given this bilinear form, the neural parameters $\theta^n = \{A, B, C\}$ can be expressed as

partial derivatives of F (in the following, $\dot{z}$ is used as a short notation for dz/dt):

$$A = \frac{\partial F}{\partial z} = \frac{\partial \dot{z}}{\partial z}$$
$$B^j = \frac{\partial^2 F}{\partial z \partial u_j} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z} \quad (19)$$
$$C = \frac{\partial F}{\partial u}.$$

The matrix A represents the effective connectivity among the regions in the absence of modulatory 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 (see Fig. 4 for a concrete example, and compare it to Fig. 1).

DCM combines this neural model with an empirically validated biophysical forward model of the transformation from neuronal activity into a BOLD response (Friston et al. 2003; Stephan et al. 2004). This haemodynamic model consists of four differential equations with five parameters (θ^h) that describe how neural activity elicits a vasodilatory signal that leads to increases in blood flow and subsequently to changes in blood volume and deoxyhaemoglobine content. The predicted BOLD signal is a non-linear function of blood volume and deoxyhaemoglobine content (for details, see Friston et al. 2000; Friston, 2002).

The combined neural and haemodynamic parameter set $\theta = \{\theta^n, \theta^h\}$ is estimated from measured BOLD data y, 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 can be found in Friston et al. (2003). Eventually, the posterior distributions of the obtained parameter estimates can be used to test hypotheses about the size and nature of modelled effects. Usually, these hypotheses concern context-dependent changes in coupling. If there is uncertainty about the connectional structure of the modelled system, or if one would like to compare competing hypotheses (represented by different DCMs), a Bayesian model selection procedure can be used to find the DCM that shows an optimal balance between model fit and model complexity (Penny et al. 2004).

6.5. Psycho-physiological interactions (PPIs)

PPI is one of the simplest models available to assess functional interactions in neuroimaging data (for details see Friston et al. 1997). Given a chosen reference time series y_0 (obtained from a seed voxel or seed region), PPI computes whole-brain connectivity maps of this seed voxel with all other voxel time series y_i in the brain according to the equation

$$y_i = ay_0 + b(y_0 \times u) + cu + X\beta. \quad (20)$$

Here, a is the strength of context-independent connectivity between y_0 and y_i. The bilinear term $y_0 \times u$ represents the interaction between physiological activity y_0 and a psychological variable u, which can be construed as a contextual input into the system, modulating the connectivity between y_0 and y_i ($\times$ represents the Hadamard product, i.e. element-by-element multiplication). The third term describes the strength c by which the input u determines activity in y_i directly, independent of y_0. Finally, β are parameters for effects of no interest X (confounds).

Equation (20) contains elementary components of system descriptions as outlined in section 3.2. In fact, there is some similarity between the form of Eq. (20) and that of the state equation of DCM (Eq. 18). However, the fact that only pair-wise interactions are considered (i.e. separately between the reference voxel and all other brain voxels) means this model is severely limited in its capacity to represent neural systems. This has also been noted in the initial description of PPIs (Friston et al. 1997). Although PPIs are thus not a proper system model, they have an important role in exploring the functional interactions of a chosen region across the whole brain; this exploratory nature renders them similar to analyses of functional connectivity. The next section shows an empirical example that demonstrates that PPIs can be very useful despite their simplicity. Unlike analyses of functional connectivity, however, PPIs model the contextual modulation of connectivity, and this modulation has a directional character, i.e. testing for a PPI from y_0 to y_i is not identical to testing for a PPI from y_i to y_0. This is because regressing $y_0 \times u$ on y_i is not identical to regressing $y_i \times u$ on y_0. In other words, the bilinear term breaks the symmetry of the regression between the regional time series.

7. Analyses of effective connectivity: what do they mean, what are the limitations and what is the empirical benefit?

In this section, I review some results from previous studies that used models of effective connectivity to analyse neuroimaging data. The aim is to demonstrate what

kind of insights can be gained by taking an explicitly system-based perspective that takes into account the interactions between individual areas, and that these insights are impossible to infer from classic 'functional specialization' analyses alone.

Before starting to discuss any particular study or result, however, it is worth reflecting on what kind of understanding models such as those described in the preceding sections can actually provide. There are several potential arguments against the usefulness of this type of models. For simplicity, let us discuss these objections using DCM as a specific case. For example, one could argue that even though models such as DCM meet the formal requirements for descriptions of SFRs as outlined in section 3.2, they are not causal in the same sense as the function of an ion channel can be derived directly from its molecular structure (e.g. Miyazawa et al. 2003). In other words, what exactly is the 'causality', bridging structure and function in models like DCM? A second and related question is what does this mean in neurobiological terms, e.g. synaptic mechanisms, if a DCM tells us that a particular connection increases its strength during a particular experimental context? And finally, a third possible objection might be that the time constants of neuroimaging techniques like fMRI (as opposed to EEG or MEG) are too slow that any model fitted to such data could reflect the processes at the underlying neural level.

The answer to the first question, the nature of the causal SFR expressed by models such as DCM, is related directly to the general state equation of dynamic systems (Eq. 4). System models in this general framework provide a causal description of how system dynamics results from system structure because they (1) have temporal precedence characteristics (embodied in the differential equations), (2) describe when and where external inputs enter the system and (3) state how changes in time induced by these inputs are determined by the system's structure, i.e. its connectivity pattern and any other time-invariant properties (e.g. time constants). With regard to temporal precedence, two details should be added: first, this principle is only partially embodied in a DCM because delays between areas are not modelled, and second, temporal relations between neural processes do not necessarily need to be reflected by analogous latency differences at the BOLD level. Instead, the information about neural activity that is reflected at the BOLD level is contained largely in the relative amplitudes and shapes of the haemodynamic responses, not in their timings (this is discussed in detail by Friston et al. 2003). One of the strengths of the combined neural and haemodynamic model in DCM is that this information can be used to estimate connectivity parameters at the neural level that implicitly specify timing relationships not otherwise observable in the data. This is possible because DCMs have knowledge-based constraints on their architecture, in the form of Bayesian priors with different precision for neural and haemodynamic parameters (Friston et al. 2003).

