CHAPTER

36

Functional integration

K. Friston

INTRODUCTION

The next chapters are about functional integration in the brain. This chapter reviews the neurobiology of functional integration, in terms of neuronal information processing and frames the sorts of question that can be addressed with analyses of functional and effective connectivity. In fact, we use empirical Bayes (see Part 4) as the basis for understanding integration among the levels of hierarchically organized cortical areas. The next two chapters (Chapters 37 and 38) deal with functional and effective connectivity. Chapters 39 and 40 deal with complementary models of functional integration, namely the Volterra or generalized convolution formulation and state-space representations. The final chapters in this section cover dynamic causal modelling. For a more mathematical treatment of these models and their interrelationships, see Appendix 2.

In this chapter, we will review the empirical evidence for functional specialization and integration, with a special focus on extrinsic connections among cortical areas and how they define cortical hierarchies. We will then look at these hierarchical architectures, using ideas from theoretical neurobiology, which clarify the potential role of forward and backward connections. Finally, we see how neuroimaging can be used to test hypotheses that arise from this theoretical treatment. Specifically, we show that functional neuroimaging can be used to test for interactions between bottom-up and top-down influences on an area and to make quantitative inferences about changes in connectivity. This chapter is quite theoretical and is used to introduce constraints from neurobiology and computational neuroscience that provide a context for the models of functional integration presented in the subsequent chapters.

Interactions and context-sensitivity

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 shift is apparent in imaging. Namely, the appreciation that functional specialization can exhibit similar extra-classical phenomena, in which a cortical area may be specialized for one thing in one context but something else in another (see McIntosh, 2000). These extra-classical phenomena have implications for theoretical ideas about how the brain might work. This chapter uses theoretical models of representational or perceptual inference as a vehicle to illustrate how imaging can be used to address important questions about functional brain architectures.

Many models of perceptual learning and inference require prior assumptions about the distribution of sensory causes. However, as seen in the previous chapters, empirical Bayes suggests 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 data are caused, can construct empirical priors and are essential for perception. Moreover, non-linear generative models require these connections to be modulatory so that 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.

Overview

We start by reviewing the two fundamental principles of brain organization, namely functional specialization

Statistical Parametric Mapping, by Karl Friston et al.
ISBN–13: 978-0-12-372560-8 ISBN–10: 0-12-372560-7

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 theory. Generative models based on predictive coding rest on hierarchies of backward and lateral projections and, critically, confer a necessary role on backward connections. The theme of context-sensitive responses is used in the subsequent section to preview different ways of measuring connectivity with functional neuroimaging. The focus of this section is evidence for the interaction of bottom-up and top-down influences and establishing the presence of backward connections. The final section reviews some of the implications for lesion studies and neuropsychology. *Dynamic diaschisis* is introduced, in which aberrant neuronal responses can be observed following damage to distal brain areas that provide enabling or modulatory afferents.

FUNCTIONAL SPECIALIZATION AND INTEGRATION

The brain appears to adhere to two fundamental principles of functional organization, functional integration and functional specialization, where the integration within and among specialized areas is mediated by effective connectivity. The distinction relates to that between 'localizationism' and '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 localization *per se* was not easy to demonstrate: for example, a meeting that took place on 4 August 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 'Localization 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 localization because localizationism 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 localizationism as a complete or sufficient explanation of cortical organization. Functional localization implies that a function can be localized in a cortical area, whereas specialization suggests that a cortical area is specialized for some aspects of perceptual or motor processing, where this *specialization* can be anatomically *segregated* within the cortex. The cortical infrastructure supporting a single function may then involve many specialized areas whose union is mediated by the functional integration among them. Functional specialization and integration are not exclusive, they are complementary. Functional specialization is only meaningful in the context of functional integration and vice versa.

Functional specialization 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 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 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 a clear relationship to functional segregation. For example, the secondary visual area, V2, has a cytochrome oxidase architecture, consisting of thick, thin and interstripes. 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 specialized for visual motion. Evidence of this nature supports the notion that patchy connectivity is the anatomical infrastructure that underpins functional segregation and specialization. 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.

The anatomy and physiology of cortico-cortical connections

If specialization rests upon connectivity then important organizational 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 organization 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 summarized in Table 36-1. The list is not exhaustive but serves to introduce some important principles that have emerged from empirical studies of visual cortex:

- *Hierarchical organization* – the organization 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 (see Figure 36.1).
- *Reciprocal connections* – although reciprocal, forward and backward connections show both a microstructural and functional asymmetry. The terminations of both show laminar specificity. Forward connections (from a low to a high level) have sparse axonal bifurcations and are topographically organized, originating in supra-granular layers and terminating largely in layer IV. 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 supra-granular 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 5–8 mm 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 the 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 than forward connections and transcend more levels. For example, the ratio of forward efferent connections to backward afferents in the lateral geniculate is about one to ten. 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, 1989) and neuroimaging (e.g. Büchel and Friston, 1997) studies suggest that forward connections are driving and always elicit a response, whereas backward connections can be modulatory. In this context, modulatory means backward connections modulate responsiveness to other inputs. At the single cell level, 'inputs from drivers can be differentiated from those of modulators. The driver can be identified as the transmitter of receptive field properties; the modulator can be identified as altering the probability of certain aspects of that transmission' (Sherman and Guillery, 1998).

The notion that forward connections are concerned with the segregation of sensory information is consistent

TABLE 36-1 Some key characteristics of extrinsic cortico-cortical connections

- The organization 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 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)

Forward connections	Backward connections
Sparse axonal bifurcations	Abundant axonal bifurcation
Topographically organized	Diffuse topography
Originate in supra-granular layers	Originate in bi-laminar/infra-granular layers
Terminate largely in layer IV	Terminate predominantly in supra-granular layers
Postsynaptic effects through fast AMPA (1.3–2.4 ms decay) and $GABA_A$ (6 ms decay) receptors	Modulatory afferents activate slow (50 ms decay) voltage-sensitive NMDA receptors

