Functional integration in the brain

Karl J Friston

The Wellcome Dept. of Cognitive Neurology,
University College London
Queen Square, London, UK WC1N 3BG
Tel (44) 020 7833 7456
Fax (44) 020 7813 1445
email k.friston@fil.ion.ucl.ac.uk

Contents

I INTRODUCTION

This section is about functional integration in the brain. This chapter introduces the neurobiological background of functional integration, in terms of neuronal information processing in cortical hierarchies. This serves to frame the sorts of question than can be addressed with analyses of functional and effective connectivity. In fact, we take the empirical Bayesian theory described in the previous chapter as a possible basis for understanding integration among the levels of hierarchically organised cortical systems. The next two chapters (**Chapter 19 and 20**) deal with the fundaments of functional and effective connectivity, that are revisited in the next two chapters. **Chapter 21 and 22** deal with two complementary perspectives on models of functional integration, namely the Volterra or generalised convolution formulation and the state-space representation used by Dynamic Causal Modelling. In the final chapter we reconcile various approaches, looking more closely at the underlying mathematics.

Self-supervised models of how the brain represents and categorises the causes of its sensory input can be divided into those that minimise the mutual information (*i.e.* redundancy) among evoked responses and those that minimise the prediction error. This chapter describes one such model and its implications for the functional anatomy of sensory cortical hierarchies in the brain. We then consider how analyses of effective connectivity can be used to look for architectures that are sufficient for perceptual learning and synthesis.

Many models of representational learning require prior assumptions about the distribution of sensory causes. However, as seen in the previous chapter, the notion of empirical Bayes, suggests that these assumptions are not necessary and that priors can be learned in a hierarchical context. The main point made in this chapter is that backward connections, mediating internal or generative models of how sensory inputs are caused, are essential and that feedforward architectures, on their own, are not sufficient. Moreover, nonlinearities in generative models require these connections to be modulatory so that estimated causes in higher cortical levels can interact to predict responses in lower levels. This is important in relation to functional asymmetries in forward and backward connections that have been demonstrated empirically.

To ascertain whether backward influences are expressed functionally requires measurements of functional integration among brain systems. This chapter summarises approaches to integration in terms of functional and effective connectivity and uses the theoretical considerations above to illustrate the sorts of questions that can be addressed.

Specifically, it will be shown that functional neuroimaging can be used to test for interactions between bottom-up and top-down inputs to an area.

In concert with the growing interest in contextual and extra-classical receptive field effects in electrophysiology (*i.e.* how the receptive fields of sensory neurons change according to the context a stimulus is presented in), a similar paradigm shift is emerging in imaging neuroscience. Namely, the appreciation that functional specialisation exhibits similar extra-classical phenomena; in which a cortical area may be specialised for one thing in one context but something else in another. These extra-classical phenomena have implications for theoretical ideas about how the brain might work. This chapter uses theoretical models of representational learning as a vehicle to illustrate how imaging can be used to address important questions about functional brain architectures.

We start by reviewing two fundamental principles of brain organisation, namely *functional specialisation* and *functional integration* and how they rest upon the anatomy and physiology of cortico-cortical connections in the brain. The second section deals with the nature and learning of representations from a theoretical or computational perspective. The key focus of this section is on the functional architectures implied by the model. Generative models based on predictive coding rest on hierarchies of backward and lateral projections and, critically, confer a necessary role on backward connections.

Empirical evidence, from electrophysiological studies of animals and functional neuroimaging studies of human subjects, is presented in the third and fourth sections to illustrate the context-sensitive nature of functional specialisation and how its expression depends upon integration among remote cortical areas. The third section looks at extra-classical effects in electrophysiology, in terms of the predictions afforded by generative models of brain function. The theme of context-sensitive evoked responses is generalised to a cortical level and human functional neuroimaging studies in the subsequent section. The critical focus of this section is evidence for the interaction of bottom-up and top-down influences in determining regional brain responses. These interactions can be considered signatures of backward connections. The final section reviews some of the implications of the forging sections for lesion studies and neuropsychology. *Dynamic diaschisis* is described, in which aberrant neuronal responses can be observed as a consequence of damage to distal brain areas providing enabling or modulatory afferents. This section uses neuroimaging in neuropsychological patients and discusses the implications for constructs based on the lesion-deficit model.

II FUNCTIONAL SPECIALISATION AND INTEGRATION

A Background

The brain appears to adhere to two fundamental principles of functional organisation, functional integration and functional specialisation, where the integration within and among specialised areas is mediated by effective connectivity. The distinction relates to that between 'localisationism' and '[dis]connectionism' that dominated thinking about cortical function in the nineteenth century. Since the early anatomic theories of Gall, the identification of a particular brain region with a specific function has become a central theme in neuroscience. However functional localisation *per se* was not easy to demonstrate: For example, a meeting that took place on August 4th 1881 addressed the difficulties of attributing function to a cortical area, given the dependence of cerebral activity on underlying connections (Phillips *et al* 1984). This meeting was entitled "Localisation of function in the cortex cerebri". Goltz, although accepting the results of electrical stimulation in dog and monkey cortex, considered that the excitation method was inconclusive, in that the behaviours elicited might have originated in related pathways, or current could have spread to distant centres. In short, the excitation method could not be used to infer functional localisation because localisationism discounted interactions, or functional integration among different brain areas. It was proposed that lesion studies could supplement excitation experiments. Ironically, it was observations on patients with brain lesions some years later (see Absher and Benson 1993) that led to the concept of 'disconnection syndromes' and the refutation of localisationism as a complete or sufficient explanation of cortical organisation. Functional localisation implies that a function can be localised in a cortical area, whereas specialisation suggests that a cortical area is specialised for some aspects of perceptual or motor processing, where this *specialisation* can be anatomically *segregated* within the cortex. The cortical infrastructure supporting a single function may then involve many specialised areas whose union is mediated by the functional integration among them. Functional specialisation and integration are not exclusive, they are complementary. Functional specialisation is only meaningful in the context of functional integration and *vice versa.*

B Functional specialisation and segregation

The functional role, played by any component (*e.g.* cortical area, sub-area, neuronal population or neuron) of the brain, is defined largely by its connections. Certain patterns of cortical projections are so common that they could amount to rules of cortical connectivity. "These rules revolve around one, apparently, overriding strategy that the cerebral cortex uses - that of functional segregation" (Zeki 1990). Functional segregation demands that cells with common functional properties be grouped together. This architectural constraint in turn necessitates both convergence and divergence of cortical connections. Extrinsic connections, between cortical regions, are not continuous but occur in patches or clusters. This patchiness has, in some instances, a clear relationship to functional segregation. For example, the secondary visual area V2 has a distinctive cytochrome oxidase architecture, consisting of thick stripes, thin stripes and inter-stripes. When recordings are made in V2, directionally selective (but not wavelength or colour selective) cells are found exclusively in the thick stripes. Retrograde (*i.e.* backward) labelling of cells in V5 is limited to these thick stripes. All the available physiological evidence suggests that V5 is a functionally homogeneous area that is specialised for visual motion. Evidence of this nature supports the notion that patchy connectivity is the anatomical infrastructure that underpins functional segregation and specialisation. If it is the case that neurons in a given cortical area share a common responsiveness (by virtue of their extrinsic connectivity) to some sensorimotor or cognitive attribute, then this functional segregation is also an anatomical one. Challenging a subject with the appropriate sensorimotor attribute or cognitive process should lead to activity changes in, and only in, the areas of interest. This is the model upon which the search for regionally specific effects with functional neuroimaging is based.

C The anatomy and physiology of cortico-cortical connections

If specialisation rests upon connectivity then important organisational principles should be embodied in the neuroanatomy and physiology of extrinsic connections. Extrinsic connections couple different cortical areas whereas intrinsic connections are confined to the cortical sheet. There are certain features of cortico-cortical connections that provide strong clues about their functional role. In brief, there appears to be a hierarchical organisation that rests upon the distinction between *forward* and *backward* connections. The designation of a connection as forward or backward depends primarily on its cortical layers of origin and termination. Some characteristics of cortico-cortical connections are presented below and are summarised in Table 1. The list is not exhaustive, nor properly qualified, but serves to

introduce some important principles that have emerged from empirical studies of visual cortex.

- *Hierarchical organisation*

The organisation of the visual cortices can be considered as a hierarchy of cortical levels with reciprocal extrinsic cortico-cortical connections among the constituent cortical areas (Felleman and Van Essen 1991). The notion of a hierarchy depends upon a distinction between reciprocal forward and backward extrinsic connections.

- *Reciprocal connections*

Although reciprocal, forward and backward connections show both a microstructural and functional asymmetry. The terminations of both show laminar specificity. Forwards connections (from a low to a high level) have sparse axonal bifurcations and are topographically organised; originating in supragranular layers and terminating largely in layer VI. Backward connections, on the other hand, show abundant axonal bifurcation and a more diffuse topography. Their origins are bilaminar/infragranular and they terminate predominantly in supragranular layers (Rockland and Pandya 1979, Salin and Bullier 1995). Extrinsic connections show an orderly convergence and divergence of connections from one cortical level to the next. At a macroscopic level, one point in a given cortical area will connect to a region 5 - 8mm in diameter in another. An important distinction between forward and backward connections is that backward connections are more divergent. For example, the divergence region of a point in V5 (*i.e.* the region receiving backward afferents from V5) may include thick and inter-stripes in V2 whereas its convergence region (*i.e.* the region providing forward afferents to V5) is limited to the thick stripes (Zeki and Shipp 1988). Backward connections are more abundant then forward connections and transcend more levels. For example the ratio of forward efferent connections to backward afferents in the lateral geniculate is about 1:10/20. Another important distinction is that backward connections will traverse a number of hierarchical levels whereas forward connections are more restricted. For example, there are backward connections from TE and TEO to V1 but no monosynaptic connections from V1 to TE or TEO (Salin and Bullier 1995).

- *Functionally asymmetric forward and backward connections*

Functionally, reversible inactivation (*e.g.* Sandell and Schiller 1982, Girard and Bullier 1988) and neuroimaging (*e.g.* Büchel and Friston 1997) studies suggest that forward connections are driving, always eliciting a response, whereas backward connections can also be modulatory. In this context, modulatory means backward connections modulate responsiveness to other inputs. The notion that forward connections are concerned with the promulgation and segregation of sensory information is consistent with; (i) their sparse axonal bifurcation, (ii) patchy axonal terminations, (iii) and topographic projections. In contradistinction, backward connections are generally considered to have a role in mediating contextual effects and in the co-ordination of processing channels. This is consistent with; (i) their frequent bifurcation, (ii) diffuse axonal terminations (iii) and more divergent topography (Salin and Bullier 1995, Crick and Koch 1998). Forward connections meditate their postsynaptic effects through fast AMPA (1.3-2.4ms decay) and $GABA_A$ (6ms decay) receptors. Modulatory effects can be mediated by NMDA receptors. NMDA receptors are voltage-sensitive, showing nonlinear and slow dynamics (~50ms decay). They are found predominantly in supragranular layers where backward connections terminate (Salin and Bullier 1995). These slow time-constants again point to a role in mediating contextual effects that are more enduring than phasic sensory-evoked responses.

There are many mechanisms that are responsible for establishing connections in the brain. Connectivity results from interplay between genetic, epigenetic and activity- or experience-dependent mechanisms. I*n utero*, epigenetic mechanisms predominate; such as the interaction between the topography of the developing cortical sheet, cell migration, gene expression and the mediating role of gene-gene interactions and gene products such as cell adhesion molecules (CAMs). Following birth, connections are progressively refined and re-modelled with a greater emphasis on activity- and use-dependent plasticity. These changes endure into adulthood with ongoing reorganisation and experience-dependent plasticity that subserves behavioural adaptation and learning throughout life. In brief, there are two basic determinants of connectivity. (i) *Structural plasticity*, reflecting the interactions between the molecular biology of gene expression, cell migration and neurogenesis in the developing brain. (ii) *Synaptic plasticity*: Activity-dependent modelling of the pattern and strength of synaptic connections. This plasticity involves changes in the form, expression and function of synapses that endure throughout life. Plasticity is an important functional attribute of connections in the brain and is thought to subserve perceptual and procedural learning and memory. A key aspect of this plasticity is that it is generally associative.