With regard to the neurobiological interpretation of DCMs, they are obviously not specified at a level of neurobiological finesse that allows one to distinguish between different processes at synaptic, cellular, columnar or laminar levels. Instead, the mechanisms represented by the model, e.g. context-dependent changes of particular connection strengths, refer to the level of large neural populations contained by one or several voxels (even a single standard size voxel contains millions of neurons). However, this relatively high degree of abstraction present in DCMs does not mean that their causal mechanisms, represented by external inputs with temporal and spatial specificity, interregional influences mediated by connections and contextual modulations of these connections, are neurobiologically meaningless. For example, there is a specific class of potential synaptic mechanisms at the level of single neurons that underlie observed context-dependent changes in coupling at the population level; see figure 1 in Penny et al. (2004) and the discussion of the study by Büchel & Friston (1997) below. Moreover, Fig. 4 demonstrates how DCMs can be used to investigate questions about the relative strength of gain control and baseline shift mechanisms during visual attention; these are questions that have previously been addressed at the level of single neurons or microcircuits in invasive recording experiments (e.g. Luck et al. 1997). Finally, there is no principled reason against DCM-like models at smaller scales where the state variables correspond, for example, to laminae or columns. This may, however, require other data modalities than fMRI.

This leads to the final objection discussed here, i.e. the time constants of BOLD and other haemodynamic signals might be too slow that models fitted to such data could reflect the processes of real interest at the underlying neural level. This could be true for very brief and transient couplings, which may be reflected poorly

in the BOLD signal. On the other hand, simulations and empirical analyses have demonstrated that the temporal precision of DCM is within the range of a few hundred milliseconds (Friston et al. 2003). The current limitations in temporal precision are likely to be overcome by extending DCM to other modalities like EEG and MEG in combination with more complex state equations that represent finer scales of cortical organization (David & Friston, 2003).

With this discussion in mind, let us now turn to some practical examples of models of effective connectivity. Given that DCM was introduced about a year ago, only a few applications have been published so far, most of which are of a methodological nature (Friston et al. 2003; Mechelli et al. 2004; Penny et al. 2004). The following section therefore largely refers to classical models of effective connectivity such as SEM.

A classic PET study of effective connectivity in the visual system was performed by McIntosh et al. (1994). They used two matching tasks for faces and locations where the volunteers had to choose which of two stimuli corresponded to a reference stimulus. Both face and location matching tasks are known to have a right-hemispheric dominance and should show a relative preference for engaging the ventral and dorsal stream of the visual system, respectively. The latter was confirmed by the results from the conventional correlation analysis, but surprisingly, the activation pattern was bilateral for both tasks. Using SEM, McIntosh et al. could explain this result by showing that the interhemispheric connectivity showed a strong asymmetry, with right→left transcallosal connections between homotopic regions being much stronger during both tasks than left→right connections. They concluded that the observed bilateral activation during the two right-lateralized tasks was due to a transcallosal recruitment of the left hemisphere by the dominant right hemisphere. Importantly, this conclusion had not been feasible on the basis of the initial correlation analysis nor by simple inspection of the system's covariance matrix.

A seminal fMRI study on top-down processes in the visual system was performed by Büchel & Friston (1997), who examined the modulatory influence of attention on effective connectivity. In their experiment, the participants were shown a radially moving starfield stimulus. In one condition, they watched this stimulus passively while in the other condition they were instructed to pay attention to allegedly subtle changes in the speed of motion (which were actually absent). By comparing the 'attention' against the 'no attention' condition, Büchel & Friston (1997) showed that V5 responses to moving stimuli increased when these stimuli were attended to instead of being passively watched. This finding at the level of population dynamics was reminiscent of the well-known gain control effects described by invasive recording studies in monkeys where neural responses in visual areas increased during selective attention to specific properties of the stimuli (e.g. Luck et al. 1997). However, the sources of this attentional top-down effect had remained largely unclear. Using a simple hierarchical SEM with psycho-physiological interactions, Büchel & Friston (1997) demonstrated that for attention to motion and at the level of cortical areas this effect could be explained by a modulation of the V1→V5 connections by the SPC, and by a modulation of the V5→SPC connections by the inferior frontal gyrus (IFG). Although their model does not detail the exact mechanism at the synaptic and microcircuit level underlying this modulation, it provides crucial constraints: the only neurobiologically plausible type of synaptic mechanism that could account for the model's behaviour at the population level is a change in the dendritic response properties of V5 neurons to inputs from V1 neurons, and this is likely to be mediated through axons from another area that target the same V5 neurons as the inputs from V1 (see Penny et al. 2004). In spite of its simplicity, this model still provides one of the most compelling and anatomically precise suggestions of where and how attentional top-down influences occur in the visual system. Remarkably, these findings were confirmed in a series of subsequent analyses using a variety of different models of effective connectivity, including PPIs (Friston et al. 1997), Kalman filtering (Büchel & Friston, 1998), Volterra series (Friston & Büchel, 2000), MAR (Harrison et al. 2003) and DCM (Friston et al. 2003; Penny et al. 2004).