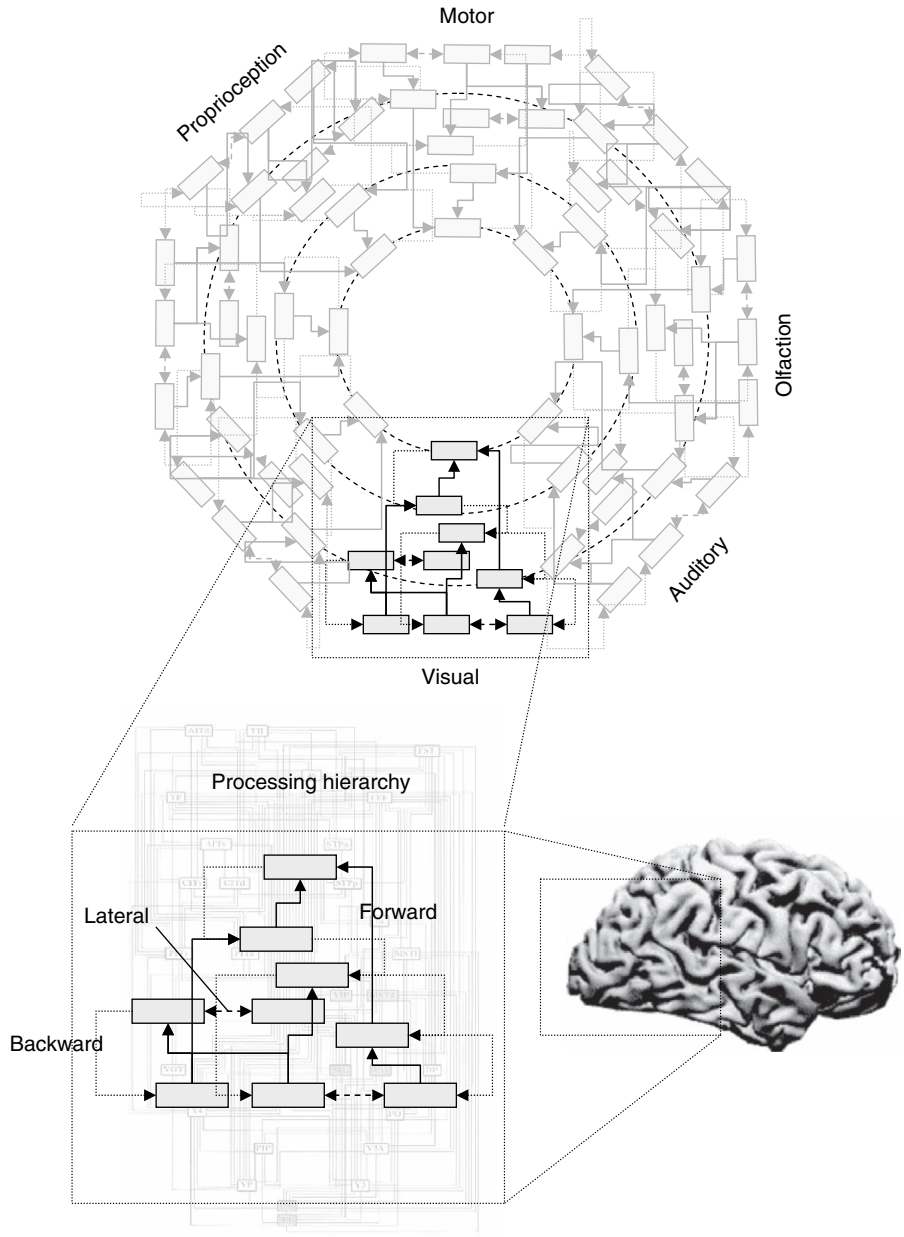

FIGURE 36.1 Schematic illustrating hierarchical structures in the brain and the distinction between forward, backward and lateral connections. This schematic is inspired by Mesulam's (1998) notion of sensory-fugal processing over 'a core synaptic hierarchy, which includes the primary sensory, upstream unimodal, downstream unimodal, hetero-modal, paralimbic and limbic zones of the cerebral cortex' (see Mesulam, 1998 for more details).

with their sparse axonal bifurcation, patchy axonal terminations and topographic projections. In contradistinction, backward connections are considered to have a role in mediating contextual effects and in the coordination of processing channels. This is consistent with their frequent bifurcation, diffuse axonal terminations and more divergent topography (Salin and Bullier, 1995; Crick and Koch, 1998). Forward connections mediate their postsynaptic effects through fast AMPA and GABA receptors. Modulatory effects can be mediated by NMDA receptors. NMDA receptors are voltage-sensitive, showing non-linear and slow dynamics ($\sim$50 ms decay). They are found predominantly in supra-granular 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. The clearest evidence for the modulatory role of backward connections (mediated by slow glutamate receptors) comes from cortico-geniculate connections; in the cat lateral geniculate nucleus, cortical

feedback is partly mediated by type 1 metabotropic glutamate receptors, which are located exclusively on distal segments of the relay-cell dendrites. Rivadulla *et al.* (2002) have shown that these backward afferents enhance the excitatory centre of the thalamic receptive field. 'Therefore, cortex, by closing this cortico-fugal loop, is able to increase the gain of its thalamic input within a focal spatial window, selecting key features of the incoming signal' (Rivadulla *et al.*, 2002) (see also Murphy and Sillito, 1987).

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. *In 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. Following birth, connections are progressively refined and remodelled with a greater emphasis on activity- and use-dependent plasticity. These changes endure into adulthood with ongoing reorganization and experience-dependent plasticity that furnish behavioural adaptation and learning throughout life. In brief, there are two basic determinants of connectivity: *cellular plasticity*, reflecting cell-migration and neurogenesis in the developing brain; and s*ynaptic 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 underlie 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 an associative aspect on synaptic changes. This is because their voltage-sensitivity only allows calcium ions to enter the cell when there is conjoint presynaptic release of glutamate and sufficient postsynaptic depolarization (i.e. the temporal association of pre- and postsynaptic events). Calcium entry renders the postsynaptic specialization eligible for future potentiation by promoting the formation of synaptic 'tags' (e.g. Frey and Morris 1997) and other calcium-dependent intracellular mechanisms.

In summary, the anatomy and physiology of cortico-cortical connections suggest that forward connections are driving and commit cells to a prespecified 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 36-1). For example, in the visual cortex Angelucci *et al.* (2002a; b) used a combination of anatomical and physiological recording methods to determine the spatial scale and retinotopic logic of intra-area V1 horizontal connections and inter-area 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.'

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 and are in a position to exert powerful effects on evoked responses, in lower levels, that define the specialization of any area or neuronal population. Modulatory effects imply the postsynaptic response evoked by presynaptic input is modulated by, or interacts with, another input. By definition this interaction must depend on non-linear synaptic or dendritic mechanisms.

Functional integration and effective connectivity

Electrophysiology and imaging neuroscience have firmly established functional specialization as a principle of brain organization in man. The functional integration of specialized areas has proven more difficult to assess. Functional integration refers to the interactions among specialized 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]

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

However, correlations can arise in a variety of ways. For example, in multiunit 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 Preil, 1991). This means effective connectivity is dynamic, i.e. activity-dependent, and depends upon a model of the interactions. Recent models of effective connectivity accommodate the modulatory or non-linear effects mentioned above. A more detailed discussion of these models is provided in subsequent chapters. In this chapter, the terms modulatory and non-linear are used almost synonymously. Modulatory effects imply the postsynaptic response evoked by one input is modulated, or interacts with, another. By definition this interaction must depend on non-linear synaptic mechanisms.

In summary, the brain can be considered as an ensemble of functionally specialized areas that are coupled in a non-linear fashion by effective connections. Empirically, it appears that connections from lower to higher areas are predominantly driving, whereas backward 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 that highlights the functional importance of backward connections and non-linear interactions.

LEARNING AND INFERENCE IN THE BRAIN

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 (2001: 359–397) to which the reader is referred for background reading. A more detailed discussion of these issues can be found in Friston (2005).

First, we will reprise empirical Bayes in the context of brain function *per se*. Having established the requisite architecture for learning and inference, neuronal implementation is considered in sufficient depth to make predictions about the structural 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.

Causes, perception and sensation

Causes are simply the states of processes generating sensory data. It is not easy to ascribe meaning to these states without appealing to the way that we categorize things. Causes may be categorical in nature, such as the identity of a face or the semantic category of an object. Others may be parametric, such as the position of an object. Even though causes may be difficult to describe, they are easy to define operationally. Causes are quantities 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 non-linear function:

$$u = g(v, \theta) \qquad \textbf{36.1}$$

where v is a vector of causes in the environment (e.g. the velocity of a particular object, direction of radiant light etc.), and u represents sensory input. $g(v, \theta)$ is a function that generates data from the causes. θ are the parameters of the generative model. Unlike the causes, they are fixed quantities that have to be learned. We shall see later that the parameters correspond to connection strengths in the brain's model of how data are caused. Non-linearities in Eqn. **36.1** represent interactions among the causes. These can often be viewed as contextual effects, where the expression of a particular cause depends on the context established by another. A ubiquitous example, from early visual processing, is the occlusion of one object by another; in a linear world the visual sensation, caused by two objects, would be a transparent overlay or superposition. Occlusion is a non-linear phenomenon because the sensory input from one object (occluded) interacts, or depends on, the other (occluder). This interaction is an example of non-linear mixing of causes to produce sensory data. At a higher level, the cause associated with the word 'hammer' depends on the semantic context (that determines whether the word is a verb or a noun).

The problem the brain has to contend with is to find a function of the data that *recognizes* the underlying causes. To do this, the brain must undo the interactions to disclose contextually invariant causes. In other words, the brain must perform a non-linear un-mixing of causes and context. The key point here is that the non-linear mixing may not be invertible and recognition may be a fundamentally ill-posed problem. For example, no amount of un-mixing can recover the parts of an object that are

occluded by another. 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). In what follows, we consider the implications of this problem. Put simply, recognition of causes from sensory data is the inverse of generating data from causes. If the generative model is not invertible, then recognition can only proceed if there is an explicit generative model in the brain.