- *Associative plasticity*

  Synaptic plasticity may be transient (*e.g.* short-term potentiation STP or depression STD) or enduring (*e.g.* long-term potentiation LTP or LTD) with many different time constants. In contrast to short-term plasticity, long-term changes rely on protein synthesis, synaptic remodelling and infrastructural changes in cell processes (*e.g.* terminal arbours or dendritic spines) that are mediated by calcium-dependent mechanisms. An important aspect of NMDA receptors, in the induction of LTP, is that they confer associatively on changes in connection strength. This is because their voltage-sensitivity only allows calcium ions to enter the cell when there is conjoint presynaptic release of glutamate and sufficient post-synaptic depolarisation (*i.e.* the temporal association of pre- and postsynaptic events). Calcium entry renders the post-synaptic specialisation eligible for future potentiation by promoting the formation of synaptic 'tags' (*e.g.* Frey and Morris 1998) and other calcium dependant intracellular mechanisms.

In summary, the anatomy and physiology of cortico-cortical connections suggest that forward connections are driving and commit cells to a pre-specified response given the appropriate pattern of inputs. Backward connections, on the other hand, are less topographic and are in a position to modulate the responses of lower areas to driving inputs from either higher or lower areas (see Table 1). For example, in the visual cortex Angelucci *et al* (2002a) used a combination of anatomical and physiological recording methods to determine the spatial scale and retinotopic logic of intra-areal V1 horizontal connections and inter-areal feedback connections to V1. "Contrary to common beliefs, these [monosynaptic horizontal] connections cannot fully account for the dimensions of the surround field [of macaque V1 neurons]. The spatial scale of feedback circuits from extrastriate cortex to V1 is, instead, commensurate with the full spatial range of centre-surround interactions. Thus these connections could represent an anatomical substrate for contextual modulation and global-to-local integration of visual signals."

Brain connections are not static but are changing at the synaptic level all the time. In many instances this plasticity is associative. Backwards connections are abundant in the brain and are in a position to exert powerful effects on evoked responses, in lower levels, that define the specialisation of any area or neuronal population. Modulatory effects imply the post-synaptic response evoked by presynaptic input is modulated, or interacts with, another. By definition this interaction must depend on nonlinear synaptic or dendritic mechanisms.

D Functional integration and effective connectivity

Electrophysiology and imaging neuroscience have firmly established functional specialisation as a principle of brain organisation in man. The functional integration of specialised areas has proven more difficult to assess. Functional integration refers to the interactions among specialised neuronal populations and how these interactions depend upon the sensorimotor or cognitive context. Functional integration is usually assessed by examining the correlations among activity in different brain areas, or trying to explain the activity in one area in relation to activities elsewhere. *Functional connectivity* is defined as correlations between remote neurophysiological events[1]. However, correlations can arise in a variety of ways. For example, in multi-unit electrode recordings they can result from stimulus-*locked* transients evoked by a common input or reflect stimulus-*induced* oscillations mediated by synaptic connections (Gerstein and Perkel 1969). Integration within a distributed system is better understood in terms of *effective connectivity*. Effective connectivity refers explicitly to the influence that one neuronal system exerts over another, either at a synaptic (*i.e.* synaptic efficacy) or population level. It has been proposed that "the [electrophysiological] notion of 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 and Preißl 1991). This speaks to two important points: (i) Effective connectivity is dynamic, *i.e.* activity- and time-dependent and (ii) it depends upon a model of the interactions. An important distinction, among models employed in functional neuroimaging, is whether these models are linear or nonlinear. Recent characterisations of effective connectivity have focussed on nonlinear models that accommodate the modulatory or nonlinear effects mentioned above. A more detailed discussion of these models is provided in subsequent chapters, after the motivation for their application is established below. In this chapter, the terms modulatory and nonlinear are used almost synonymously. Modulatory effects imply the post-synaptic response evoked by one input is modulated, or interacts with, another. By definition this interaction must depend on nonlinear synaptic mechanisms.

In summary, the brain can be considered as an ensemble of functionally specialised areas that are coupled in a nonlinear fashion by effective connections. Empirically, it appears that connections from lower to higher areas are predominantly driving whereas backwards connections, that mediate top-down influences, are more diffuse and are capable of exerting modulatory influences. In the next section we describe a theoretical perspective, provided by

[1] More generally any statistical dependency as measured by the mutual information

'generative models', that highlights the functional importance of backwards connections and nonlinear interactions.

III REPRESENTATIONAL LEARNING

This section describes the heuristics behind self-supervised learning based on *empirical Bayes.* This approach is considered within the framework of *generative models* and follows Dayan and Abbott (pp359-397, 2001) to which the reader is referred for more detailed background. A more heuristic discussion of these issues can be found in Friston (2002)

An important focus of this section is the interaction among causes of sensory input. These interactions create a problem of contextual invariance. In brief, it will be shown that this problem points to the adoption of generative models where interactions among causes of a percept are modelled explicitly in backward connections. First, we will reprise empirical Bayes in the context of brain function *per se*. Having established the requisite architectures for representational learning, neuronal implementation is considered in sufficient depth to make predictions about the anatomical and functional anatomy that would be needed to implement empirical Bayes in the brain. We conclude by relating theoretical predictions with the four neurobiological principles listed in the previous section.

A The nature of inputs, causes and representations

Here a representation is taken to be a neuronal event that represents some 'cause' in the sensorium. Causes are simply the states of processes generating sensory data or input. It is not easy to ascribe meaning to these states without appealing to the way that we categorise things, perceptually or conceptually. High-level conceptual causes may be categorical in nature, such as the identity of a face in the visual field or the semantic category a perceived object belongs to. In a hierarchical setting, high-level causes may induce priors on lower-level causes that are more parametric in nature. For example, the perceptual cause "moving quickly" may show a one-to-many relationship with representations of different velocities in V5 (MT) units. Causes have relationships to each other (*e.g.* 'is part of') that often have a hierarchical structure. This hierarchical ontology is attended by ambiguous many-to-one and one-to-many mappings (*e.g.* a table has legs but so do horses; a wristwatch is a watch

irrespective of the orientation of its hands). This ambiguity can render the problem of inferring causes from sensory information under-determined or ill posed.

Even though causes may be difficult to describe they are easy to define operationally. Causes are the variables or states that are necessary to specify the products of a process generating sensory information. To keep things simple, let us frame the problem of representing causes in terms of a deterministic nonlinear generative function.

$$u = G(v,\theta) \qquad 1$$

where v is a vector of underlying causes in the environment (*e.g.* the velocity of a particular object, direction of radiant light *etc*), and u represents some sensory inputs. $G(v,\theta)$ is a function that generates inputs from the causes. Nonlinearities in Eq(1) represent interactions among the causes. Second-order interactions are formally identical to interaction terms in conventional statistical models of observed data. These can often be viewed as contextual effects, where the expression of a particular cause depends on the context established by another. For example, the extraction of motion from the visual field depends upon there being sufficient luminance or wavelength contrast to define the surface moving. Another ubiquitous example, from early visual processing, is the occlusion of one object by another. In the absence of interactions we would see a linear superposition of both objects but the visual input, caused by the nonlinear mixing of these two causes, render one occluded by the other. At a more cognitive level the cause associated with the word 'HAMMER' will depend on the semantic context (that determines whether the word is a verb or a noun). These contextual effects are profound and must be discounted before the representations of the underlying causes can be considered veridical.

The problem the brain has to contend with is to find a function of the input that recognises or represents the underlying causes. To do this, the brain must effectively undo the interactions to disclose contextually invariant causes. In other words, the brain must perform some form of nonlinear unmixing of causes and context without knowing either. The key point here is that this nonlinear mixing may not be invertible and that the estimation of causes from input may be fundamentally ill posed. For example, no amount of unmixing can discern the parts of an object that are occluded by another. The mapping $u = v^2$ provides a trivial example of this non-invertibility. Knowing u does not uniquely determine v. The corresponding indeterminacy, in probabilistic learning, rests on the combinatorial explosion of ways in which stochastic generative models can generate input patterns (Dayan *et al*

1995). The combinatorial explosion represents another example of the uninvertible 'many to one' relationship between causes and inputs.

In probabilistic learning one allows for stochastic components in the generation of inputs and recognising a particular cause becomes probabilistic. Here the issue of deterministic invertibility is replaced by the existence of an inverse conditional probably [*i.e.* recognition] density that can be parameterised. Although not a mathematical fundament, parameterisation is critical for the brain because it has to encode the parameters of these densities with biophysical attributes of its nervous tissue. In what follows we consider the implications of this problem. In brief, we will show that one needs separate [approximate] recognition and generative models that induces the need for both forward and backward influences. Separate recognition and generative models resolve the problem caused by generating processes that are difficult to invert and speak to a possible role for backward connections in the brain.

B Generative models and representational learning

Generative models afford a generic formulation of representational leaning in a supervised or self-supervised context. There are many forms of generative models that range from conventional statistical models (*e.g.* factor and cluster analysis) and those motivated by Bayesian inference and learning (*e.g.* Dayan *et al* 1995, Hinton *et al* 1995). The goal of generative models is "to learn representations that are economical to describe but allow the input to be reconstructed accurately" (Hinton *et al* 1995). Representational learning is framed in terms of estimating probability densities of the causes. This is referred to as posterior density analysis in the estimation literature and posterior mode analysis if the inference is restricted to estimating the most likely cause (See **Chapter 17: Classical and Bayesian Inference**). Although density learning is formulated at a level of abstraction that eschews many issues of neuronal implementation *(e.g.* the dynamics of real-time learning), it provides a unifying framework that connects the various schemes considered below.

Figure 2 about here

1 Inference vs. learning

Equation (1) relates the unknown state of the causes v and some unknown parameters θ, to observed inputs u. The objective is to make *inferences* about the causes and *learn* the parameters. Inference may be simply estimating the most likely state of the causes and is based on the products of learning. A useful way of thinking about the distinction between

inference and learning is in terms of how one accounts for the patterns or distribution of inputs encountered. Figure 1 shows a very simply example with a univariate cause and a bivariate observation. Observations are denoted by dots in the right hand panel and cluster around a curvilinear line. A parsimonious way of generating dots like these would be move up and down the line and add a small amount of observation error. The position on the line corresponds to the state of the single cause and the probability of selecting a particular position to the probability density of the causes on the right. *Inference* means ascertaining the probability of each potential cause given an observation. *Estimation* refers to estimating the most likely cause, denoted in Figure 1 by $\dot{v}$. This estimate is the closest point on the line to the observation that *a priori* has a reasonable probability of being selected. This simple example introduces the notion of representing observations in terms of points that lie on a low dimensional manifold in observation space, in this case a line. The dimensions of this manifold are the causes. The shape and position of the manifold depends on the parameters θ. These have to be known or learned before inference about any particular observation can proceed. This learning requires multiple observations so that the manifold can be placed to transect the highest density of observations. In short, representational learning can be construed as learning a low dimensional manifold onto which data can be projected with minimum loss of information. This manifold is an essential component of generative models.

The goal of learning is to acquire a recognition model for inference that is effectively the inverse of a generative model. Learning a generative model corresponds to making the density of the inputs, implied by a generative model $p(u;\theta)$, as close as possible to those observed $p(u)$. The generative model is specified in terms of a *prior* distribution over the causes $p(v;\theta)$ and the *generative* distribution or likelihood of the inputs given the causes $p(u \mid v;\theta)$. Together, these define the marginal distribution that has to be matched to the input distribution

$$p(u;\theta) = \int p(u \mid v;\theta)\, p(v;\theta) dv \qquad 2$$

See Figure 1. Once the parameters of the generative model have been learned, through this matching, the posterior density of the causes, given the inputs are given by the recognition model, which is defined in terms of the *recognition* distribution

$$p(v \mid u;\theta) = \frac{p(u \mid v;\theta)\, p(v;\theta)}{p(u;\theta)} \qquad 3$$

However, as considered above, the generative model may not be easily inverted and it may not be possible to parameterise the recognition distribution. This is crucial because the endpoint of learning is the acquisition of a useful recognition model that can be applied to sensory inputs. One solution is to posit an approximate recognition distribution $q(v;u,\phi)$ that is consistent with the generative model and that can be learned at the same time. The approximate recognition distribution has some parameters ϕ, for example, the strength of forward connections or its mode (*i.e.* most likely value). The first question addressed in this section is whether forward connections are sufficient for representational leaning.