Beyond the particular study by Büchel & Friston (1997), the investigation of top-down effects has been a particular topic of interest for models of effective connectivity. Conventional neuroimaging studies of top-down effects like selective attention or maintenance of a particular cognitive set have consistently demonstrated the involvement of certain cortical areas, for example the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC) (e.g. Kastner et al. 1999; Ishai et al. 2000; Luks et al. 2002). They could not, however, (1) disentangle the differential

roles of the candidate source areas of top-down modulatory processes, (2) establish whether there was a single or multiple distinguishable modulatory process, or (3) clarify where the exact targets of the modulatory processes were located. For example, usually more than one putative source of top-down effects is found. Likewise, there is often more than one candidate target area where context-dependent changes in activity are observed. So how do these sources interact with each other and where and how do they modulate activity elsewhere in the brain? This question cannot be answered by conventional analyses, but, as demonstrated by Büchel & Friston's (1997) study and in the following two examples, is precisely what models of effective connectivity can address.

The first example is an fMRI study on the mechanisms underlying hemispheric specialization (Stephan et al. 2003). This study addressed the question of whether lateralization of brain activity depends on the nature of the sensory stimuli or on the nature of the cognitive task. For example, microstructural differences between hemispheres that favour the processing of certain stimulus characteristics and disadvantage others (Jenner et al. 1999) might mediate stimulus-dependent lateralization in a bottom-up fashion (Sergent, 1983). On the other hand, processing demands, mediated through cognitive control processes, might determine in a top-down fashion which hemisphere takes precedence over the other in accomplishing a given task (Levy & Trevarthen, 1976; Fink et al. 1996). To decide between these two possibilities, Stephan et al. (2003) used a protocol in which the stimuli were kept constant throughout the experiment, and subjects were alternately instructed to attend to certain stimulus features and ignore others. The stimuli were concrete German nouns (of four letters length each) in which either the second or the third letter was red. In a letter decision task, the subjects had to ignore the position of the red letter and indicate whether or not the word contained the target letter 'A'. In a visuospatial decision task they were required to ignore the language-related properties of the word and to judge whether the red letter was located left or right of the word centre.

The results of the conventional GLM analysis were clearly in favour of the top-down hypothesis: despite the use of identical word stimuli in all conditions, comparing letter to visuospatial decisions showed strongly left-lateralized activity, including classic language areas like Broca's area in the left IFG, whereas comparing visuospatial to letter decisions showed strongly right-lateralized activity in the parietal cortex. Yet it did not manage to clarify the actual mechanisms by which information processing was biased towards one hemisphere in a task-dependent fashion. The stimuli contained both letter and visuospatial information and thus required subjects to process only information that was meaningful for the current task and inhibit processing of any other information. Could this cognitive control process be the decisive 'switch' determining the relative involvement of the two hemispheres? If so, it should lead to task- and hemisphere-specific changes in functional coupling between control areas in the frontal lobe and areas related to the execution of the tasks. Comparisons between the two tasks and a baseline condition (a simple reaction time task on the same type of stimuli) showed that the only putative control area was the ACC. This area showed increased activity in both hemispheres during both tasks (Fig. 5A). However, when ACC connectivity with the rest of the brain was analysed, using a simple model of effective connectivity (PPIs; Friston et al. 1997), a striking hemispheric dissociation was found: left ACC specifically increased its coupling during letter decisions with the left IFG, an important language area (Fig. 5B), whereas the right ACC specifically increased its connectivity during visuospatial decisions with areas in the right parietal cortex known to be involved in spatial judgments (Fig. 5C). No other brain area showed significant task-dependent changes in coupling with either left or right ACC. Even though this analysis of effective connectivity did not detail the interactions between the areas involved in executing the two tasks, it provided a simple mechanistic description of the system that controlled the enhancement of activity in the task-relevant hemisphere.

The second example of how system models based on effective connectivity can elucidate top-down mechanisms is an fMRI study by Rowe et al. (2004). In this study, the authors examined the role of the DLPFC in free selection of a response among several options. The hypothesis was that DLPFC should be activated during free selection regardless of the modality of the selected item, but should convey the outcome of the selection process to modality-specific areas by means of modality-dependent changes in effective connectivity. This hypothesis was tested by contrasting selection tasks from two different domains: in a motor selection task, the participants could freely choose to press one of four