The specific model considered here rests on empirical Bayes. This model can be regarded as a mathematical formulation of the long-standing notion (Locke, 1690) that: 'our minds should often change the idea of its sensation into that of its judgement, and make one serve only to excite the other'. In a similar vein, Helmholtz (1860) distinguishes between perception and sensation: 'It may often be rather hard to say how much from perceptions as derived from the sense of sight is due directly to sensation, and how much of them, on the other hand, is due to experience and training' (see Pollen, 1999). In short, there is a distinction between a percept, which is the product of recognizing the causes of sensory input, and sensation *per se*. Recognition, i.e. inferring causes from sensation, is the inverse of generating sensory data from their causes. It follows that recognition rests on models, learned through experience, of how sensations are caused.

Conceptually, empirical Bayes and generative models are related to 'analysis-by–synthesis' (Neisser, 1967). This approach to perception, from cognitive psychology, involves adapting an internal model of the world to match sensory input and was suggested by Mumford (1992) as a way of understanding hierarchical neuronal processing. The idea is reminiscent of Mackay's epistemological automata (MacKay, 1956) which perceive by comparing expected and actual sensory input, (Rao, 1999). These models emphasize the role of backward connections in mediating predictions of lower level input, based on the activity of higher cortical levels.

Generative models and perception

This section introduces a basic framework for understanding learning and inference. This framework rests upon generative and recognition models, which are functions that map causes to sensory input and vice versa. Generative models afford a generic formulation of representational learning and inference 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) to 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). The distinction between reconstructing data and learning efficient representations relates directly to the distinction between inference and learning.

Inference vs. learning

Generative models relate unknown causes v and unknown parameters θ to observed sensory data u. The objective is to make *inferences* about the causes and *learn* the parameters. Inference may be simply estimating the most likely cause and is based on estimates of the parameters from learning. A generative model is specified in terms of a *prior* distribution over the causes $p(v;\theta)$ and the *generative* distribution or likelihood of the data given the causes $p(u|v;\theta)$. Together, these define the marginal distribution of data implied by a generative model:

$$p(u;\theta) = \int p(u|v;\theta)p(v;\theta)dv \qquad \textbf{36.2}$$

The conditional density of the causes, given the data, are given by the recognition model, which is defined in terms of the *recognition* or conditional distribution:

$$p(v|u;\theta) = \frac{p(u|v;\theta)p(v;\theta)}{p(u;\theta)} \qquad \textbf{36.3}$$

However, as considered above, the generative model may not be inverted easily and it may not be possible to parameterize this recognition distribution. This is crucial because the endpoint of learning is the acquisition of a useful recognition model that can be applied to sensory data. One solution is to posit an approximate recognition or conditional density $q(v)$ that is consistent with the generative model and that can be parameterized. Estimating the moments (e.g. expectation) of this density corresponds to *inference*. Estimating the parameters of the underlying generative model corresponds to *learning*. This distinction maps directly onto the two steps of expectation–maximization (EM) (Dempster *et al.*, 1977).

Expectation maximization

To keep things simple, assume that we are only interested in the first moment or expectation of $q(v)$ which we will denote by μ. This is the conditional mean or expected cause. EM is a coordinate ascent scheme that comprises an **E**-step and an **M**-step. In the present context, the **E**-step finds the conditional expectation of the causes (i.e. inference), while the **M**-step identifies the maximum likelihood value of the parameters (i.e. learning). Critically, both adjust the conditional causes and parameters to maximize the same thing.

The free energy formulation

EM provides a useful procedure for density estimation that has direct connections with statistical mechanics. Both steps of the EM algorithm involve maximizing a function of the densities above. This function corresponds to the negative free energy in physics (see Chapter 24 and the Appendices for more details):

$$F = \ln p(u;\theta) - KL\{q(v), p(v|u;\theta)\} \qquad \textbf{36.4}$$

This objective function has two terms. The first is the marginal likelihood of the data under the generative model. The second term is the Kullback-Leibler divergence[2] between the approximate and true recognition densities. Critically, the second term is always positive, rendering F a lower-bound on the expected log-likelihood of the data. This means maximizing the objective function (i.e. minimizing the free energy) is simply minimizing our surprise about the data. The **E**-step increases F with respect to the expected cause, ensuring a good approximation to the recognition distribution implied by the parameters θ. This is inference. The **M**-step changes θ, enabling the generative model to match the likelihood of the data and corresponds to learning:

$$\begin{aligned} &\text{Inference: } \mathbf{E} \qquad \mu = \max_{\mu} F \\ &\text{Learning: } \mathbf{M} \qquad \theta = \max_{\theta} F \end{aligned} \qquad \textbf{36.5}$$

The remarkable thing is that both inference and learning are driven in exactly the same way, namely to minimize the free energy. This is effectively the same as minimizing surprise about sensory data encountered. The implication, as we will see below, is that the same principle can explain phenomena as wide ranging as the mis-match negativity (MMN) in evoked electrical brain responses to Hebbian plasticity during perceptual learning.

Predictive coding

We have now established an objective function that is maximized to enable inference and learning in **E**- and **M**-steps respectively. Here, we consider how that maximization might be implemented. In particular, we will look at predictive coding, which is based on minimizing prediction error (Rao and Ballard, 1998). Prediction error is the difference between the data observed and that predicted by the inferred causes. We will see that minimizing the free energy is equivalent to minimizing prediction error. Consider any static non-linear generative model under Gaussian assumptions:

$$\begin{aligned} u &= g(v,\theta) + \varepsilon^{(1)} \\ v &= \eta + \varepsilon^{(2)} \end{aligned} \qquad \textbf{36.6}$$

where $Cov(\varepsilon^{(1)}) = \Sigma^{(1)}$ is the covariance of random fluctuations in the sensory data. Priors on the causes are specified in terms of their expectation η and covariance $Cov(\varepsilon^{(2)}) = \Sigma^{(2)}$. This form will be useful in the next section when we generalize to hierarchical models. For simplicity, we will approximate the recognition density with a point mass. From Eqn. **36.4**:

$$\begin{aligned} F &= -\frac{1}{2}\xi^{(1)T}\xi^{(1)} - \frac{1}{2}\xi^{(2)T}\xi^{(2)} - \frac{1}{2}\ln|\Sigma^{(1)}| - \frac{1}{2}\ln|\Sigma^{(2)}| \\ \xi^{(1)} &= \Sigma^{(1)^{-1/2}}(u - g(\mu,\theta)) \\ \xi^{(2)} &= \Sigma^{(2)^{-1/2}}(\mu - \eta) \end{aligned} \qquad \textbf{36.7}$$

The first term in Eqn. **37.7** is the prediction error that is minimized in predictive coding. The second corresponds to a prior term that constrains or regularizes conditional estimates of the causes. The need for this term stems from the ill-posed nature of recognition discussed above and is a ubiquitous component of inverse solutions.

Predictive coding schemes can be seen in the context of forward and inverse models adopted in machine vision (Ballard *et al.*, 1983; Kawato *et al.*, 1993). Forward models generate data from causes (cf. generative models), whereas inverse models approximate the reverse transformation of data to causes (cf. recognition models). This distinction embraces the ill-posed nature of inverse problems. As with all underdetermined inverse problems, the role of constraints is central. In the inverse literature, *a priori* constraints usually enter in terms of regularized solutions. For example: 'Descriptions of physical properties of visible surfaces, such as their distance and the presence of edges, must be recovered from the primary image inputs. Computational vision aims to understand how such descriptions can be obtained from inherently ambiguous and noisy inputs. A recent development in this field sees early vision as a set of ill-posed problems, which can be solved by the use of regularization methods' (Poggio *et al.*, 1985). The architectures that emerge from these schemes suggest that: 'Feedforward connections from the lower visual cortical area to the higher visual cortical area provide an approximated inverse model of the imaging process (optics)'. Conversely: 'while the back-projection connection from the higher area to the lower area provides a forward model of the optics' (Kawato *et al.*, 1993). This perspective highlights the importance of backward connections and the role of priors in enabling predictive coding schemes.