C Density estimation and EM

In density learning, representational learning has two components that are framed in terms of expectation maximisation (EM, Dempster *et al* 1977). Iterations of an **E**-Step ensure the recognition approximates the inverse of the generative model and the **M**-Step ensures that the generative model can predict the observed inputs. Probabilistic recognition proceeds by using $q(v;u,\phi)$ to determine the probability that v caused the observed sensory inputs. EM provides a useful procedure for density estimation that helps relate many different models within a framework that has direct connections with statistical mechanics. Both steps of the EM algorithm involve maximising a function of the densities that corresponds to the negative free energy in physics.

$$\begin{aligned} F &= \left\langle l(u) \right\rangle_u \\ l &= \int q(v;u,\phi) \ln \frac{p(v,u;\theta)}{q(v;u,\phi)} dv \\ &= \left\langle \ln p(v,u;\theta) \right\rangle_q - \left\langle \ln q(v;u,\phi) \right\rangle_q \\ &= \ln p(u;\theta) - KL\{q(v;u,\phi), p(v \mid u;\theta)\} \end{aligned} \qquad 4$$

This objective function comprises two terms. The first is the expected log likelihood of the inputs under the generative model. The second term is the Kullback-Leibler (KL)

divergence[2] between the approximating and true recognition densities. Critically, the KL term is always positive, rendering F a lower bound on the expected log likelihood of the inputs. Maximising F encompasses two components of representational learning; (i) it increases the likelihood of the inputs produced by the generative model and (ii) minimises the discrepancy between the approximate recognition model and that implied by the generative model. The **E**-Step increases F with respect to the recognition parameters ϕ, ensuring a veridical approximation to the recognition distribution implied by the generative parameters θ. The **M**-Step changes θ, enabling the generative model to reproduce the inputs.

$$\begin{aligned} \mathbf{E} \qquad & \phi = \max_{\phi} F \\ \mathbf{M} \qquad & \theta = \max_{\theta} F \end{aligned} \tag{5}$$

There are a number of ways of motivating the free energy formulation in Eq(4). A useful one, in this context, rests upon the problem posed by non-invertible models. This problem is finessed by assuming it is sufficient to match the joint probability of inputs and causes under the generative model $p(u,v;\theta) = p(u \mid v;\theta)p(v;\theta)$ with that implied by recognising the causes of inputs encountered $p(u,v;\phi) = q(v;u,\phi)p(u)$. Both these distributions are well defined even when $p(v \mid u;\theta)$ is not easily parameterised. This matching minimises the divergence.

$$\begin{aligned} KL\{p(v,u;\phi), p(v,u;\theta)\} &= \int q(v;u,\phi)p(u)\ln\frac{q(v;u,\phi)p(u)}{p(v,u;\theta)}dvdu \\ &= -F - H(u) \end{aligned} \tag{6}$$

This is equivalent to maximising F because the entropy of the inputs $H(u)$ is fixed. This perspective is used in Figure 2 to illustrate the **E** and **M** steps schematically. The **E**-Step adjusts the recognition parameters to match the two joint distributions, while the **M**-Step does exactly the same thing but by changing the generative parameters. The dependency of the generative parameters, on the input distribution, is mediated vicariously in the **M**-Step through the recognition. In the setting of invertibility, where $q(v;u,\phi) = p(v \mid u;\theta)$ the divergence in Eq(6) reduces to $KL\{p(u), p(u;\theta)\}$. As above, the **M**-Step then finds

[2] a measure of the distance or difference between two probability densities

parameters that allow the model to simply match the observed input distribution (*i.e.* maximise the expected likelihood).

Figure 2 about here

1 Invertibility

This formulation of representational leaning is critical for the thesis of this section because it suggests that backward and lateral connections, parameterising a generative model, are essential when the model is not invertible. If the generative model is invertible then the KL term in Eq(4) can be discounted by setting $q(v;u,\phi) = p(v \mid u;\theta)$ with Eq(3) and learning reduces to the **M**-Step (*i.e.* maximising the expected likelihood).

$$F = \langle \ln p(u;\theta) \rangle_u \qquad 7$$

In principle, this could be done using a feedforward architecture corresponding to the inverse of the generative model. However, when processes generating inputs are non-invertible (in terms of the parameterisation of the recognition density) a generative model and approximate recognition model are required that are updated in **M**- and **E**-Steps respectively. In short, non-invertibility enforces an explicit parameterisation of the generative model in representational learning. In the brain this parameterisation may be embodied in backward connections.

2 Deterministic recognition

Another special case arises when the recognition is deterministic. The recognition becomes deterministic when $q(v;u,\phi)$ is a Dirac δ-function over its mode $v(u,\phi)$. In this instance, posterior density analysis reduces to a posterior mode analysis at which point inference and estimation coincide. They are equivalent in the sense that inferring the posterior distribution of causes is the same as estimating the most likely cause given the inputs (the maximum *a posteriori* or MAP estimator). Here the integral in Eq(4) disappears, leaving the joint probability of the inputs and their cause to be maximised

$$
\begin{aligned}
F &= \left\langle \ln p(v(u),u;\theta) \right\rangle_u \\
&= \left\langle \ln p(u \mid v(u);\theta) + \ln p(v(u);\theta) \right\rangle_u
\end{aligned}
\qquad 8
$$

Notice, again, that this objective function does not require $p(v \mid u;\theta)$ and eschews the inversion in Eq(3). An illustration of the **E**-Step for deterministic recognition is shown in Figure 4 (lower panel). Here, the distinction between deterministic and stochastic relates to inference and refers to form of the recognition density. It should be noted that learning could also employ a deterministic or stochastic ascent on *F*. We will deal largely with deterministic learning schemes.

3. Summary

EM enables exact and approximate maximum likelihood density estimation for a whole variety of generative models that can be specified in terms of prior and generative distributions. Dayan and Abbott (2001) work though a series of didactic examples from cluster analysis to independent component analyses, within this unifying framework. For example, factor analysis corresponds to the generative model

$$
\begin{aligned}
p(v;\theta) &= N(v:0,1) \\
p(u|v;\theta) &= N(u:\theta v,\Sigma)
\end{aligned}
\qquad 9
$$

Namely, the underlying causes of inputs are independent normal variates that are mixed linearly and added to Gaussian noise to form inputs. In the limiting case of $\Sigma \to 0$ the ensuing model become deterministic and conforms to PCA. By simply assuming non-Gaussian priors one can specify generative models for sparse coding of the sort proposed by Olshausen and Field (1996)

$$
\begin{aligned}
p(v;\theta) &= \prod p(v_i;\theta) \\
p(u|v;\theta) &= N(u:\theta v,\Sigma)
\end{aligned}
\qquad 10
$$

where $p(v_i;\theta)$ are chosen to be suitably sparse (*i.e.* heavy-tailed) with a cumulative density function that corresponds to the squashing function in **Chapter 19 (Functional connectivity**). The deterministic equivalent of sparse coding is ICA that obtains when $\Sigma \to 0$. The relationships among different models are rendered apparent under the

perspective of generative models. In what follows we consider a series of models entailing assumptions about the generation of sensory inputs that are relaxed one by one. At each point we consider whether they could be implemented plausibly in the brain.

Figure 3 about here

D Cortical hierarchies and empirical Bayes

Empirical Bayes harnesses the hierarchical structure of a generative model, treating the estimates at one level as prior expectations for the subordinate level (Efron and Morris 1973). This provides a natural framework within which to treat cortical hierarchies in the brain, each providing constraints on the level below. This approach models the world as a hierarchy of systems where supraordinate causes induce, and moderate, changes in subordinate causes. For example, the presence of a particular object in the visual field changes the incident light falling on a particular part of the retina. A more intuitive example is provided in Figure 3. These priors offer contextual guidance towards the most likely cause of the input. Note that predictions at higher levels are subject to the same constraints, only the highest level, if there is one in the brain, is free to be directed solely by bottom-up influences (although there are always implicit priors). If the brain has evolved to recapitulate the casual structure of its environment, in terms of its sensory infrastructures, it is interesting to reflect on the possibility that our visual cortices reflect the hierarchical casual structure of our environment.

1 The nature of hierarchical models

Consider any level *i* in a hierarchy whose causes v_i are induced by corresponding causes in the level above v_{i+1}. The hierarchical form of the implicit generative model is

$$\begin{aligned} u &= G_1(v_2, \theta_1) + \varepsilon_1 \\ v_2 &= G_2(v_3, \theta_2) + \varepsilon_2 \\ v_3 &= \ldots \end{aligned} \qquad 11$$

with $u = v_1$ *c.f.* Eq(1). Technically, these models fall into the class of conditionally independent hierarchical models when the stochastic terms are independent at each level (Kass and Steffey 1989). These models are also called *parametric empirical Bayes* (PEB) models because the obvious interpretation of the higher-level densities as priors led to the development of PEB methodology (Efron and Morris 1973). Often, in statistics, these

hierarchical models comprise just two levels, which is a useful way to specify simple shrinkage priors on the parameters of single-level models (see **Section II, Part II**). We will assume the stochastic terms are Gaussian with covariance $\Sigma_i = \Sigma(\lambda_i)$. Therefore, θ_i and λ_i parameterise the means and covariances of the likelihood at each level.

$$p(v_i|v_{i+1};\theta) = N(v_i : G_i(v_{i+1},\theta_i),\Sigma_i) \qquad 12$$

This likelihood of v_i *also plays the role of a prior on* v_i that is jointly maximised with the likelihood of the level below $p(v_{i-1}\,|\,v_i;\theta)$. This is key to understanding the utility of hierarchical models; By learning the parameters of the generative distribution of level *i* one is implicitly learning the parameters of the prior distribution for level $i-1$. This enables this learning of prior densities.

Figure 4 about here

The hierarchical nature of these models lends an important context-sensitivity to recognition densities not found in single-level models. This is illustrated in Figure 4, which should be compared with Figure 1. The key point here is that high-level causes v_{i+1} determine the prior expectation of causes v_i in the subordinate level. This can completely change the marginal $p(v_{i-1};\theta)$ and recognition $p(v_i\,|\,v_{i-1};\theta)$ distributions upon which inference in based. From the manifold perspective on inference, the part of the manifold $G_{i-1}(v_i;\theta_{i-1})$ highlighted by prior expectations, changes from input to input in a context-dependent way (see Figure 4). The context established by priors is not determined by preceding events but is immediate and conferred by higher hierarchical levels. For example, in Figure 3 the semantic context induced by reading one of the sentences has a profound effect on the most likely graphemic cause of the visual input subtended by 'ev'. The dual role of $p(v_i\,|\,v_{i+1};\theta)$ as a likelihood or generative density for level *i* and a prior density for level *i* - 1 is recapitulated by a dual role for MAP estimates of v_i. From a bottom-up perspective these correspond to parameters [modes] of the recognition densities. However, from a top-down perspective they also act as parameters of the generative model by interacting with θ_{i-1} in $G_{i-1}(v_i,\theta_{i-1})$ to give the prior expectation of v_{i-1}.

Although λ_i are parameters of the forward model we have referred to as hyperparameters in previous chapters and, in classical statistics, correspond to variance components. We will preserve the distinction between θ_i and λ_i because they may correspond to backward and lateral connections strengths respectively.

2 Implementation

The biological plausibility of the empirical Bayes in the brain can be established fairly simply. To do this a hierarchical scheme is described in some detail. For the moment, we will address neuronal implementation at a purely theoretical and somewhat heuristic level, using the framework developed above.

For simplicity, we will assume deterministic recognition such that $q(\phi(u);u) = 1$. In this setting, with conditional independence, *F* comprises a series of log likelihoods

$$\begin{aligned} l(u) &= \left\langle \ln p(u,v;\theta) \right\rangle_q = \ln p(u,\phi_2,\ldots;\theta) \\ &= \ln p(u \mid \phi_2;\theta) + \ln(\phi_2 \mid \phi_3;\theta) + \ldots \\ &= -\tfrac{1}{2}\xi_1^T\xi_1 - \tfrac{1}{2}\xi_2^T\xi_2 - \ldots - \tfrac{1}{2}\ln|\Sigma_1| - \tfrac{1}{2}\ln|\Sigma_2| - \ldots \end{aligned}$$

$$\begin{aligned} \xi_i &= \phi_i - G_i(\phi_{i+1},\theta_i) - \lambda_i\xi_i \\ &= (1+\lambda_i)^{-1}(\phi_i - G_i(\phi_{i+1},\theta_i)) \end{aligned} \qquad 13$$

Here $\Sigma_i^{1/2} = 1 + \lambda_i$. In the setting of neuronal models the [whitened] prediction error is encoded by the activities of units denoted by ξ_i. These error units receive a prediction from units in the level above[3] and connections from the principal units ϕ_i being predicted. Horizontal interactions among the error units serve to de-correlate them (*c.f.* Foldiak 1990), where the symmetric lateral connection strengths λ_i hyper-parameterise the covariances of the errors Σ_i, which are the prior covariances for level $i-1$.