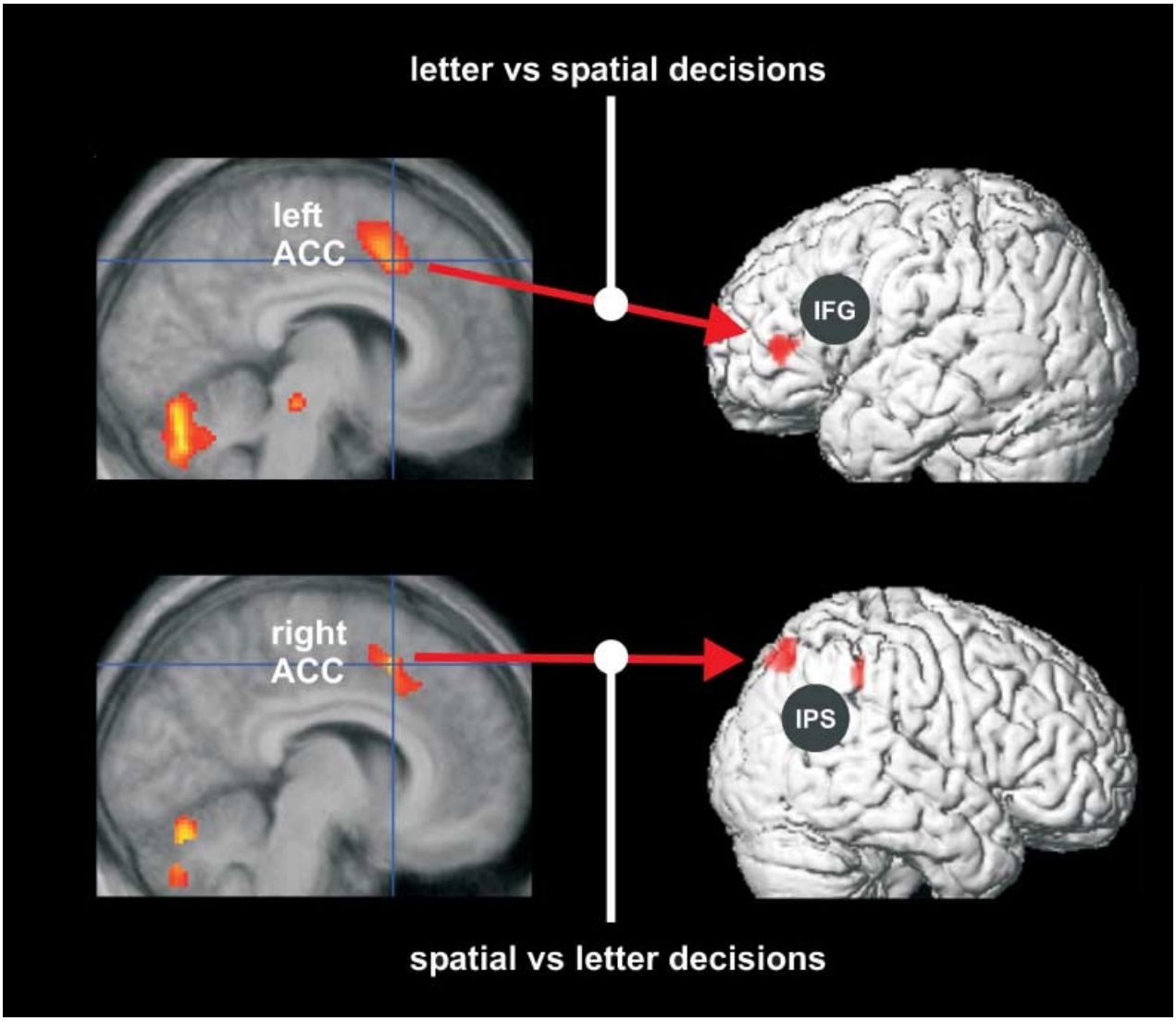


Fig. 5 Schematic summary of the results by Stephan et al. (2003). (A) Brain areas that were significantly activated during both letter and spatial decisions (contrast between the letter decision task and the baseline condition, masked by the contrast between the spatial decision task and the baseline condition; $P < 0.05$ cluster-level corrected). The anterior cingulate cortex was bilaterally activated during both conditions. Coordinates of the local maxima (left ACC: –6/16/42; right ACC: 8/16/48; see cross-hairs) refer to the space defined by the Montreal Neurological Institute (MNI) and differ marginally from those reported by Stephan et al. (2003) since they resulted from a re-analysis of the data by a different software package (SPM2). (B) Results from an analysis of effective connectivity of the left ACC using psycho–physiological interactions (PPIs) with SPM99. Left ACC specifically increased its coupling with left inferior frontal gyrus during letter decisions ($P < 0.05$, small-volume corrected). (C) Results from an analysis of effective connectivity of the right ACC using PPIs (SPM99). Right ACC specifically increased its coupling with anterior and posterior parts of right intraparietal sulcus during spatial decisions ($P < 0.05$, small-volume corrected).

buttons, whereas in a colour task, they could select one of four colours (and communicate this choice by button press). As a control, both tasks were supplemented by conditions in which the response was externally specified. In the conventional GLM analysis the DLPFC showed higher activity during free than externally specified selections, regardless of modality. Examination of the interactions between the two factors 'selection' and 'modality' revealed that there was no prefrontal region that was specifically engaged in action selection only or colour selection only. However, using a simple SEM of the putative neural system including DLPFC, motor, parietal and prestriate areas, DLPFC connectivity was found to be significantly modulated by modality: during action selection, the DLPFC influence on the motor cortex increased, whereas during colour selection,

DLPFC connectivity to prestriate areas (including the putative V4 region) increased. In addition, the modulation of connection strengths by modality was in itself modulated by the selection factor, e.g. the increase of the prefrontal–prestriate connection strength during the colour task was larger during free than during externally specified selection. Again, as in the examples above, the nature and topography of these complex top-down effects could not have been inferred from the GLM analysis but required a proper system model that allowed us to assess context-specific changes in connectivity.

With the advent of DCM, more sophisticated models of top-down and bottom-up processes have become possible (for example see Mechelli et al. 2004; Penny et al. 2004). Another example from ongoing work at our laboratory is given in Fig. 4, which shows how different types of top-down processes, i.e. gain control effects and baseline shifts, can be modelled using DCM. This model also demonstrates an issue highlighted by Penny et al. (2004): one has to be careful with apparent analogies between different levels of system modelling and consider what mechanisms at the neural level are actually represented by certain model components. For example, in DCM, the modulation of a forward connection (from a hierarchically lower to a hierarchically higher area) can both model a bottom-up or a top-down process at the neural level, depending on the nature of the modulatory factor (compare figure 1 in Penny et al. 2004 with Fig. 4 here).