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

Predictive coding and Bayes

Predictive coding is a strategy that has some compelling [Bayesian] underpinnings. To finesse the inverse problem posed by non-invertible generative models, constraints or priors are required. These resolve the ill-posed problems that confound recognition based on purely forward architectures. It has long been assumed that sensory units adapt to the statistical properties of the signals to which they are exposed (see Simoncelli and Olshausen, 2001 for review). In fact, the Bayesian framework for perceptual inference has its origins in Helmholtz's notion of perception as unconscious inference. Helmholtz realized that retinal images are ambiguous and that prior knowledge was required to account for perception (Kersten *et al.*, 2004). Kersten *et al.* (2004) provide an excellent review of object perception as Bayesian inference and ask a fundamental question: 'Where do the priors come from. Without direct input, how does image-independent knowledge of the world get put into the visual system?' In the next section, we answer this question and show how empirical Bayes allows priors to be learned and induced online, during inference.

Cortical hierarchies and empirical Bayes

The problem with fully Bayesian inference is that the brain cannot construct the prior expectation and variability, η and $\Sigma^{(2)}$ *de novo*. They have to be learned and, furthermore, adapted to the current experiential context. This calls for empirical Bayes, in which priors are estimated from data. Empirical Bayes harnesses the hierarchical structure of a generative model, treating the estimates at one level as priors on the subordinate level (Efron and Morris, 1973). Empirical Bayes provides a natural framework within which to treat cortical hierarchies in the brain, each level 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. Empirical priors offer contextual guidance towards the most likely cause of the data. Note that predictions at higher levels are subject to the same constraints; only the highest level, if there is one in the brain, is unconstrained.

Next, we extend the generative model to cover empirical priors. This means that constraints, required by predictive coding, are absorbed into the learning scheme. This hierarchical extension induces extra parameters that encode the variability or precision of the causes. These are referred to as hyperparameters in the classical covariance component literature. Hyperparameters are updated in the **M**-step and are treated in exactly the same way as the parameters.

Hierarchical models

Consider any level i in a hierarchy whose causes $v^{(i)}$ are elicited by causes in the level above $v^{(i+1)}$. The hierarchical form of the generative model is:

$$\begin{aligned} u &= \\ v^{(1)} &= g(v^{(2)}, \theta^{(1)}) + \varepsilon^{(1)} \\ v^{(2)} &= g(v^{(3)}, \theta^{(2)}) + \varepsilon^{(2)} \\ v^{(3)} &= \ldots \end{aligned} \qquad \textbf{36.8}$$

Technically, these models fall into the class of conditionally independent hierarchical models, when the stochastic terms are independent (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. We will assume the stochastic terms are Gaussian with covariance $\Sigma^{(i)}$. Therefore, the means and covariances determine the likelihood at each level:

$$p(v^{(i)}|v^{(i+1)}; \theta^{(i)}) = N(g^{(i)}, \Sigma^{(i)}) \qquad \textbf{36.9}$$

This likelihood *also plays the role of an empirical prior on* $v^{(i)}$ at the level below, where it is jointly maximized with the likelihood $p(v^{(i-1)}|v^{(i)}; \theta^{(i-1)})$. This is the key to understanding the utility of hierarchical models; by inferring the generative distribution of level i one is implicitly estimating the prior for level $i-1$. This enables the learning of prior densities. The hierarchical nature of these models lends an important context-sensitivity to recognition densities not found in single-level models. Because high-level causes determine the prior expectation of causes in the subordinate level, they change the distributions upon which inference is based, in a data and context-dependent way.

Implementation

The biological plausibility of empirical Bayes in the brain can be established fairly simply. The objective function is now:

$$\begin{aligned} F &= -\frac{1}{2}\xi^{(1)T}\xi^{(1)} - \frac{1}{2}\xi^{(2)T}\xi^{(2)} - \ldots \\ &\quad -\frac{1}{2}\ln|\Sigma^{(1)}| - \frac{1}{2}\ln|\Sigma^{(2)}| - \ldots \\ \xi^{(i)} &= \mu^{(i)} - g(\mu^{(i+1)}, \theta^{(i)}) - \lambda^{(i)}\xi^{(i)} \\ &= (I + \lambda^{(i)})^{-1}(\mu^{(i)} - g(\mu^{(i+1)}, \theta^{(i)})) \\ \Sigma^{(i)} &= (I + \lambda^{(i)})^2 \end{aligned} \qquad \textbf{36.10}$$

In neuronal models, the prediction error is encoded by the activities of units denoted by $\zeta^{(i)}$. These error units receive a prediction from units in the level above[3] via *backward* connections and *lateral* influences from the representational units $\mu^{(i)}$ being predicted. Horizontal interactions among the error units serve to de-correlate them (cf. Foldiak 1990), where the symmetric lateral connection strengths $\lambda^{(i)}$ hyperparameterize the covariances $\Sigma^{(i)}$, which are the prior covariances for level $i-1$.

The conditional causes and parameters perform a gradient ascent on the objective function:[4]

$$\begin{aligned}
\mathbf{E} \qquad \dot{\mu}^{(i)} &= \frac{\partial F}{\partial \mu^{(i)}} = -\frac{\partial \xi^{(i-1)T}}{\partial \mu^{(i)}}\xi^{(i-1)} - \frac{\partial \xi^{(i)T}}{\partial \mu^{(i)}}\xi^{(i)} \\
\xi^{(i)} &= \mu^{(i)} - g(\mu^{(i+1)}, \theta^{(i)}) - \lambda^{(i)}\xi^{(i)} \\
\mathbf{M} \qquad \dot{\theta}^{(i)} &= \frac{\partial F}{\partial \theta^{(i)}} = -\left\langle \frac{\partial \xi^{(i)T}}{\partial \theta^{(i)}}\xi^{(i)} \right\rangle_u \\
\dot{\lambda}^{(i)} &= \frac{\partial F}{\partial \lambda^{(i)}} = -\left\langle \frac{\partial \xi^{(i)T}}{\partial \lambda^{(i)}}\xi^{(i)} \right\rangle_u - (1+\lambda^{(i)})^{-1}
\end{aligned} \qquad \textbf{36.11}$$

Inferences, mediated by the **E**-step, rest on changes in units encoding expected causes $\mu^{(i)}$ that are mediated by forward connections from error units in the level below and lateral interactions with error units in the same level. Similarly, prediction error is constructed by comparing the activity of these units with the activity predicted by backward connections.

This is the simplest version of a very general learning algorithm. It is general in the sense that it does not require the parameters of either the generative or prior distributions. It can learn non-invertible, non-linear generative models and encompasses complicated hierarchical processes. Furthermore, each of the learning components has a relatively simple neuronal interpretation (see below).

IMPLICATIONS FOR CORTICAL INFRASTRUCTURE AND PLASTICITY

The scheme implied by Eqn. **36.11** has four clear implications for the functional architecture required to implement it. We review these in relation to cortical organization in the brain. A schematic summarizing these points is provided in Figure 36.2. In short, we arrive at exactly the same four points presented at the end of the first section.