The estimators ϕ_i and the connection strength parameters perform a gradient ascent on the compound log probability.

[3] Clearly, in the brain, backward connections are not inhibitory but, after mediation by inhibitory interneurons, their effective influence could be rendered so.

$$\mathbf{E} \qquad \dot{\phi}_{i+1} = \frac{\partial l(u)}{\partial \phi_{i+1}} = -\frac{\partial \xi_i^T}{\partial \phi_{i+1}}\xi_i - \frac{\partial \xi_{i+1}^T}{\partial \phi_{i+1}}\xi_{i+1}$$

$$\mathbf{M} \qquad \dot{\theta}_i = \frac{\partial F}{\partial \theta_i} = -\left\langle \frac{\partial \xi_i^T}{\partial \theta_i}\xi \right\rangle_u \qquad 14$$

$$\dot{\lambda}_i = \frac{\partial F}{\partial \lambda_i} = -\left\langle \frac{\partial \xi_i^T}{\partial \lambda_i}\xi \right\rangle_u - (1+\lambda_i)^{-1}$$

Each of the learning components has a relatively simple neuronal interpretation (see below)

Figure 5 about here

E Implications for neuronal implementation

The scheme implied by Eq(14) has four clear implications or predictions about the functional architectures required for its implementation. We now review these in relation to cortical organisation in the brain. A schematic summarising these points in provided in Figure 5. In short, we arrive at exactly the same four points presented in the previous section.

- *Hierarchical organisation*

  Hierarchical models enable empirical Bayesian learning of prior densities and provide a plausible model for sensory inputs. Single-level models that do not show any conditional independence (*e.g.* those used by connectionist and infomax schemes) depend on prior constraints for unique inference and do not call upon a hierarchical cortical organisation. On the other hand, if the causal structure of generative processes is hierarchical, this will be reflected, literally, by the hierarchical architectures trying to minimise prediction error, not just at the level of sensory input but at all levels (notice the deliberate mirror symmetry in Figure 5). The nice thing about this architecture is that the responses of units at the *i*th level ϕ_i depend only on the error for the current level and the immediately preceding level. This follows from conditional independence and is important because it permits a biologically plausible implementation, where the connections driving the error minimisation only run forward from one level to the next.

- *Reciprocal connections*

As established at the beginning of his section the non-invertibility of processes generating sensory data induces a need for both forward and backward connections. In the hierarchical model, the dynamics of principal units ϕ_{i+1} are subject to two, locally available, influences. A likelihood or recognition term mediated by forward afferents from the error units in the level below and an empirical prior conveyed by error units in the same level. Critically, the influences of the error units in both levels are meditated by linear connections with a strength that is exactly the same as the [negative] effective connectivity of the *reciprocal* connections from ϕ_{i+1} to ξ_i and ξ_{i+1}. Functionally, forward and lateral connections are reciprocated, where backward connections generate predictions of lower-level responses. Effective connectivity is simply the change in a neuronal unit (neuron, assembly or cortical area) induced by inputs from another (Friston 1995). In this case $\partial\xi_i/\partial\phi_{i+1}$ and $\partial\xi_{i+1}/\partial\phi_{i+1}$

Effective connectivity in the forward direction is the reciprocal (negative transpose) of that in the backward direction $\partial\xi_i/\partial\phi_{i+1} = -\partial G_i(\phi_{i+1},\beta_i)_i/\partial v_{i+1}$ that is a function of the generative parameters. Lateral connections, within each level, mediate the influence of error units on the principal units and intrinsic connections λ_i among the error units decorrelate them, allowing competition among prior expectations with different precisions (precision is the inverse of variance). In short, lateral, forwards and backward connections are all reciprocal, consistent with anatomical observations.

- *Functionally asymmetric forward and backward connections*

The forward connections are the reciprocal of the backward effective connectivity from the higher level to the lower level, extant at that time. However, the functional attributes of forward and backward influences are different. The influences of units ϕ_{i+1} on error units in the lower level ξ_i instantiate the forward model $\xi_i = \phi_i - G_i(\phi_{i+1},\theta_i) - \lambda_i\xi_i$. These can be nonlinear, where each unit in the higher level *may modulate or interact with the influence of others*, according to the nonlinearities in $G_i(\phi_{i+1},\theta_i)$. In contradistinction, the influences of units in lower levels do not interact when producing changes in the higher level because their effects are linearly separable [see Eq(27)]. This is a key observation because the empirical

evidence, reviewed in the previous section, suggests that backward connections are in a position to interact (*e.g.* though NMDA receptors expressed predominantly in the supragranular layers receiving backward connections). Forward connections are not. It should be noted that, although the implied forward connections $-\partial\xi_i/\partial\phi_{i+1}{}^T$ mediate linearly separable effects of ξ_i on ϕ_{i+1}, these connections might be activity- and time-dependent because of their dependence on ϕ_{i+1}. In summary, nonlinearities, in the way sensory inputs are produced, necessitate nonlinear interactions in the generative model that are mediated by backward influences but do not require forward connections to be modulatory.

- *Associative plasticity*

Changes in the parameters correspond to plasticity in the sense that the parameters control the strength of backward and lateral connections. The backward connections parameterise the prior expectations of the forward model and the lateral connections hyper-parameterise the prior covariances. Together they parameterise the Gaussian densities that constitute the priors (and likelihoods) of the model. The plasticity implied can be seen more clearly with an explicit parameterisation of the connections. For example, let $G_i(v_{i+1},\theta_i) = \theta_i v_{i+1}$. In this instance

$$\begin{aligned}\dot{\theta}_i &= (1+\lambda_i)^{-1}\left\langle \xi_i \phi_{i+1}^T \right\rangle_u \\ \dot{\lambda}_i &= (1+\lambda_i)^{-1}(\left\langle \xi_i \xi_i^T \right\rangle_u - 1)\end{aligned} \qquad 15$$

This is just Hebbian or associative plasticity where the connection strengths change in proportion to the product of pre and post-synaptic activity. An intuition about Eq(15) obtains by considering the conditions under which the expected change in parameters is zero (*i.e.* after learning). For the backward connections this implies there is no component of prediction error that can be explained by estimates at the higher level $\left\langle \xi_i \phi_{i+1}^T \right\rangle = 0$. The lateral connections stop changing when the prediction error has been whitened $\left\langle \xi_i \xi_i^T \right\rangle = 1$

It is evident that the predictions of the theoretical analysis coincide almost exactly with the empirical aspects of functional architectures in visual cortices highlighted by the previous section (hierarchical organisation, reciprocity functional asymmetry and associative

plasticity).. Although somewhat contrived, it is pleasing that purely theoretical considerations and neurobiological empiricism converge so precisely.

VI GENERATIVE MODELS AND THE BRAIN

In summary, generative models lend themselves naturally to a hierarchical treatment, which considers the brain as an empirical Bayesian device. The dynamics of the units or populations are driven to minimise prediction error at all levels of the cortical hierarchy and implicitly render themselves posterior modes of the causes given the data. The overall scheme implied by Eq. (14) sits comfortably the hypothesis (Mumford, 1992). "on the role of the reciprocal, topographic pathways between two cortical areas, one often a 'higher' area dealing with more abstract information about the world, the other 'lower', dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops".

A Context, causes and representations

The Bayesian perspective suggests something quite profound for the classical view of receptive fields. If neuronal responses encompass a bottom-up likelihood term and top-down priors, then responses evoked by bottom-up input should change with the context established by prior expectations from higher levels of processing. Consider the example in Figure 3. Here a unit encoding the visual form of 'went' responds when we read the first sentence at the top of this figure. When we read the second sentence 'The last event was cancelled' it would not. If we recorded from this unit we might infer that our 'went' unit was, in some circumstances, selective for the word 'event'. This might be difficult to explain without an understanding of hierarchical inference and the semantic context the stimulus was presented in. In short, under a predictive coding scheme, the receptive fields of neurons should be context-sensitive. The remainder of this subsection deals with empirical evidence for these extra-classical receptive field effects.

Generative models suggest that the role of backward connections is to provide contextual guidance to lower levels through a prediction of the lower level's inputs. When this prediction is incomplete or incompatible with the lower area's input, an error is generated that engenders changes in the area above until reconciliation. When, and only when, the bottom-up driving inputs are in harmony with top-down prediction, error is suppressed and a consensus between the prediction and the actual input is established. Given this conceptual model a stimulus-related response or 'activation' corresponds to some transient error signal that drives the appropriate change in higher areas until a veridical higher-level representation emerges and the error is 'cancelled' by backwards connections. Clearly the prediction error will depend on the context and consequently the backward connections confer context-sensitivity on the functional specificity of the lower area. In short, the activation does not just depend on bottom-up input but on the difference between bottom-up input and top-down predictions.

The prevalence of nonlinear or modulatory top-down effects can be inferred from the fact that context interacts with the content of representations. Here context is established simply through the expression of causes other than the one in question. Backward connections from one higher area can be considered as providing contextual modulation of the prediction from another area. Because the effect of context will only be expressed when the thing being predicted is present these contextual afferents should not elicit a response by themselves. Effects of this sort, which change the responsiveness of units but do not elicit a response, are a hallmark of modulatory projections. In summary, hierarchical models offer a scheme that allows for contextual effects; firstly through biasing responses towards their prior expectation and secondly by conferring a context-sensitivity on these priors through the modulatory component of backward projections. Next we consider the nature of real neuronal responses and whether they are consistent with this perspective.

B Extra-classical and context-sensitive effects

Classical models (*e.g.* classical receptive fields) assume that evoked responses will be expressed invariably in the same units or neuronal populations irrespective of the context. However, real neuronal responses are not invariant but depend upon the context in which they are evoked. For example, visual cortical units have dynamic receptive fields that can change from moment to moment [*c.f.* the non-classical receptive field effects modelled in (Rao and Ballard 1999)]. A useful synthesis of data for the macaque visual system that highlights the anatomical and physiological substrates of context-dependent responses can be

found in Angelucci *et al* (2002b). A key conclusion of the authors is that "feedback from extrastriate cortex (possibly together with overlap or interdigitation of coactive lateral connectional fields within V1) can provide a large and stimulus-specific surround modulatory field. The stimulus specificity of the interactions between the centre and surround fields, may be due to the orderly, matching structure and different scales of intra-areal and feedback projection excitatory pathways."

There are numerous examples of context-sensitive neuronal responses. Perhaps the simplest is short-term plasticity. Short-term plasticity refers to changes in connection strength, either potentiation or depression, following pre-synaptic inputs (*e.g.* Abbot 1997). In brief, the underlying connection strengths, that define what a unit represents, are a strong function of the immediately preceding neuronal transient (*i.e.* preceding representation). A second, and possibly richer, example is that of attentional modulation that can change the sensitivity of neurons to different perceptual attributes (*e.g.* Treue and Maunsell 1996). . It has been shown, both in single unit recordings in primates (Treue and Maunsell 1996) and human functional fMRI studies (Büchel and Friston 1997), that attention to specific visual attributes can profoundly alter the receptive fields or event-related responses to the same stimuli.

These sorts of effects are commonplace in the brain and are generally understood in terms of the dynamic modulation of receptive field properties by backward and lateral afferents. There is clear evidence that lateral connections in visual cortex are modulatory in nature (Hirsch and Gilbert 1991), speaking to an interaction between the functional segregation implicit in the columnar architecture of V1 and the neuronal dynamics in distal populations. These observations, suggests that lateral and backwards interactions may convey contextual information that shapes the responses of any neuron to its inputs (*e.g.* Kay and Phillips 1996, Phillips and Singer 1997) to confer on the brain the ability to make conditional inferences about sensory input. See also McIntosh (2000) who develops the idea from a cognitive neuroscience perspective "that a particular region in isolation may not act as a reliable index for a particular cognitive function. Instead, the *neural context* in which an area is active may define the cognitive function." His argument is predicated on careful characterisations of effective connectivity using neuroimaging.