Finally, I would like to comment on one particularly promising application of system models, i.e. the characterization of drug effects on connectivity. Given that many drugs used in psychiatry and neurology change synaptic transmission and thus functional coupling between neurons, a full understanding of their therapeutic effects cannot be achieved without models of how these drugs change the connectivity in neural systems of interest. So far, relatively few studies have studied pharmacologically induced changes in connectivity, ranging from simple analyses of functional connectivity (e.g. Stephan et al. 2001a) to proper system models, mainly based on SEM (e.g. Honey et al. 2003). As highlighted in a recent review by Honey & Bullmore (2004), one particularly exciting option for the future is to use system models at the early stage of drug development in order to screen for substances that induce desired changes of connectivity in neural systems that are reasonably well understood.

8. Future clinical applications of neuroimaging-based system modelling

The rise of explicit system models in neuroimaging represents the beginning of a merging of the field with traditional modelling approaches in computational neuroscience. It can be expected that this trend will be considerably reinforced and accelerated during the next few years, fuelled by the need for mechanistic explanations of how cognition is mediated by neural systems and by the availability of more powerful modelling techniques. One particular line of progress is expected in the domain of MEG and EEG where neural mass models of measured responses will be able to exploit the temporal resolution of these techniques in order to analyse synchronization and coherence phenomena that are, at best, only indirectly accessible by fMRI (Robinson et al. 2001; Breakspear et al. 2003, 2004; David & Friston, 2003). Another important extension will be to join approaches that use predictions from computational models (e.g. temporal difference learning models, O'Doherty et al. 2003) as regressors in conventional GLM analyses with system models based on connectivity. One of the most promising developments in this context is the formulation of predictive coding models. These models combine anatomical specificity (allowing for representation of different neural subpopulations, different types of connections and potentially different receptor types) with a precise model of local neural computations. Although previous implementations of predictive coding models have referred to more-or-less abstract neural systems (see Rao & Ballard, 1999; Lee & Mumford, 2003), ongoing work combines these models with modality-specific forward models that make it possible to fit them to measured EEG/MEG or fMRI data (Friston, 2004).

A particularly exciting possibility is that these advanced models may once be used as diagnostic tools in a clinical context. This option seems particularly attractive for psychiatric diseases whose phenotypes are often confusingly heterogeneous due to strong interactions between genotype and environmental influences. One hope is that we may find disease-specific endophenotypes, i.e. biological markers at intermediate levels between genome and behaviour (e.g. particular neurophysiological, neurochemical or endocrinological signatures). Such specific markers, if found, could allow for more precise categorization of patients and help to bridge the two distant levels of genetics

and behaviour (Gottesman & Gould, 2003). The endophenotype concept postulates that if a given psychiatric disease is indeed a homogeneous entity, its biological cause must be expressed at the level of a particular structure–function relation in the brain. Given the lack of focal structural changes in almost all psychiatric diseases, the biological cause therefore must reside in the dysfunctional structure of a particular neural system, i.e. in its connectivity. This 'disconnection hypothesis', which has received particular attention in the field of schizophrenia research (Friston, 1998), has been investigated in various forms by a series of imaging studies (e.g. Friston et al. 1996; Stephan et al. 2001a; Lawrie et al. 2002). Although robust connectivity differences have been reported by these studies for schizophrenic patients at the population level, connectivity parameters in classic system models like SEM have so far proved to be a poor predictor of genetic risk at the individual level (Winterer et al. 2003). More promising results have recently been obtained in research on major depression where an SEM, fitted to PET data, has been presented in which a few parameters were sufficient to distinguish patients who responded to pharmacotherapy from those patients who responded to behavioural therapy (Seminowicz et al. 2004).

The challenge will therefore be to establish neural systems models that are sensitive enough that their connectivity parameters can be used reliably for the diagnostic classification and treatment response prediction of individual patients. Ideally, such models should be used in conjunction with protocols that are minimally dependent on patient compliance and are not confounded by differences in performance, e.g. mismatch negativity protocols (Baldeweg et al. 2004). Given established validity and sufficient sensitivity of such a model, one could use it in analogy to a biochemical laboratory test in internal medicine, i.e. to compare a particular model parameter (or combinations thereof) against a reference distribution derived from a healthy population. Such procedures could help to decompose current psychiatric entities like schizophrenia into subgroups that are characterized by common SFRs in the brain and may facilitate the search for genetic underpinnings.

Acknowledgements

This article is based on a talk given at a Symposium of the Anatomical Society of Great Britain and Ireland in January 2004, entitled 'Functional anatomy of the human brain', organized by John Marshall. I would like to thank several of my colleagues for many stimulating discussions that have shaped my view on system analysis over the years, particularly Karl Friston, Lee Harrison, Claus Hilgetag, Rolf Kötter, Will Penny and Malcolm Young. Furthermore, I am grateful to Nancy Andreasen, Gereon Fink, John Marshall, Dick Passingham, James Rowe and Karl Zilles for giving me the opportunity to contribute to some of the neuroimaging experiments referred to in this article. I would also like to thank two reviewers for their helpful comments. This work was supported by a Travelling Research Fellowship (grant number 069468/Z/02/Z) from the Wellcome Trust to the author.

References

Absher JR, Benson DF (1993) Disconnection syndromes: An overview of Geschwind's contributions. *Neurology* **43**, 862–867.

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

Albright TD, Stoner GR (2002) Contextual influences on visual processing. *Ann. Rev. Neurosci.* **25**, 339–379.

Amunts K, Malikovic A, Mohlberg M, Schormann T, Zilles K (2000) Brodmann's Areas 17 and 18 brought into stereotaxic space – where and how variable? *Neuroimage* **11**, 66–84.

Ashby WR (1956) *An Introduction to Cybernetics*. London: Chapman & Hall.