- *Hierarchical organization* – hierarchical models enable empirical Bayesian estimation of prior densities and provide a plausible model for sensory data. Models that do not show conditional independence (e.g. those used by connectionist and infomax schemes) depend on prior constraints for inference and do not invoke a hierarchical cortical organization. The nice thing about the architecture in Figure 36.2 is that the responses of units at the i-th level $\mu^{(i)}$ depend only on the error at the current level and the immediately preceding level. Similarly the error units $\zeta^{(i)}$ are only connected to units in the current level and the level above. This hierarchical organization follows from conditional independence and is important because it permits a biologically plausible implementation, where the connections driving inference run only between neighbouring levels.
- *Reciprocal connections* – in the hierarchical scheme, the dynamics of units $\mu^{(i+1)}$ 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 mediated by linear connections with strengths that are exactly the same as the [negative] *reciprocal* connections from $\mu^{(i+1)}$ to $\zeta^{(i)}$ and $\zeta^{(i+1)}$. From Eqn. **36.11**:

$$\begin{aligned}
\frac{\partial \dot{\mu}^{(i+1)}}{\partial \xi^{(i)}} &= -\frac{\partial \xi^{(i)}}{\partial \mu^{(i+1)}}^T \\
\frac{\partial \dot{\mu}^{(i+1)}}{\partial \xi^{(i+1)}} &= -\frac{\partial \xi^{(i+1)}}{\partial \mu^{(i+1)}}^T
\end{aligned} \qquad \textbf{36.12}$$

  Functionally, forward and lateral connections are reciprocated, where backward connections generate predictions of lower-level responses. Forward connections allow prediction error to drive units in supraordinate levels. Lateral connections, within each level, mediate the influence of error units on the predicting units and intrinsic connections $\lambda^{(i)}$ among the error units de-correlate them, allowing competition among prior expectations with different precisions (precision is the inverse of variance). In short, lateral, forward and backward connections are all reciprocal, consistent with anatomical observations.
- *Functionally asymmetric forward and backward connections* – although the connections are reciprocal, the functional attributes of forward and backward influences are different. The top-down influences of units $\mu^{(i+1)}$ on error units in the lower level $\xi^{(i)} = \mu^{(i)} - g(\mu^{(i+1)}, \theta^{(i)}) - \lambda^{(i)}\xi^{(i)}$ instantiate the forward model. These can be non-linear, where each unit in the higher

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

[4] For simplicity, we have ignored conditional uncertainty about the causes that would otherwise induce further terms in the **M**-step.

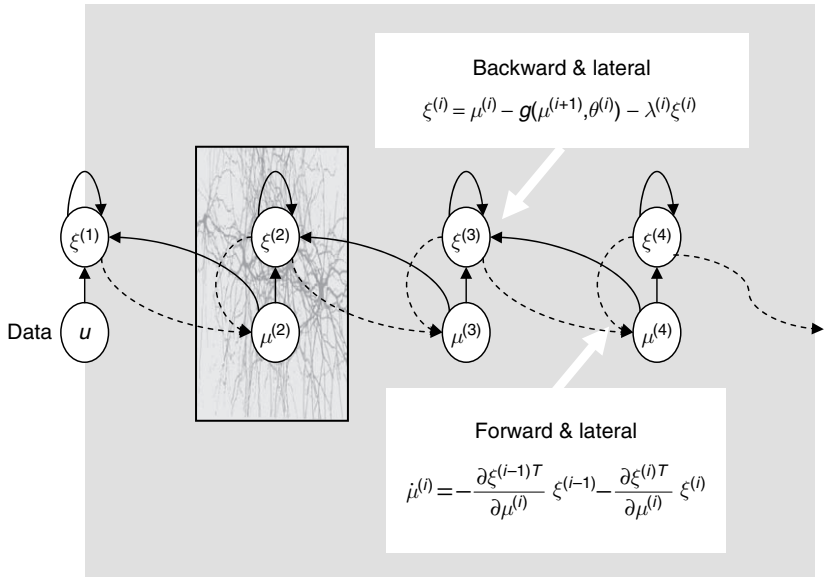

FIGURE 36.2 Schematic depicting a hierarchical predictive coding architecture. Here, hierarchical arrangements within the model serve to provide predictions or priors to representations in the level below. The upper circles represent error units and the lower circles functional subpopulations encoding the conditional expectation of causes. These expectations change to minimize both the discrepancy between their predicted value and the mismatch incurred by their prediction of the level below. These two constraints correspond to prior and likelihood terms respectively (see main text).

level *may modulate or interact with the influence of others*, according to the non-linearities in $g^{(i)}$. In contrast, the bottom-up influences of units in lower levels do not interact when producing changes at the higher level because, according to Eqn. **36.11**, their effects are linearly separable. This is a key observation because the empirical evidence, reviewed in the first section, suggests that backward connections are in a position to interact (e.g. through NMDA receptors expressed predominantly in supra-granular layers that are in receipt of backward connections). Forward connections are not. In summary, non-linearities, in the way sensory data are produced, necessitate non-linear interactions in the generative model. These are mediated by backward connections 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 parameterize the prior expectations and the lateral connections hyperparameterize the prior covariances. Together they parameterize the Gaussian densities that constitute the empirical priors. The plasticity implied can be seen more clearly with an explicit model. For example, let $g(v^{(i+1)}, \theta^{(i)}) = \theta^{(i)}v^{(i+1)}$. In this instance:

$$\begin{aligned}\dot{\theta}^{(i)} &= (1+\lambda^{(i)})^{-1}\left\langle \xi^{(i)}\mu^{(i+1)T}\right\rangle_u \\ \dot{\lambda}^{(i)} &= (1+\lambda^{(i)})^{-1}(\left\langle \xi^{(i)}\xi^{(i)T}\right\rangle_u - I)\end{aligned} \qquad \textbf{36.13}$$

This is just Hebbian or associative plasticity where the connection strengths change in proportion to the product of pre- and postsynaptic activity. An intuition about Eqn. **36.13** 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 $\left\langle \xi^{(i)}\phi^{(i+1)T}\right\rangle_u = 0$. The lateral connections stop changing when the prediction error is spherical or *IID* $\left\langle \xi^{(i)}\xi^{(i)T}\right\rangle_u = I$.

It is evident that the predictions of the theoretical analysis coincide almost exactly with the empirical aspects of functional architectures in visual cortices: hierarchical organization; reciprocity; functional asymmetry; and associative plasticity. Although somewhat contrived, it is pleasing that purely theoretical considerations and neurobiological empiricism converge so precisely.

Summary

In summary, perceptual inference and learning lends itself naturally to a hierarchical treatment, which considers the brain as an empirical Bayesian device. The dynamics of the units or populations are driven to minimize error at all levels of the cortical hierarchy and implicitly render themselves posterior modes (i.e. most likely values) of the causes given the data. In contradistinction to supervised learning, hierarchical prediction does not require any desired output. Unlike information theoretic approaches, they do not assume independent causes. In contrast to regularized inverse solutions, they do not depend on *a priori* constraints. These emerge spontaneously as empirical priors from higher levels.

The overall scheme sits comfortably with 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.

We have tried to show that this sort of hierarchical prediction can be implemented in brain-like architectures using mechanisms that are biologically plausible (Figure 36.3).

Backward or feedback connections?