C Conclusion

In conclusion, the representational capacity and inherent function of any neuron, neuronal population or cortical area in the brain is dynamic and context-sensitive. Functional integration, or interactions among brain systems, that employ driving (bottom-up) and backward (top-down) connections, mediate this adaptive and contextual specialisation. Most models of representational learning require prior assumptions about the distribution of causes. However, empirical Bayes suggests that these assumptions can be relaxed and that priors can be learned in a hierarchical context. We have tried to show that this hierarchical prediction can be implemented in brain-like architectures and in a biologically plausible fashion.

A key point, made above, is that backward connections, mediating internal or generative models of how sensory inputs are caused, are essential if the processes generating inputs are difficult to invert. This non-invertibility demands an explicit parameterisation of both the generative model (backward connections) and approximate recognition (forward connections). This suggests that feedforward architectures are not sufficient for representational learning or perception. Moreover, nonlinearities in generative models, that make backward connections necessary, require these connections to be modulatory; so that estimated causes in higher cortical levels can interact to predict responses in lower levels. This is important in relation to asymmetries in forward and backward connections that have been characterised empirically.

The arguments in this section were developed under hierarchical models of brain function, where high-level systems provide a prediction of the inputs to lower-levels. Conflict between the two is resolved by changes in the high-level representations, which are driven by the ensuing error in lower regions, until the mismatch is 'cancelled'. From this perspective the specialisation of any region is determined both by bottom-up driving inputs and by top-down predictions. Specialisation is therefore not an intrinsic property of any region but depends on both forward and backward connections with other areas. Because the latter have access to the context in which the inputs are generated they are in a position to modulate the selectivity or specialisation of lower areas. The implications for classical models (*e.g.* classical receptive fields in electrophysiology, classical specialisation in neuroimaging and connectionism in cognitive models) are severe and suggest these models may provide incomplete accounts of real brain architectures. On the other hand, representational learning, in the context of hierarchical generative models not only accounts for extra-classical

phenomena seen empirically but enforces a view of the brain as an inferential machine through its empirical Bayesian motivation.

V ASSESSING FUNCTIONAL ARCHITECTURES WITH BRAIN IMAGING

Clearly, it would be nice to demonstrate the existence of backward influences with neuroimaging. This is a slightly deeper problem than might be envisaged. This is because making causal inferences about effective connectivity is not straightforward (see Pearl 2000). It might be thought that showing regional activity was partially predicted by activity in a higher level would be sufficient to confirm the existence of backward influences, at least at a population level. The problem is that this statistical dependency does not permit any causal inference. Statistical dependencies could easily arise in a purely forward architecture because the higher level activity is predicated on activity in the lower level. One resolution of this problem is to perturb the higher level directly using transmagnetic stimulation or pathological disruptions (see below). However, discounting these interventions, one is left with the difficult problem of inferring backward influences, based on measures that could be correlated because of forward connections. Although there are causal modelling techniques that can address this problem we will take a simpler approach and note that interactions between bottom-up and top-down influences cannot be explained by a purely feedforward architecture. This is because the top-down influences have no access to the bottom-up inputs. An interaction, in this context, can be construed as an effect of backward connections on the driving efficacy of forward connections. In other words, the response evoked by the same driving bottom-up inputs depends upon the context established by top-down inputs. This interaction is used below simply as evidence for the existence of backward influences. There are instances of predictive coding that emphasises this phenomenon. For example, the "Kalman filter model demonstrates how certain forms of attention can be viewed as an emergent property of the interaction between top-down expectations and bottom-up signals" (Rao 1999).

The remainder of this chapter focuses on the evidence for these interactions. From the point of view of functionally specialised responses these interactions manifest as context-sensitive or contextual specialisation, where modality-, category- or exemplar-specific responses, driven by bottom up inputs are modulated by top-down influences induced by perceptual set. The first half of this section adopts this perceptive. The second part of this

section uses measurements of effective connectivity to establish interactions between bottom-up and top-down influences. All the examples presented below rely on attempts to establish interactions by trying to change sensory-evoked neuronal responses through putative manipulations of top-down influences. These include inducing independent changes in perceptual set, cognitive [attentional] set and, in the last section through the study of patients with brain lesions

A Context-sensitive specialisation

If functional specialisation is context-dependent then one should be able to find evidence for functionally-specific responses, using neuroimaging, that are expressed in one context and not in another. The first part of this section provides an empirical example. If the contextual nature of specialisation is mediated by backwards modulatory afferents then it should be possible to find cortical regions in which functionally-specific responses, elicited by the same stimuli, are modulated by activity in higher areas. The second example shows that this is indeed possible. Both of these examples depend on multifactorial experimental designs.

1 Multifactorial designs

Factorial designs combine two or more factors within a task or tasks. Factorial designs can be construed as performing subtraction experiments in two or more different contexts. The differences in activations, attributable to the effects of context, are simply the interaction. Consider an implicit object recognition experiment, for example naming (of the object's name or the non-object's colour) and simply saying "yes' during passive viewing of objects and non-objects. The factors in this example are implicit object recognition with two levels (objects *vs.* non-objects) and phonological retrieval (naming vs. saying "yes"). The idea here is to look at the interaction between these factors, or the effect that one factor has on the responses elicited by changes in the other. Noting that object-specific responses are elicited (by asking subjects to view objects relative to meaningless shapes), with and without phonological retrieval, reveals the factorial nature of this experiment. This 'two by two' design allows one to look specifically at the interaction between phonological retrieval and object recognition. This analysis identifies not regionally specific activations but regionally specific *interactions*. When we actually performed this experiment these interactions were evident in the left posterior, inferior temporal region and can be associated with the integration of phonology and object recognition (see Figure 6 and Friston *et al* 1996 for details). Alternatively this region can be thought of as expressing recognition-dependent

responses that are realised in, and only in, the context of having to name the object seen. These results can be construed as evidence of contextual specialisation for object-recognition that depends upon modulatory afferents [possibly from temporal and parietal regions] that are implicated in naming a visually perceived object. There is no empirical evidence in these results to suggest that the temporal or parietal regions are the source of this top-down influence but in the next example the source of modulation is addressed explicitly using psychophysiological interactions.

B Psychophysiological Interactions

Psychophysiological interactions speak directly to the interactions between bottom-up and top-down influences, where one is modelled as an experimental factor and the other constitutes a measured brain response. In an analysis of psychophysiological interactions one is trying to explain a regionally specific response in terms of an interaction between the presence of a sensorimotor or cognitive process and activity in another part of the brain (Friston *et al* 1997). The supposition here is that the remote region is the source of backward modulatory afferents that confer functional specificity on the target region. For example, by combining information about activity in the posterior parietal cortex, mediating attentional or perceptual set pertaining to a particular stimulus attribute, can we identify regions that respond to that stimulus when, and only when, activity in the parietal source is high? If such an interaction exists, then one might infer that the parietal area is modulating responses to the stimulus attribute for which the area is selective. This has clear ramifications in terms of the top-down modulation of specialised cortical areas by higher brain regions.

The statistical model employed in testing for psychophysiological interactions is a simple regression model of effective connectivity that embodies nonlinear (second-order or modulatory effects). As such, this class of model speaks directly to functional specialisation of a nonlinear and contextual sort. Figure 7 illustrates a specific example (see Dolan *et al* 1997 for details). Subjects were asked to view [degraded] faces and non-face (object) controls. The interaction between activity in the parietal region and the presence of faces was expressed most significantly in the right infero-temporal region not far from the homologous left infero-temporal region implicated in the object naming experiment above. Changes in parietal activity were induced experimentally by pre-exposure of the [un-degraded] stimuli before some scans but not others to prime them. The data in the right panel of Figure 7 suggests that the infero-temporal region shows face-specific responses, relative to non-face objects, when, and only when, parietal activity is high. These results can be

interpreted as a priming-dependent face-specific response, in infero-temporal regions that are mediated by interactions with medial parietal cortex. This is a clear example of contextual specialisation that depends on top-down effects.

C Effective connectivity

The previous examples, demonstrating contextual specialisation, are consistent with functional architectures implied by generative models. However, they do not provide definitive evidence for an interaction between top-down and bottom-up influences. In this subsection we look for direct evidence of these interactions using functional imaging. This rests upon being able to measure effective connectivity in a way that is sensitive to interactions among inputs. This requires a plausible model of coupling among brain regions that can accommodate nonlinear effects. We will illustrate the use of a model that is based on the Volterra expansion described in **Chapter 20** (**Effective Connectivity**) and expanded on in the subsequent chapter.

1 Nonlinear coupling among brain areas

Linear models of effective connectivity assume that the multiple inputs to a brain region are linearly separable. This assumption precludes activity-dependent connections that are expressed in one context and not in another. The resolution of this problem lies in adopting nonlinear models like the Volterra formulation that include interactions among inputs. These interactions can be construed as a context- or activity-dependent modulation of the influence that one region exerts over another (Büchel and Friston 1997). In the Volterra model, second order kernels model modulatory effects. Within these models the influence of one region on another has two components. (i) The direct or *driving* influence of input from the first (*e.g.* hierarchically lower) region, irrespective of the activities elsewhere and (ii) an activity-dependent, *modulatory* component that represents an interaction with inputs from the remaining (*e.g.* hierarchically higher) regions. These are mediated by the first and second order kernels respectively. The example provided in Figure 8 addresses the modulation of visual cortical responses by attentional mechanisms (*e.g.* Treue and Maunsell 1996) and the mediating role of activity-dependent changes in effective connectivity. This is the same example used in the introduction (**Chapter 1**) and in subsequent chapters.

The right panel in Figure 8 shows a characterisation of this modulatory effect in terms of the increase in V5 responses, to a simulated V2 input, when posterior parietal activity is zero

(broken line) and when it is high (solid lines). In this study subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes). The brain regions and connections comprising the model are shown in the upper panel. The lower panel shows a characterisation of the effects of V2 inputs on V5 and their modulation by posterior parietal cortex (PPC) using simulated inputs at different levels of PPC activity. It is evident that V2 has an activating effect on V5 and that PPC increases the responsiveness of V5 to these inputs. The insert shows all the voxels in V5 that evidenced a modulatory effect ($p<0.05$ uncorrected). These voxels were identified by thresholding statistical parametric maps of the F statistic testing for the contribution of second order kernels involving V2 and PPC while treating all other components as nuisance variables. The estimation of the Volterra kernels and statistical inference procedure is described in Friston and Büchel (2000).

This sort of result suggests that backward parietal inputs may be a sufficient explanation for the attentional modulation of visually evoked extrastriate responses. More importantly, they are consistent with the functional architecture implied by predictive coding because they establish the existence of functionally expressed backward connections. V5 cortical responses evidence an interaction between bottom-up input from early visual cortex and top-down influences from parietal cortex. In the final section the implications of this sort of functional integration are addressed from the point of view of the lesion-deficit model and neuropsychology.

VI. FUNCTIONAL INTEGRATION AND NEUROPSYCHOLOGY

If functional specialisation depends on interactions among cortical areas then one might predict changes in functional specificity in cortical regions that receive enabling or modulatory afferents from a damaged area. A simple consequence is that aberrant responses will be elicited in regions hierarchically below the lesion if, and only if, these responses depend upon inputs from the lesion site. However, there may be other contexts in which the region's responses are perfectly normal (relying on other, intact, afferents). This leads to the notion of a context-dependent region-specific abnormality, caused by, but remote from, a lesion (*i.e.* an abnormal response that is elicited by some tasks but not others). We have

referred to this phenomenon as 'dynamic diaschisis' (Price *et al* 2000). See **Section V**, (**Language and Semantics; Part I**) for a more psychologically finessed discussion.

A Dynamic diaschisis

Classical diaschisis, demonstrated by early anatomical studies and more recently by neuroimaging studies of resting brain activity, refers to regionally specific reductions in metabolic activity at sites that are remote from, but connected to, damaged regions. The clearest example is 'crossed cerebellar diaschisis' (Lenzi *et al* 1982) in which abnormalities of cerebellar metabolism are seen characteristically following cerebral lesions involving the motor cortex. Dynamic diaschisis describes the context-sensitive and task-specific effects that a lesion can have on the *evoked responses* of a distant cortical region. The basic idea behind dynamic diaschisis is that an otherwise viable cortical region expresses aberrant neuronal responses when, and only when, those responses depend upon interactions with a damaged region. This can arise because normal responses in any given region depend upon inputs from, and reciprocal interactions with, other regions. The regions involved will depend on the cognitive and sensorimotor operations engaged at any particular time. If these regions include one that is damaged, then abnormal responses may ensue. However, there may be situations when the same region responds normally, for instance when its dynamics depend only upon integration with undamaged regions. If the region can respond normally in some situations then forward driving components must be intact. This suggests that dynamic diaschisis will only present itself when the lesion involves a hierarchically equivalent or higher area.