Baldeweg T, Klugman A, Gruzelier J, Hirsch SR (2004) Mismatch negativity potentials and cognitive impairment in schizophrenia. *Schizophr. Res.* **69**, 203–217.

Bar-Yam Y (1997) *Dynamics of Complex Systems*. Reading, MA: Addison-Wesley.

von Bertalanffy L (1950) An outline of General System Theory. *Br. J. Philos. Sci.* **1**, 1389–1164.

von Bertalanffy L (1969) *General System Theory*. New York: George Braziller.

Bodke ALW, Tagamets MA, Friedman RB, Horwitz B (2001) Functional interactions of the inferior frontal cortex during the processing of words and word-like stimuli. *Neuron* **30**, 609–617.

Bollen KA (1989) *Structural Equations with Latent Variables*. New York: John Wiley.

von Bonin G, Bailey P (1947) *The Neocortex of Macaca Mulatta*. Urbana: University of Illinois Press.

Bossel H (1992) *Modellbildung und Simulation*. Braunschweig: Vieweg.

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: Comput. Neural Syst.* **14**, 703–732.

Breakspear M, Williams LM, Stam CJ (2004) A novel method for the topographic analysis of neural activity reveals formation

and dissolution of 'Dynamic Cell Assemblies'. *J. Comput. Neurosci*. **16**, 49–68.

Brodmann K (1909) *Vergleichende Lokalisationslehre der Grosshirnrinde in Ihren Prinzipien Dargestellt Auf Grund Des Zellenbaues*. Leipzig: Barth.

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

Büchel C, Friston KJ (1998) Dynamic changes in effective connectivity characterized by variable parameter regression and Kalman filtering. *Hum. Brain Mapp*. **6**, 403–408.

Büchel C, Coull JT, Friston KJ (1999) The predictive value of changes in effective connectivity for human learning. *Science* **283**, 1538–1541.

Bullmore E, Horwitz B, Honey G, Brammer M, Williams S, Sharma T (2000) How good is good enough in path analysis of fMRI data? *Neuroimage* **11**, 289–301.

Chawla D, Rees G, Friston KJ (1999) The physiological basis of attentional modulation in extrastriate visual areas. *Nat. Neurosci*. **2**, 671–676.

Chong L, Ray LB (2002) Whole-istic biology. *Science* **295**, 1661.

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

Dayan P, Abott LF (2001) *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

Dolcos F, LaBar KS, Cabeza R (2004) Interaction between the amygdala and the medial temporal lobe memory system predicts better memory for emotional events. *Neuron* **42**, 855–863.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47.

Fink GR, Halligan PW, Marshall JC, Frith CD, Frackowiak RS, Dolan RJ (1996) Where in the brain does visual attention select the forest and the trees? *Nature* **382**, 626–628.

Friston KJ, Jezzard PJ, Turner R (1994) Analysis of functional MRI time-series. *Hum. Brain Mapp*. **1**, 153–171.

Friston KJ (1995) Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp*. **2**, 56–78.

Friston KJ, Frith CD, Fletcher P, Liddle PF, Frackowiak RS (1996) Functional topography: multidimensional scaling and functional connectivity in the brain. *Cereb. Cortex* **6**, 156–164.

Friston KJ, Büchel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229.

Friston KJ (1998) The disconnection hypothesis. *Schizophr. Res*. **30**, 115–125.

Friston KJ (2000) The labile brain. I. Neuronal transients and nonlinear coupling. *Phil. Trans. R. Soc. Lond. B Biol. Sci*. **355**, 215–236.

Friston KJ, Büchel C (2000) Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proc. Natl Acad. Sci. USA* **97**, 7591–7596.

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) Beyond phrenology: what can neuroimaging tell us abut distributed circuitry? *Ann. Rev. Neurosci*. **25**, 221–250.

Friston KJ (2003) Learning and inference in the brain. *Neural Netw*. **16**, 1325–1352.

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. *Neuroimage* **19**, 1273–1302.

Friston KJ (2004) A theory of cortical responses. *Proc. R. Soc. Lond. B, in Press*.

Friston KJ, Büchel C (2004) Functional connectivity: eigenimages and multivariate analyes. In *Human Brain Function*, 2nd edn (eds Frackowiack R et al.), pp. 999–1018. New York: Elsevier.

Gitelman DR, Penny WD, Ashburner J, Friston KJ (2003) Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage* **19**, 200–207.

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. *Magn. Reson. Imaging* **21**, 1251–1261.

Gold BT, Buckner RL (2002) Common prefrontal regions coactivate with dissociable posterior regions during controlled semantic and phonological tasks. *Neuron* **35**, 803–812.

Gottesman II, Gould TD (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* **160**, 636–645.

Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438.

Grill-Spector K, Knouf N, Kanwisher N (2004) The fusiform face area subserves face perception, not generic within-category identification. *Nat. Neurosci*. **7**, 555–562.

Hamzei F, Rijntjes M, Dettmers C, Glauche V, Weiller C, Büchel C (2003) The human action recognition system and its relationship to Broca's area: an fMRI study. *Neuroimage* **19**, 637–644.

Handwerker DA, Ollinger JM, D'Esposito M (2004) Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* **21**, 1639–1651.

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

Henson RN (2004) Analysis of fMRI time series: Linear time-invariant models, event-related fMRI, and optimal experimental design. In *Human Brain Function*, 2nd edn (eds Frackowiack R et al.), pp. 793–823. New York: Elsevier.

Honey GD, Fu CHY, Kim J, et al. (2002) Effects of verbal working memory load on corticocortical connectivity modeled by path analysis of functional magnetic resonance imaging data. *Neuroimage* **17**, 573–582.