There is something slightly counterintuitive about empirical Bayes in the brain. In this view, cortical hierarchies are trying to generate sensory data from high-level causes. This means the causal structure of the world is embodied in the backward connections. Forward connections simply provide feedback by conveying prediction error to higher levels. In short, forward connections are the *feedback* connections. This is why we have been careful not to ascribe a functional label like feedback to backward connections. Perceptual inference emerges from recurrent top-down and bottom-up processes that enable sensation to constrain perception. This self-organizing process is distributed throughout the hierarchy. Similar perspectives have emerged in cognitive neuroscience on the basis of psychophysical findings. For example, *Reverse Hierarchy Theory* distinguishes between early explicit perception and implicit low level vision, where: 'our initial conscious percept – vision at a glance – matches a high-level, generalized, categorical scene interpretation, identifying

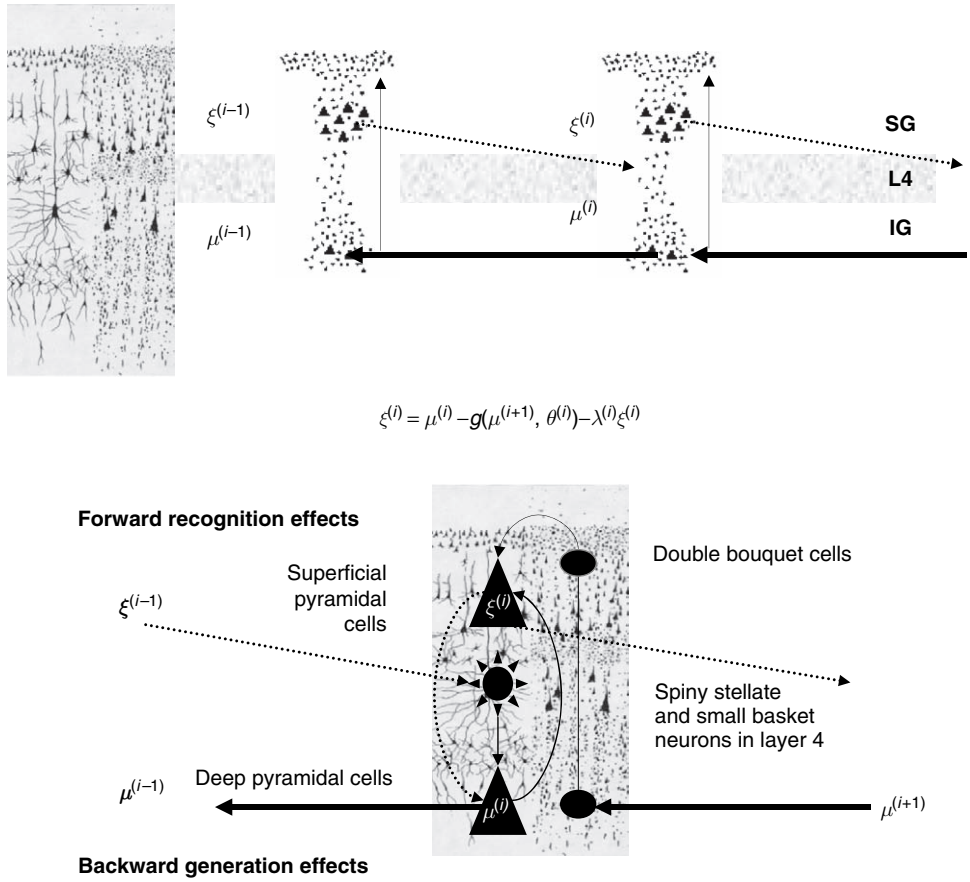

FIGURE 36.3 A more detailed picture of the influences among units. Here, we have associated various cells with different roles in the scheme described in the main text. The key constraint on this sort of association is that superficial pyramidal cells are the source of forward connections, which, according to the theory, should encode prediction error.

' "forest before trees" ' (Hochstein and Ahissar, 2002). On the other hand, backward connections are responsible for predicting; lower level responses embed the generative or *forward* model. The effect of these predictions on lower-level responses is the focus of the next section.

ASSESSING FUNCTIONAL ARCHITECTURES WITH BRAIN IMAGING

In the previous section, we have seen one approach to understanding the nature of functional integration in the brain. We will use this framework to preview the different ways in which integration can be measured empirically, with a special focus on the interaction between forward and backward connections. The examples used will appear again in later chapters that take us from simple measure of statistical correlations among different brain areas though to dynamic causal models of cortical hierarchies.

Clearly, it would be nice to demonstrate the existence of top-down 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 is not sufficient to show regional activity is partially predicted by activity in a higher level to confirm the existence of backward connections because statistical dependency does, in itself, not permit causal inference. Statistical dependencies could easily arise in a purely forward architecture because the higher-level activity is caused by activity in the lower level. Although there are causal modelling techniques (i.e. dynamic causal modelling (DCM)) that can address this problem, we will start with a simpler approach and note that interactions between bottom-up and top-down influences cannot be explained by purely feed-forward architectures. 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 influence depends upon the context established by top-down influence. This interaction is used below simply as evidence for the existence of backward influences. There are instances of predictive coding that emphasize 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).

This section focuses on the evidence for these interactions. From the point of view of functionally specialized responses, these interactions manifest as context-sensitive or contextual specialization, where modality-, category- or exemplar-specific responses, driven by bottom-up input 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, perceptual learning and, in the last section, through the study of patients with brain lesions

Context-sensitive specialization

If functional specialization 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 empirical example provides such evidence. If the contextual nature of specialization is mediated by backwards connections 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 factorial experimental designs.

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 versus non-objects) and phonological retrieval (naming versus 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. In our experiment, object-specific responses are elicited (by asking subjects to view objects relative to meaningless shapes), with and without phonological retrieval. This 'two-by-two' design allows one to look at the interaction between phonological retrieval and object recognition. This analysis identifies not regionally specific activations but regionally specific *interactions*. When we 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 36.4 and Friston *et al.*, 1996 for details). Alternatively, this region can be thought of as expressing recognition-dependent responses that are realized in, and only in, the context of having to name the object. These results can be construed as evidence of contextual specialization 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.

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 specialized 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 non-linear (second-order or modulatory effects). As such, this class of model speaks directly to functional specialization of a non-linear and contextual sort. Figure 36.5 illustrates a specific example (see Dolan *et al.*, 1997 for details). Subjects were asked to view degraded face and non-face control stimuli. 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 to un-degraded stimuli before some scans but not others. The data in the right panel of Figure 36.5 suggest 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 specialization that depends on top-down effects.

Effective connectivity

The previous examples, demonstrating contextual specialization, are consistent with functional architectures

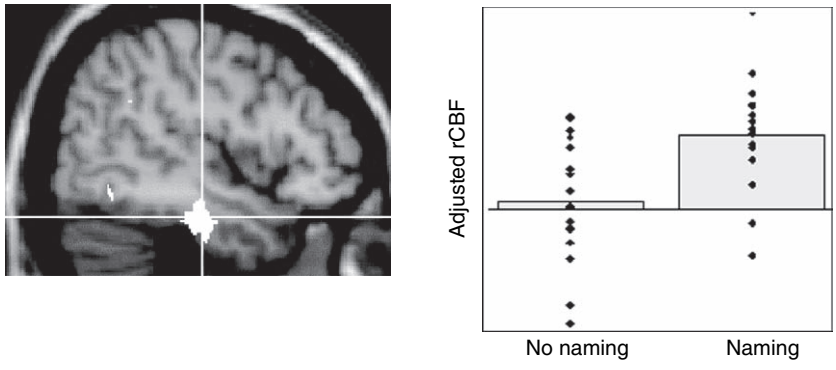

FIGURE 36.4 This example of regionally specific interactions comes from an experiment where subjects were asked to view coloured non-object shapes or coloured objects and say 'yes', or to name either the coloured object or the colour of the shape. Left: a regionally specific interaction in the left infero-temporal cortex. The SPM (statistical parametric map) 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.