1 An empirical demonstration

We investigated this possibility in a functional imaging study of four aphasic patients, all with damage to the left posterior inferior frontal cortex, classically known as Broca's area (see Figure 9 - upper panels). These patients had speech output deficits but relatively preserved comprehension. Generally functional imaging studies can only make inferences about abnormal neuronal responses when changes in cognitive strategy can be excluded. We ensured this by engaging the patients in an explicit task that they were able to perform normally. This involved a keypress response when a visually presented letter string contained a letter with an ascending visual feature (e.g.: h, k, l, or t). While the task remained constant, the stimuli presented were either words or consonant letter strings. Activations detected for words, relative to letters, were attributed to implicit word

processing. Each patient showed normal activation of the left posterior middle temporal cortex that has been associated with semantic processing (Price 1998). However, none of the patients activated the left posterior inferior frontal cortex (damaged by the stroke), or the left posterior inferior temporal region (undamaged by the stroke) (see Figure 4b). These two regions are crucial for word production (Price 1998). Examination of individual responses in this area revealed that all the normal subjects showed increased activity for words relative to consonant letter strings while all four patients showed the reverse effect. The abnormal responses in the left posterior inferior temporal lobe occurred even though this undamaged region lies adjacent and posterior to a region of the left middle temporal cortex that activated normally (see middle column of Figure 9b). Critically, this area thought to be involved in an earlier stage of word processing than the damaged left inferior frontal cortex (*i.e.* is hierarchically lower than the lesion). From these results we can conclude that, during the reading task, responses in the left basal temporal language area rely on afferent inputs from the left posterior inferior frontal cortex. When the first patient was scanned again, during an explicit semantic task, the left posterior inferior temporal lobe responded normally. The abnormal implicit reading related responses were therefore task-specific.

These results serve to illustrate the concept of dynamic diaschisis; namely the anatomically remote and context-specific effects of focal brain lesions. Dynamic diaschisis represents a form of functional disconnection where regional dysfunction can be attributed to the loss of enabling inputs from hierarchically equivalent or higher brain regions. Unlike classical or anatomical disconnection syndromes its pathophysiological expression depends upon the functional brain state at the time responses are evoked. Dynamic diaschisis may be characteristic of many regionally specific brain insults and may have implications for neuropsychological inference.

CONCLUSION

In conclusion, the representational capacity and inherent function of any neuron, neuronal population or cortical area in the brain is dynamic and context-sensitive. Functional integration, or interactions among brain systems, that employ driving (bottom up) and backward (top-down) connections, mediate this adaptive and contextual specialisation. A critical consequence is that hierarchically organised neuronal responses, in any given cortical area, can represent different things at different times. Although most models of

representational learning require prior assumptions about the distribution of causes; empirical Bayes suggests that these assumptions can be relaxed and that priors can be learned in a hierarchical context. We have tried to show that this hierarchical prediction based on can be implemented in brain-like architectures and in a biologically plausible fashion. The arguments in this chapter were developed under generative models of brain function, where higher-level systems provide a prediction of the inputs to lower-level regions. Conflict between the two is resolved by changes in the higher-level representations, which are driven by the ensuing error in lower regions, until the mismatch is 'cancelled'. From this perspective the specialisation of any region is determined both by bottom-up driving inputs and by top-down predictions. Specialisation is therefore not an intrinsic property of any region but depends on both forward and backward connections with other areas. Because the latter have access to the context in which the inputs are generated they are in a position to modulate the selectivity or specialisation of lower areas.

The emphasis on theoretical neurobiology has been used to expose the usefulness of being able to measure effective connectivity and the importance of modulatory or nonlinear coupling in the brain. These nonlinear aspects of effective connectivity will be a recurrent theme in subsequent chapters that discuss functional and effective connectivity from an operational point of view.

References

Abbot LF, Varela JA, Karmel Sen, and Nelson SB (1997) Synaptic depression and cortical gain control Science 275:220-223

Absher JR and Benson DF. (1993) Disconnection syndromes: an overview of Geschwind's contributions. Neurology 43:862-867

Aertsen A and Preißl H. (1991) Dynamics of activity and connectivity in physiological neuronal Networks. in Non Linear Dynamics and Neuronal Networks. Ed Schuster HG VCH publishers Inc. New York NY USA p281-302

Ballard DH, Hinton GE, Sejnowski TJ (1983) Parallel visual computation. Nature 306:21-6

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

Crick F and Koch C (1998) Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. Nature 391:245-250

Dayan P, Hinton GE and Neal RM (1995) The Helmholtz machine. Neural Computation 7:889-904

Dempster AP, Laird NM and Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Series B 39;1-38

Dolan RJ Fink GR Rolls E Booth M Holmes A Frackowiak RSJ Friston KJ (1997) How the brain learns to see objects and faces in an impoverished context Nature 389: 596-598

Efron B and Morris C (1973) Stein's estimation rule and its competitors – an empirical Bayes approach. J. Am. Stats. Assoc. 68:117-130

Felleman DJ and Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1:1-47

Friston KJ (1995) Functional and effective connectivity in neuroimaging: A synthesis Human Brain Mapping 2;56-78

Friston KJ Price CJ Fletcher P Moore C Frackowiak RSJ and Dolan RJ. (1996) The trouble with cognitive subtraction. NeuroImage 4:97-104

Friston KJ Büchel C Fink GR Morris J Rolls E and Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6:218-229

Friston KJ and Büchel C (2000) Attentional modulation of V5 in human Pro Natl Acad. Sci USA 97:7591-7596

Gerstein GL and Perkel DH. (1969) Simultaneously recorded trains of action potentials: Analysis and functional interpretation. Science 164: 828-830

Girard P and Bullier J. (1989) Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. J Neurophysiol. 62:1287-1301

Hinton GE Dayan P Frey BJ and Neal RM (1995) The "Wake-Sleep" algorithm for unsupervised neural networks. Science 268;1158-1161

Hirsch JA and Gilbert CD. (1991) Synaptic physiology of horizontal connections in the cat's visual cortex. J. Neurosci. 11:1800-1809

Kass RE and Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407:717-726

Kay J and Phillips WA (1996) Activation functions, computational goals and learning rules for local processors with contextual guidance. Neural Computation 9:895-910

Lenzi, GL, Frackowiak, R.S.J., Jones, T. (1982) Cerebral oxygen metabolism and blood flow in human cerebral ischaemic infarction. J. Cereb. Blood Flow and Metab. 2: 321-335

Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern 66:241-51

McIntosh AR (2000) Towards a network theory of cognition. Neural Networks 13:861-870

Olshausen BA and Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607-609

Pearl J (2000) Causality, Models, reasoning and inference. Cambridge University Press, UK.

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

Phillips WA and Singer W (1997) In search of common foundations for cortical computation. Behavioural and Brain Sciences. 20:57-83

Price, CJ (1998) The functional anatomy of word comprehension and production. Trends. Cog. Sci. 2:281-288.

Price CJ Warburton EA, Moore, CJ, Frackowiak RSJ, and Friston KJ. (2000) Dynamic Diaschisis: Anatomically remote and context-specific human brain lesions. Journal of Cognitive Neuroscience 00:00-00

Rao RP & Ballard DH (1998) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. Nature Neuroscience 2, 79-87

Rao RP (1999). An optimal estimation approach to visual perception and learning. Vision Res. 39:1963-89

Rockland K. S. and Pandya D. N. (1979) Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain-Res. 179: 3-20

Salin P-A and Bullier J (1995) Corticocortical connections in the visual system: Structure and function. Psychol. Bull. 75:107-154

Sandell JH and Schiller PH (1982) Effect of cooling area 18 on striate cortex cells in the squirrel monkey. J. Neurophysiol. 48:38-48

Treue S and Maunsell HR. (1996) Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382: 539-41

Zeki S and Shipp S (1988) The functional logic of cortical connections. Nature 335:311-317

Zeki S. (1990) The motion pathways of the visual cortex. in "Vision: coding and efficiency" (C Blakemore Ed.) Cambridge University Press UK p321-345

Figure 1

Schematic of a simple model with a univariate cause and a bivariate observation. Observations are denoted by dots in the right hand panel and cluster around a curvilinear line. A parsimonious way of generating dots like these would be move up and down the line and add a small amount of random error. The position on the line corresponds to the state of the single cause and the probability of selecting a particular position the probability density of the causes on the right.

Figure 2

Schematic illustrating the two components of EM. In the **E**-Step the joint distribution of causes and inputs under the recognition model changes to approximate that under the generative model. This refines the recognition model. In the **M**-Step the joint distribution under the generative model changes to approximate that under the recognition model. This reduces the difference between the distribution of inputs implied by the generative model and that observed.

Figure 3

Schematic illustrating the role of priors in biasing towards one representation of an input or another. Upper panel: On reading the first sentence 'Jack and Jill went up the hill' we perceive the word 'event' as 'went' despite the fact it is 'event' (as in the second sentence). However, in the absence of any hierarchical inference the best explanation for the pattern of visual stimulation incurred by the text is the grapheme 'ev'. This would correspond to the maximum likelihood estimate and would be the most appropriate in the absence of prior information, from the lexical and semantic context, about which is the most likely grapheme. However, within hierarchical inference the semantics (provided by the sentence) provide top-down predictions about the word, which in turn predicts the graphemes and finally the visual input. The posterior estimate is accountable to all these levels. When the semantic prior biases in favour of 'went' and 'w' we tolerate a small error as a lower level of visual analysis to minimise the overall prediction error. Lower panel: (left) The grapheme 'ev' is selected as the most likely cause of visual input. (right) The letter 'w' is selected, as it is (i) a reasonable explanation for the sensory input and (ii) conforms to prior expectations induced by lexico-semantic context. The bars represent prediction error, which is minimised over all levels to attain the most likely cause.

Figure 4

Hierarchical models embody context-sensitivity not found in single-level models (*c.f.* Figure 1). High-level causes v_{i+1} determine the prior expectation of causes v_i in the subordinate level. Changes in v_{i+1} can completely change the marginal $p(v_{i-1};\theta)$ and recognition $p(v_i \mid v_{i-1};\theta)$ distributions upon which inference in based.

Figure 5

Upper panel: Schematic depicting a hierarchical extension to the predictive coding architecture. Hierarchical arrangements within the model serve to provide predictions or priors to representations in the level below. The open circles are the error units and the filled circles are the states encoding the conditional expectation of causes in the environment. These change to minimise both the discrepancies between their predicted value and the mismatch incurred by their own prediction of the level below. These two constraints correspond to prior and likelihood terms respectively (see main text). Lower panel: A more detailed picture of the influences on principal and error units.

Figure 6

This example of regionally specific interactions comes from an experiment where subjects were asked to view coloured non-object shapes or coloured objects and say "yes", or to name either the coloured object or the colour of the shape. Left: A regionally specific interaction in the left infero-temporal cortex. The SPM threshold is $p < 0.05$ (uncorrected). Right: The corresponding activities in the maxima of this region are portrayed in terms of object recognition-dependent responses with and without naming. It is seen that this region shows object recognition responses when, and only when, there is phonological retrieval. The 'extra' activation with naming corresponds to the interaction. These data were acquired from six subjects scanned 12 times using PET.

Figure 7

Top: Examples of the stimuli presented to subjects. During the measurement of brain responses only degraded stimuli where shown (*e.g.* the right hand picture). In half the scans the subject was given the underlying cause of these stimuli, through presentation of the original picture (*e.g.* left) before scanning. This priming induced a profound difference in perceptual set for the primed, relative to non-primed, stimuli,

Right: Activity observed in a right infero-temporal region, as a function of [mean corrected] PPC activity. This region showed the most significant interaction between the presence of faces in visually presented stimuli and activity in a reference location in the posterior medial parietal cortex (PPC). This analysis can be thought of as finding those areas that are subject to top-down modulation of face-specific responses by medial parietal activity. The crosses correspond to activity whilst viewing non-face stimuli and the circles to faces. The essence of this effect can be seen by noting that this region differentiates between faces and non-faces when, and only when, medial parietal activity is high. The lines correspond to the best second-order polynomial fit. These data were acquired from six subjects using PET. Left: Schematic depicting the underlying conceptual model in which driving afferents from ventral form areas (here designated as V4) excite infero-temporal (IT) responses, subject to permissive modulation by PPC projections.