Honey GD, Suckling J, Zelaya F, et al. (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 Pharmacol. Sci*. **25**, 366–374.

Horwitz B, Duara R, Rapoport SI (1984) Intercorrelations of glucose metabolic rates between brain regions: application to healthy males in a state of reduced sensory input. *J. Cereb. Blood Flow Metab*. **4**, 484–499.

Horwitz B, Rumsey JM, Donohue BC (1998) Functional connectivity of the angular gyrus in normal reading and dyslexia. *Proc. Natl Acad. Sci. USA* **95**, 8939–8944.

Horwitz B, Tagamets BA, McIntosh AR (1999) Neural modelling, functional brain imaging, and cognition. *Trends Cogn. Sci.* **3**, 91–98.

Hupe JM, James AC, Girard P, Lomber SG, Payne BR, Bullier J (2001) Feedback connections act on the early part of the responses in monkey visual cortex. *J. Neurophysiol.* **85**, 134–145.

Hupe JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J (1998) Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* **394**, 794–797.

Ishai A, Ungerleider LG, Haxby JV (2000) Distributed neural systems for the generation of visual images. *Neuron* **28**, 979–990.

Jenner AR, Rosen GD, Galaburda AM (1999) Neuronal asymmetries in primary visual cortex of dyslexic and nondyslexic brains. *Ann. Neurol.* **46**, 189–196.

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

Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* **22**, 751–761.

Kleinbaum DG, Kupper LL, Muller KE (1988) *Applied Regression Analysis and Other Multivariable Methods*. Belmont, CA: Duxbury Press.

Kosslyn SM (1999) If neuroimaging is the answer, what is the question? *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **354**, 1283–1294.

Kötter R, Stephan KE (1997) Useless or helpful? The 'limbic system' concept. *Rev. Neurosci.* **8**, 139–146.

Kötter R, Nielsen P, Johnsen D, Sommer FT, Northoff G (2002) Multi-level neuron and network modeling in computational neuroanatomy. In *Computational Neuroanatomy: Principles and Methods* (ed. Ascoli G), pp. 359–382. Totowa, NJ: Humana Press.

Lawrie SM, Büchel C, Whalley HC, Frith CD, Friston KJ, Johnstone EC (2002) Reduced frontotemporal functional connectivity in schizophrenia associated with auditory hallucinations. *Biol. Psychiatry* **51**, 1008–1011.

LeDoux JE (1991) Emotion and the limbic system concept. *Concepts Neurosci.* **2**, 169–199.

Lee TS (2003) Computations in the early visual cortex. *J. Physiol. Paris* **97**, 121–139.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.* **20**, 1434–1448.

Levy J, Trevarthen C (1976) Metacontrol of hemispheric function in human split-brain patients. *J. Exp. Psychol. Hum. Percept. Perform.* **2**, 299–312.

Li W, Piech V, Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. *Nat. Neurosci.* **7**, 651–657.

Lomber SG, Payne BR, Hilgetag CC, Rushmore J (2002) Restoration of visual orienting into a cortically blind hemifield by reversible deactivation of posterior parietal cortex or the superior colliculus. *Exp. Brain Res.* **142**, 463–474.

Luck SJ, Chelazzi L, Hillyard SA, Desimone R (1997) Neural mechanisms of spatial selective attention in areas V1, V2 and V4 of macaque visual cortex. *J. Neurophysiol.* **77**, 24–42.

Luks TL, Simpson GV, Feiwell RJ, Miller WL (2002) Evidence for anterior cingulate cortex involvement in monitoring preparatory attentional set. *Neuroimage* **17**, 792–802.

Lund JS (2002) Specificity and non-specificity of synaptic connections within mammalian visual cortex. *J. Neurocytol.* **31**, 203–209.

Manjaly ZM, Marshall JC, Stephan KE, Gurd JM, Zilles K, Fink GR (2003) In search of the hidden: an fMRI study with implications for the study of patients with autism and with acquired brain injury. *Neuroimage* **19**, 674–683.

Manjaly ZM, Marshall JC, Stephan KE, Gurd JM, Zilles K, Fink GR (2004) Context–dependent interactions of left posterior inferior frontal gyrus in a local visual search task unrelated to language. *Cogn. Neuropsychol.* in press.

Marrelec G, Benali H, Ciuciu P, Pelegrini-Issac M, Poline JB (2003) Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Hum. Brain Mapp.* **19**, 1–17.

McIntosh AR, Gonzalez-Lima F (1991) Metabolic activation of the rat visual system by patterned light and footshock. *Brain Res.* **547**, 295–302.

McIntosh AR, Grady CL, Ungerleider LG, Haxby JV, Rapoport SI, Horwitz B (1994) Network analysis of cortical visual pathways mapped with PET. *J. Neurosci.* **14**, 655–666.

McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* **3**, 143–157.

McIntosh AR, Rajah MN, Lobaugh NJ (1999) Interactions of prefrontal cortex in relation to awareness in sensory learning. *Science* **284**, 1531–1533.

McIntosh AR (2000) Towards a network theory of cognition. *Neural Netw.* **13**, 861–870.

McIntosh AR, Rajah MN, Lobaugh NJ (2003) Functional connectivity of the medial temporal lobe relates to learning and awareness. *J. Neurosci.* **23**, 6520–6528.

Mechelli A, Penny WD, Price CJ, Gitelman DR, Friston KJ (2002) Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities. *Neuroimage* **17**, 1459–1469.

Mechelli A, Price CJ, Friston KJ, Ishai A (2004) Where bottom-up meets top–down: neuronal interactions during perception and imagery. *Cereb. Cortex* **14**, 1256–1265.