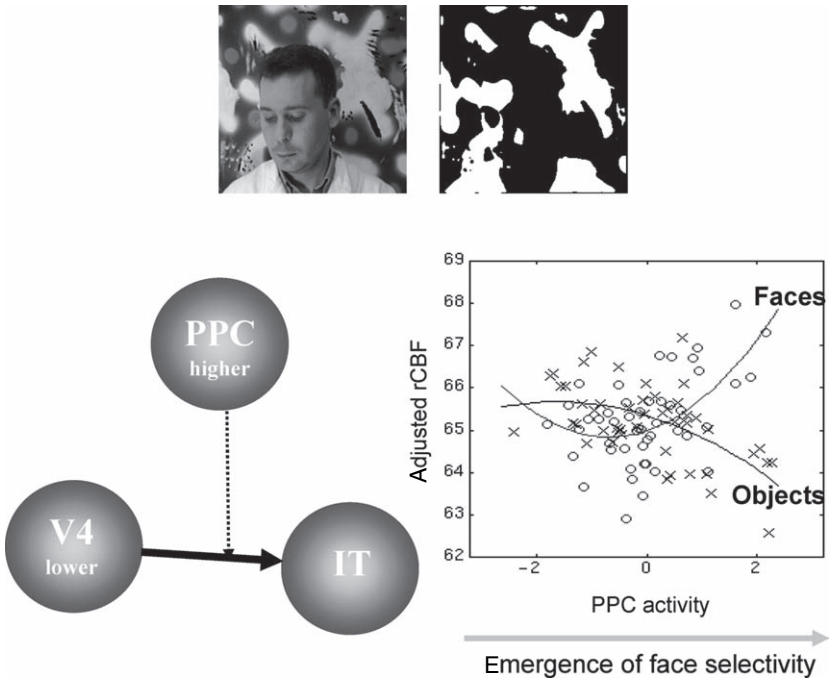

FIGURE 36.5 Top: examples of the stimuli presented to subjects. During the measurement of brain responses only degraded stimuli where shown (e.g. the right-hand picture). In half the scans, the subject was given the underlying cause of these stimuli, by presenting the original picture (e.g. left-hand picture) before scanning. This priming induced a profound difference in perceptual set for the primed, relative to non-primed, stimuli. Lower right: activity observed in a right infero-temporal region, as a function of mean-corrected posterior parietal cortex (PPC) activity. This region showed the most significant interaction between the presence of faces and activity in a reference location in the posterior medial parietal cortex. 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 while 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. Lower 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.

implied by empirical Bayes. 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 models of effective connectivity. This requires a plausible model of coupling among brain regions that can accommodate non-linear effects. We will illustrate the use of models based on the Volterra expansion and conclude with an example using DCM for event-related potentials (ERP). These examples change context using attention and perceptual learning respectively.

Non-linear 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 non-linear models like the Volterra formulation. Non-linearities 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: the direct or *driving* influence of input from the first (e.g. hierarchically lower) region, irrespective of the activities elsewhere, and a *modulatory* component that represents an interaction with input from the remaining (e.g. hierarchically higher) regions. These are mediated by first- and second-order kernels respectively. The example provided in Figure 36.6 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.

The lower panel in Figure 36.6 shows a characterization 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 line). In this study, subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) while manipulating the attentional component of the task (detection of velocity changes). The brain regions and connections comprising

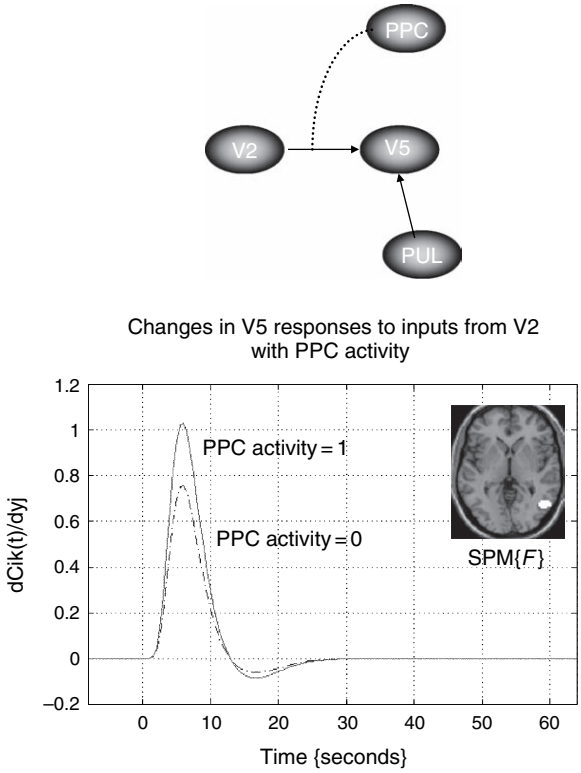

FIGURE 36.6 Upper panel: brain regions and connections comprising a model of distributed responses. Lower panel: characterization of the effects of V2 on V5 and their modulation by posterior parietal cortex (PPC). The broken line represents estimates of V5 responses when PPC activity is zero, according to a second order Volterra model of effective connectivity with input to V5 from V2, PPC and the pulvinar (PUL). The solid curve represents 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 data. The insert shows all the voxels in V5 that evidenced a modulatory effect ($p < 0.05$ uncorrected). These voxels were identified by thresholding an SPM (statistical parametric map) 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) while manipulating the attentional component of the task (detection of velocity changes).

the model are shown in the upper panel. The lower panel shows a characterization of the effects of V2 data on V5 and their modulation by posterior parietal cortex (PPC) using simulated data 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 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 are described in Friston and Büchel (2000).

These results suggest that top-down parietal influences 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. In our final example, we use perceptual learning to induce changes in connections and DCM to measure those changes. Unlike Volterra formulations of effective connectivity, dynamic causal models parameterize the coupling among brain areas explicitly. This means that one can make inference about directed influences that are causal and quantitative in nature. We will illustrate this using an example from electroencephalography.

Perceptual learning, prediction error and the MMN

The mismatch negativity (MMN) is a negative component of the ERP elicited by a change in some repetitive aspect of auditory stimulation. The MMN can be seen in the absence of attention and is generally thought to reflect pre-attentive processing in the temporal and frontal system (Näätänen, 2003). The MMN is elicited by stimulus change at about 100–200 ms after the stimulus, and is presumed to reflect an automatic comparison of stimuli to sensory memory representations encoding the repetitive aspects of auditory inputs. This prevailing theory assumes that there are distinct change-specific neurons in auditory cortex that generate the MMN. The alternative view is that preceding stimuli adapt feature-specific neurons. In this *adaptation hypothesis*, the response is delayed and suppressed on exposure to repeated stimuli, giving rise to the MMN (Jääskeläinen *et al.*, 2004).

The empirical Bayes scheme would suggest that a component of the event-related potential (ERP) corresponding to prediction error, is suppressed more efficiently after learning-related plasticity in backward and lateral connections (and implicitly forward connections by Eqn. **36.12**). This suppression would be specific for the repeated aspects of the stimuli and would be a selective suppression of prediction error. Recall that error suppression (i.e. minimization of free energy) is the motivation for plasticity in the **M**-step. The ensuing repetition suppression hypothesis suggests the MMN is simply the attenuation of evoked prediction error. As noted above, prediction error may be encoded by superficial pyramidal cells (see Figure 36.3), which are a major contributor to the ERP.

In summary, both the **E**-step and **M**-step try to minimize free energy; the **E**-step does this during perceptual

inference, on a time-scale of milliseconds, and the **M**-step, during perceptual learning, over seconds or longer. If the ERP is an index of prediction error (i.e. free energy), the ERP evoked by the first, in a train of repeated stimuli, will decrease with each subsequent presentation. This decrease discloses the MMN evoked by a new (oddball) stimulus. In this view, the MMN is subtended by a *positivity* that increases with the number of standards.

DCM and perceptual learning

We elicited event-related potentials that exhibited a strong modulation of late components, on comparing responses to frequent and rare stimuli, using an auditory oddball paradigm. Auditory stimuli, 1000 or 2000 Hz tones with 5 ms rise and fall times and 80 ms duration, were presented binaurally. The tones were presented for 15 minutes, every 2 s in a pseudo-random sequence with 2000 Hz tones on 20 per cent of occasions and 1000 Hz tones for 80 per cent of the time (standards). The subject was instructed to keep a mental record of the number of 2000 Hz tones (oddballs). Data were acquired using 128 electrodes with 1000 Hz sample-frequency. Before averaging, data were referenced to mean earlobe activity and bandpass filtered between 1 and 30 Hz. Trials showing ocular artefacts and bad channels were removed from further analysis.