Figure 8

Upper panel: Brain regions and connections comprising the model. Lower panel: Characterisation of the effects of V2 inputs on V5 and their modulation by posterior parietal cortex (PPC). The broken lines represent estimates of V5 responses when PPC activity is zero, according to a second order Volterra model of effective connectivity with inputs to V5 from V2, PPC and the pulvinar (PUL). The solid curves represent the same response when PPC activity is one standard deviation of its variation over conditions. It is evident that V2 has an activating effect on V5 and that PPC increases the responsiveness of V5 to these inputs. The insert shows all the voxels in V5 that evidenced a modulatory effect ($p < 0.05$ uncorrected). These voxels were identified by thresholding a SPM (Friston *et al* 1995b) of the F statistic testing for the contribution of second order kernels involving V2 and PPC (treating all other terms as nuisance variables). The data were obtained with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes).

Figure 9

a) Top: These renderings illustrate the extent of cerebral infarcts in four patients, as identified by voxel-based morphometry. Regions of reduced grey matter (relative to neurologically normal controls) are shown in white on the left hemisphere. The SPMs were thresholded at $P<0.001$ uncorrected. All patients had damage to Broca's area. The first (upper left) patient's left middle cerebral artery infarct was most extensive encompassing temporal and

parietal regions as well as frontal and motor cortex. b) Bottom: SPMs illustrating the functional imaging results with regions of significant activation shown in black on the left hemisphere. Results are shown for: (i) Normal subjects reading words (left). (ii) Activations common to normal subjects and patients reading words using a conjunction analysis (middle-top). (iii) Areas where normal subjects activate significantly more than patients reading words, using the group times condition interaction (Middle lower). (iv) The first patient activating normally for a semantic task. Context-sensitive failures to activate are implied by the abnormal activations in the first patient, for the implicit reading task, despite a normal activation during a semantic task.

Table 1

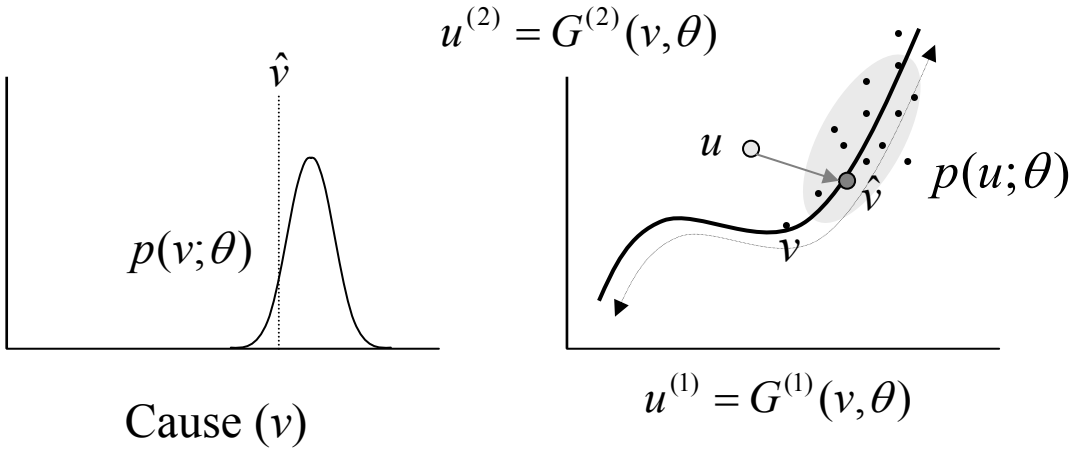
Some key characteristics of extrinsic cortico-cortical connections in the brain

Hierarchical organisation

- The organisation of the visual cortices can be considered as a hierarchy (Felleman and Van Essen 1991).
- The notion of a hierarchy depends upon a distinction between forward and backward extrinsic connections.
- This distinction rests upon different laminar specificity (Rockland and Pandya 1979, Salin and Bullier 1995).
- Backward connections are more numerous and transcend more levels
- Backward connections are more divergent than forward connections (Zeki and Shipp 1988).

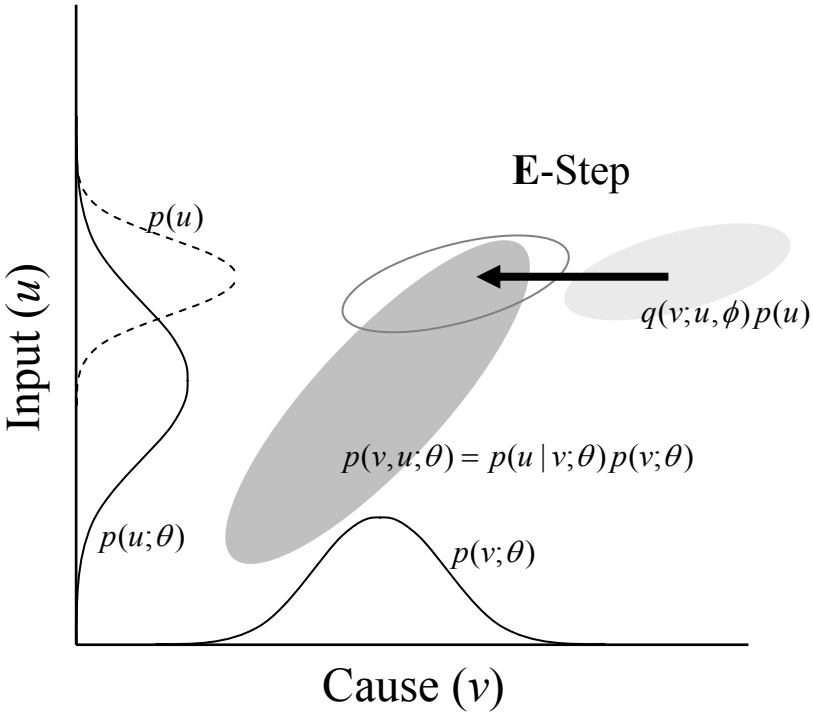
Forwards connections	**Backwards connections**
Sparse axonal bifurcations	Abundant axonal bifurcation
Topographically organised	Diffuse topography
Originate in supragranular layers	Originate in bilaminar/infragranular layers
Terminate largely in layer VI	Terminate predominantly in supragranular layers
Postsynaptic effects through fast AMPA (1.3-2.4ms decay) and $GABA_A$ (6ms decay) receptors.	Modulatory afferents activate slow (50ms decay) voltage-sensitive NMDA receptors

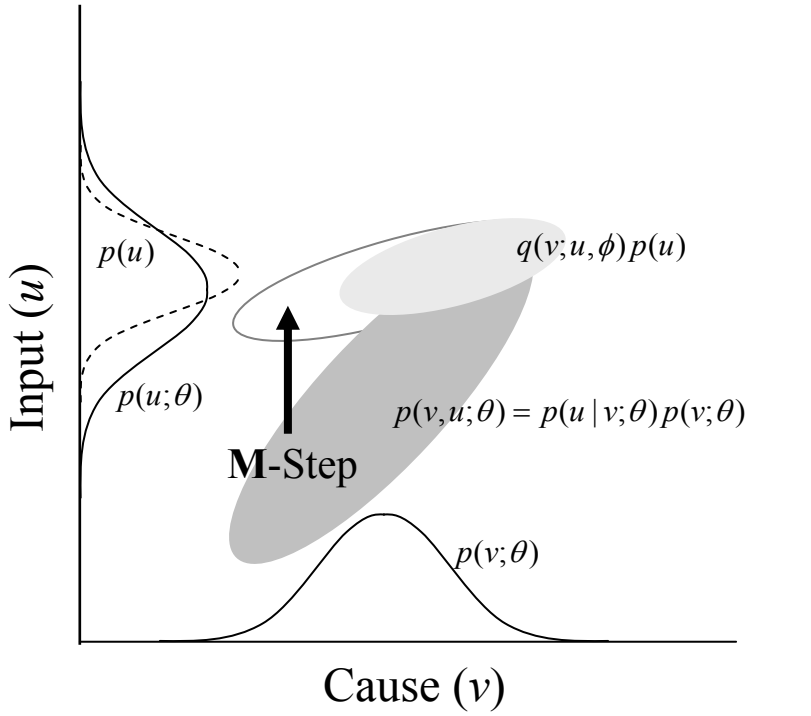
Inference and learning



Cause (v)