Mehta AD, Ulbert I, Schroeder CE (2000) Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas. *Cereb. Cortex* **10**, 343–358.

Miller KL, Luh WM, Liu TT, et al. (2001) Nonlinear temporal dynamics of the cerebral blood flow response. *Hum. Brain Mapp.* **13**, 1–12.

Miyazawa A, Fujiyoshi Y, Unwin N (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**, 949–955.

Moore T, Armstrong KM (2003) Selective gating of visual signals by microstimulation of frontal cortex. *Nature* **421**, 370–373.

Motter BC (1993) Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70**, 909–919.

Murray SO, Wojciulik E (2004) Attention increases neural selectivity in the human lateral occipital complex. *Nat. Neurosci.* **7**, 70–74.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337.

Olson IR, Chun MM, Allison T (2001) Contextual guidance of attention: human intracranial event-related potential evidence for feedback modulation in anatomically early, temporally late stages of visual processing. *Brain* **124**, 1417–1425.

Passingham RE, Stephan KE, Kötter R (2002) The anatomical basis of functional localization in the cortex. *Nature Rev. Neurosci*. **3**, 606–616.

Payne BR, Lomber SG, Geeraerts S, van der Gucht E, Vandenbussche E (1996) Reversible visual hemineglect. *Proc. Natl Acad. Sci. USA* **93**, 290–294.

Payne BR, Lomber SG (2001) Reconstructing functional systems after lesions of cerebral cortex. *Nat. Rev. Neurosci*. **2**, 911–919.

Penny WD, Roberts SJ (2002) Bayesian multivariate autoregressive models with structured priors. *IEEE Proc. Vis. Image Signal Proc*. **149**, 33–41.

Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. *Neuroimage* **22**, 1157–1172.

Phillips CG, Zeki S, Barlow HB (1984) Localisation of function in the cerebral cortex. Past present and future. *Brain* **107**, 327–361.

Pitt MA, Myung IJ (2002) When a good fit can be bad. *Trends Cogn. Sci*. **6**, 421–425.

Price CJ, Friston KJ (2002) Degeneracy and cognitive anatomy. *Trends Cogn. Sci*. **6**, 416–421.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci*. **2**, 79–87.

Rieke F, Warland D, Ruyter van Steveninck R, Bialek W (1997) *Spikes. Exploring the Neural Code*. Cambridge, MA: MIT Press.

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

Rowe JB, Stephan KE, Friston KJ, Frackowiak RJ, Lees A, Passingham RE (2002) Attention to action in Parkinson's disease. Impaired effective connectivity among frontal cortical regions. *Brain* **125**, 276–289.

Rowe JB, Stephan KE, Friston KJ, Frackowiak RS, Passingham RE (2004) The prefrontal cortex shows context-specific changes in effective connectivity to motor or visual cortex during the selection of action or colour. *Cereb. Cortex* in press.

Seminowicz DA, Mayberg HS, McIntosh AR, et al. (2004) Limbic-frontal circuitry in major depression: a path modeling metanalysis. *Neuroimage* **22**, 409–418.

Sergent J (1983) Role of the input in visual hemispheric asymmetries. *Psychol. Bull*. **93**, 481–512.

Seymour B, O'Doherty JP, Dayan P, et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667.

Sprague JM (1966) Interaction of cortex and superior colliculus in mediation of visually guided behavior in the cat. *Science* **153**, 1544–1547.

Stephan KE, Magnotta VA, White TJ, et al. (2001a) Effects of Olanzapine on cerebellar functional connectivity in schizophrenia measured by fMRI during a simple motor task. *Psychol. Med*. **31**, 1065–1078.

Stephan KE, Kamper L, Bozkurt A, Burns GAPC, Young MP, Kötter R (2001b) Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Phil. Trans. R. Soc. Lond. B Biol. Sci*. **356**, 1159–1186.

Stephan KE, Marshall JC, Friston KJ, et al. (2003) Lateralized cognitive processes and lateralized task control in the human brain. *Science* **301**, 384–386.

Stephan KE, Harrison LM, Penny WD, Friston KJ (2004) Biophysical models of fMRI responses. *Curr. Opin. Neurobiol*. **14**, 629–635.

Vogt C, Vogt O (1919) Ergebnisse unserer Hirnforschung. Vierte Mitteilung: Die physiologische Bedeutung der architektonischen Rindenreizungen. *J. Psychol. Neurol*. **25**, 279–461.

Wiener N (1948) *Cybernetics*. New York: Wiley.

Winterer G, Coppola R, Egan MF, Goldberg TE, Weinberger DR (2003) Functional and effective frontotemporal connectivity and genetic risk for schizophrenia. *Biol. Psychiatry* **54**, 1181–1192.

Yamashita O, Galka A, Ozaki T, Biscay R, Valdes-Sosa P (2004) Recursive penalized least squares solution for dynamical inverse problems of EEG generation. *Hum. Brain Mapp*. **21**, 221–235.

Young MP (1992) Objective analysis of the topological organization of the primate cortical visual system. *Nature* **358**, 152–155.

Young MP, Hilgetag CC, Scannell JW (2000) On imputing function to structure from the behavioural effects of brain lesions. *Phil. Trans. R. Soc. Lond. B Biol. Sci*. **355**, 147–161.

Zilles K, Palomero-Gallagher N, Grefkes C, et al. (2002) Architectonics of the human cerebral cortex and transmitter receptor fingerprints: reconciling functional neuroanatomy and neurochemistry. *Eur. Neuropsychopharmacol*. **12**, 587–599.