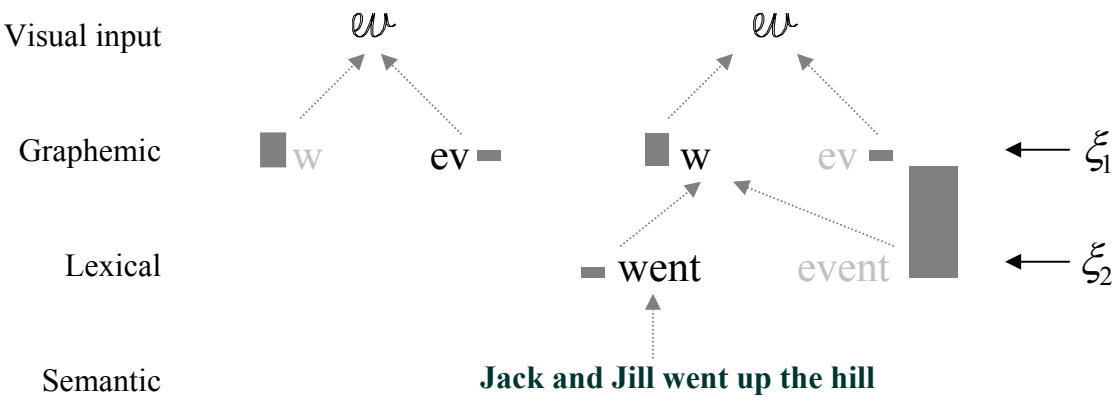
Expectation Maximisation



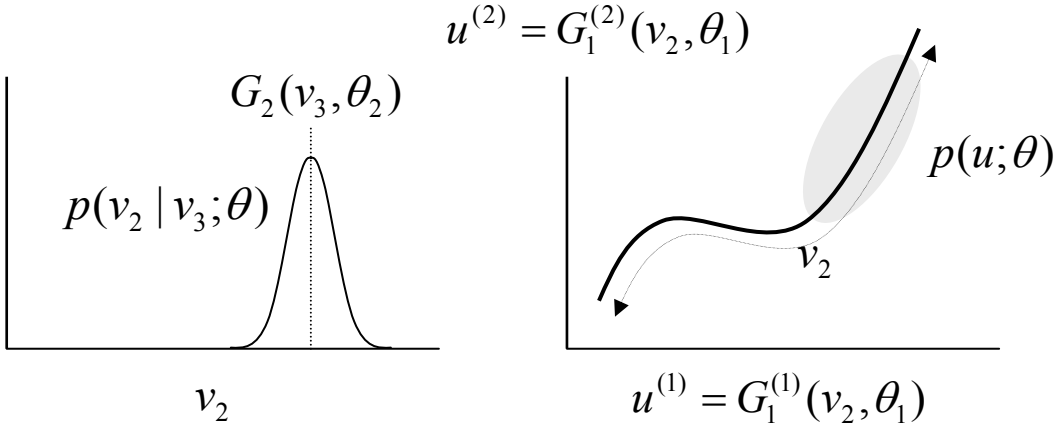


Jack and Jill event up the hill

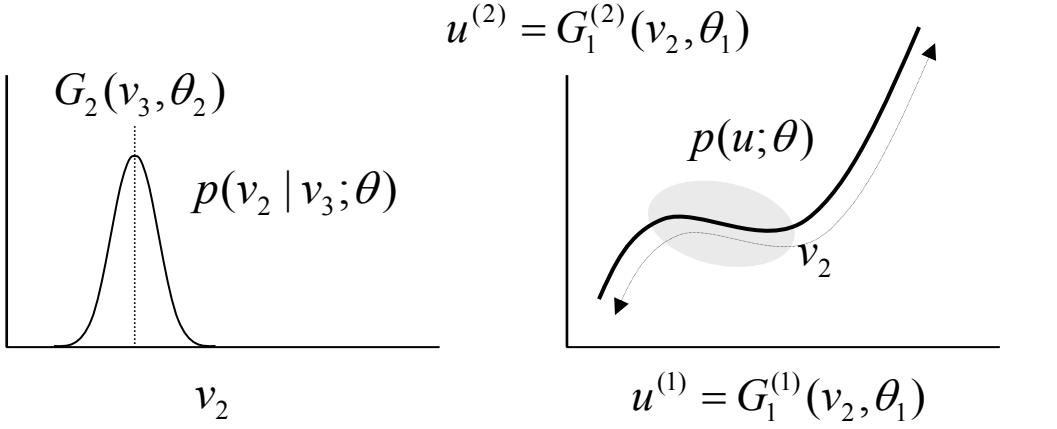
The last event was cancelled



Hierarchical models



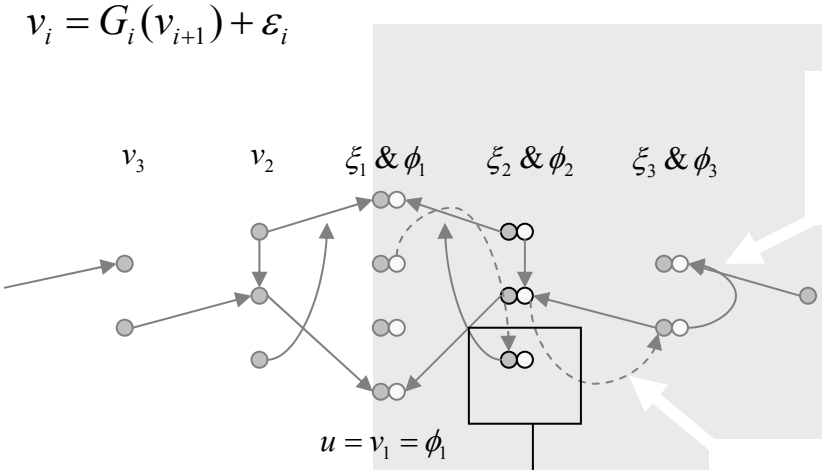
Context 1



Context 2

Hierarchical architectures for Empirical Bayes

$$v_i = G_i(v_{i+1}) + \varepsilon_i$$

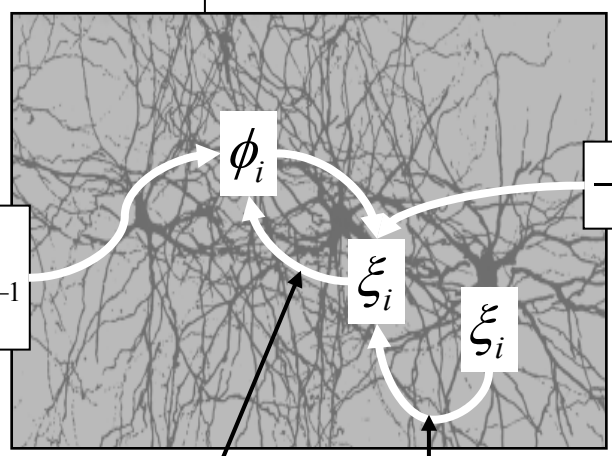


Backward & lateral

$$\xi_i = \phi_i - G_i(\phi_{i+1}, \theta_i) - \lambda_i \xi_i$$

Forward & lateral

$$\dot{\phi}_{i+1} = -\frac{\partial \xi_i^T}{\partial \phi_{i+1}} \xi_i - \frac{\partial \xi_{i+1}^T}{\partial \phi_{i+1}} \xi_{i+1}$$



Forward recognition effects

$$-\frac{\partial \xi_{i-1}^T}{\partial \phi_i} \xi_{i-1}$$

$$-G_i(\phi_{i+1}, \theta_i)$$

Backward generation effects

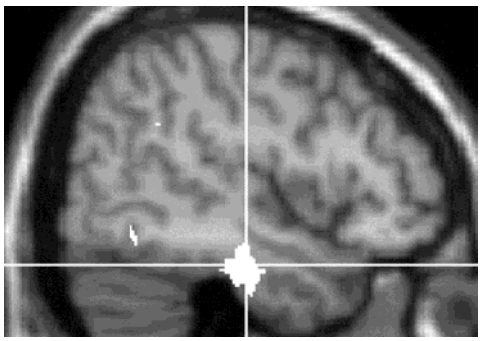
lateral interactions mediating priors

$$-\frac{\partial \xi_i^T}{\partial \phi_i} \xi_i$$

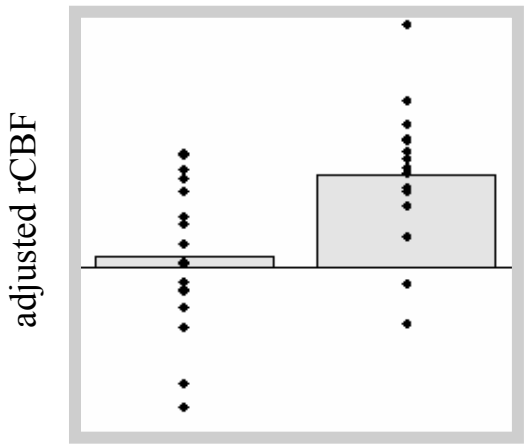
$$-\lambda_i \xi_i$$

De-correlating lateral interactions

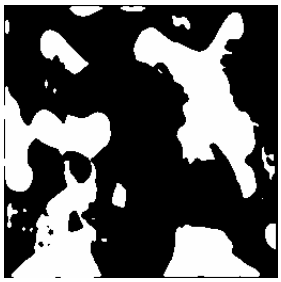
Regionally-specific interactions

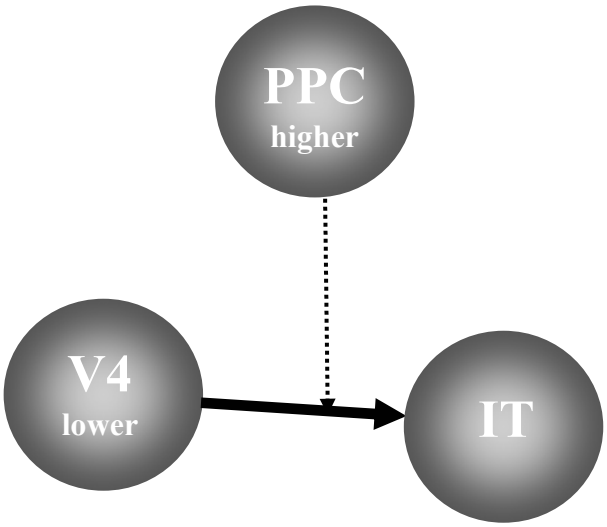


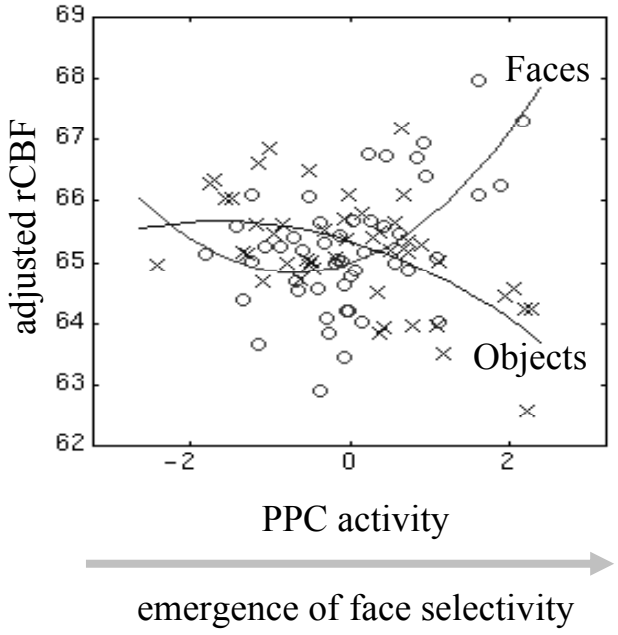
Object-specific activations

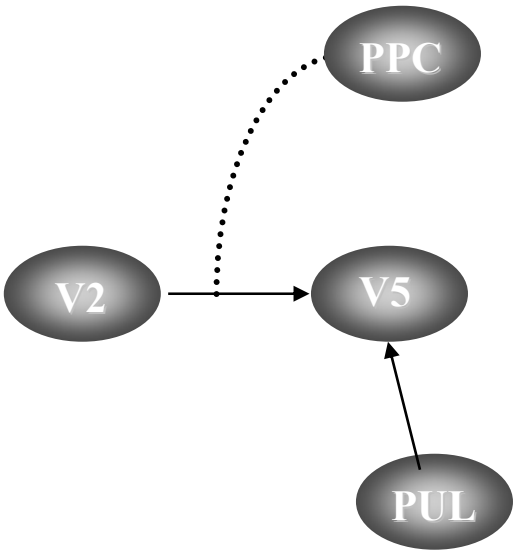


Modulation of face-selectivity by PPC

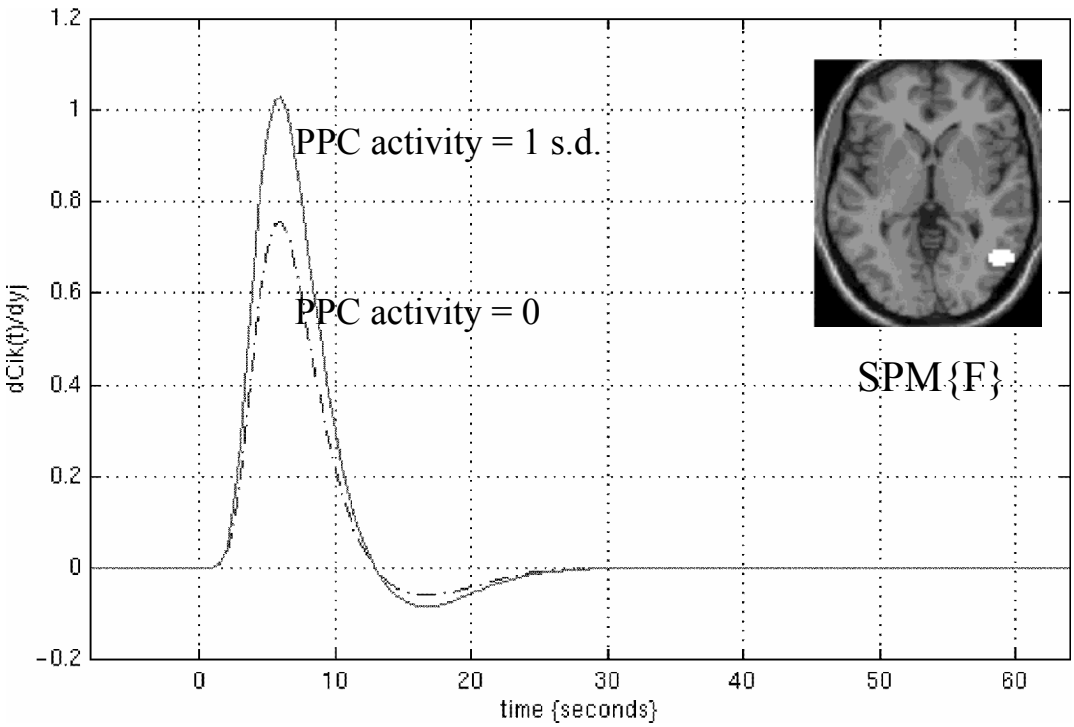






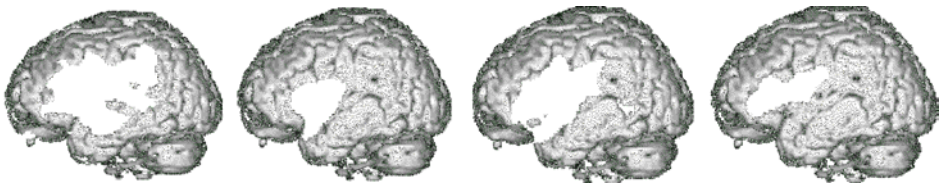


Changes in V5 responses to inputs from V2 with PPC activity



Dynamic diaschisis

a) Lesion sites in four patients



b) Patterns of activation