Six sources were identified using conventional procedures and used to construct four dynamic causal models (see Figure 36.7 and Chapter 42). To establish evidence for

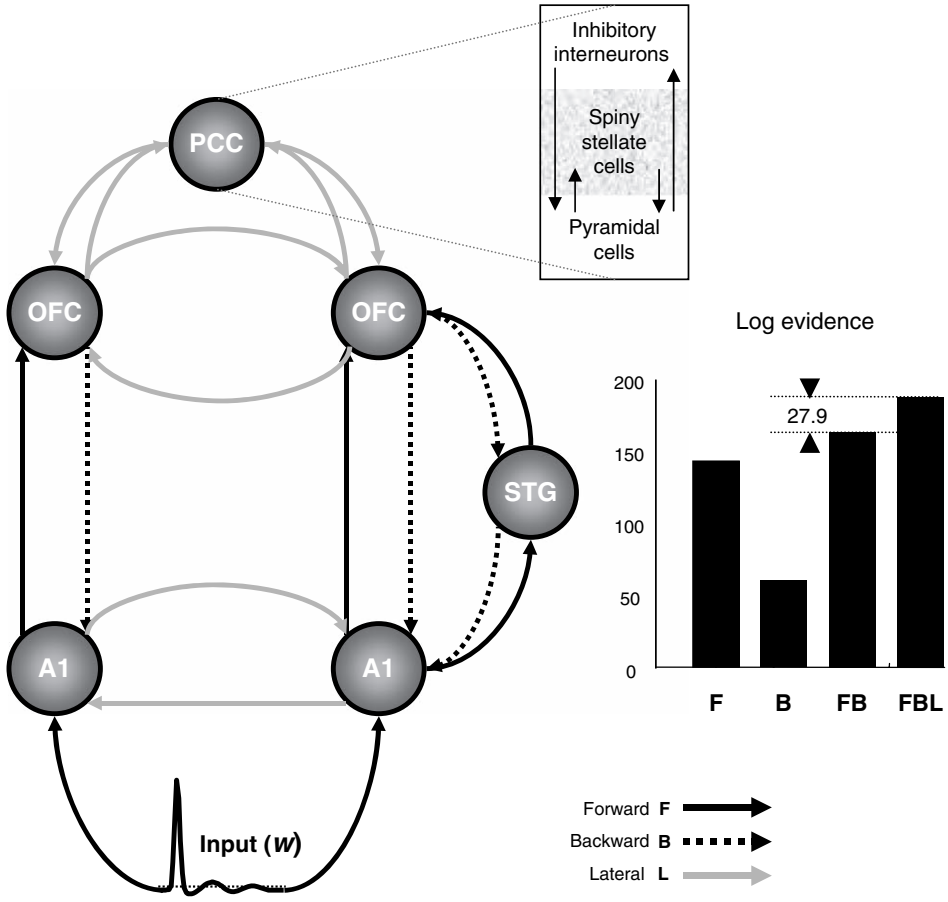

FIGURE 36.7 Left: schematic showing the extrinsic connections of a DCM (dynamic causal model) used to explain ERP data; a bilateral extrinsic input acts on primary auditory cortices which project reciprocally to orbitofrontal regions. In the right hemisphere, an indirect pathway was specified via a relay in the superior temporal gyrus. At the highest level, orbitofrontal and left posterior cingulate cortices were assumed to be laterally and reciprocally connected. Sources were coupled with extrinsic cortico-cortical connections following the rules of Felleman and van Essen (1991) – upper insert. A1: primary auditory cortex, OFC: orbitofrontal cortex, PCC: posterior cingulate cortex, STG: superior temporal gyrus (right is on the right and left on the left). The free parameters of this model included extrinsic connection strengths that were adjusted to explain best the observed ERPs. Critically, these parameters allowed for differences in connections between standard and oddball trials. Right: the results of a Bayesian model selection are shown in terms of the log-evidence for models allowing changes in forward (**F**), backward (**B**), forward and backward (**FB**) and forward, backward and lateral (**FBL**) connections. In this example, there is strong evidence that forward, backward and lateral connections change with perceptual learning.

changes in backward and lateral connections, above and beyond changes in forward connections, we employed Bayesian model selection (see Chapter 35). This entailed specifying four models that allowed for changes in forward, backward, forward and backward and in all connections. These, and only these, changes in extrinsic connectivity could explain the differences in the ERP, elicited by standard relative to oddball stimuli. The models were compared using the negative free energy as an approximation to the log-evidence for each model: if, after inversion, we assume the approximating conditional density is the true conditional density, the free energy reduces to the log-evidence (see Eqn. **36.4**). In Bayesian model selection, a difference in log-evidence of three or more can be considered as strong evidence for the model with the greater evidence, relative to the one with less. The log evidences for the four models are shown in Figure 36.7. The model with the highest evidence (by a margin of 27.9) is the DCM that allows for learning-related changes in forward, backward and lateral connections. These results provide clear evidence that changes in backward and lateral connections are needed to explain the observed differences in cortical responses.

In the final section, the implications of hierarchically organized connections are considered from the point of view of the lesion-deficit model and neuropsychology.

FUNCTIONAL INTEGRATION AND NEUROPSYCHOLOGY

If functional specialization 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 input 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.*, 2001).

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 following cerebral lesions involving the motor cortex. Dynamic diaschisis describes the task-specific effects that a lesion can have on the *evoked responses* of a distant cortical region. The basic idea 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 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.

An example from neuropsychology

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 (Figure 36.8; 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 key-press 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 36.4). 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,

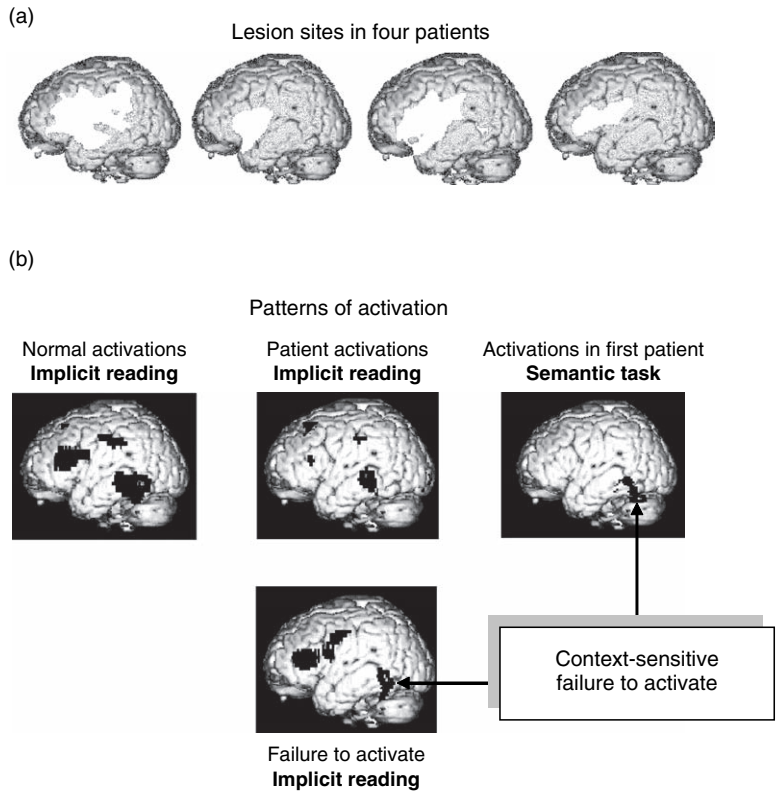

FIGURE 36.8 (a) Top: these renderings illustrate the extent of cerebral infarcts in four patients, as identified by voxel-based morphometry. Regions of reduced grey matter (relative to neurologically normal controls) are shown in white on the left hemisphere. The SPMs (statistical parametric maps) were 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 normal subjects reading words (left); activations common to normal subjects; and patients reading words using a conjunction analysis (middle-top); areas where normal subjects activate significantly more than patients reading words, using the group times condition interaction (middle lower); and 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.

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 36.8(b)). Critically, this area is thought to mediate 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 afferents 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 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 function of any neuron, neuronal population or cortical area is context-sensitive. Functional integration, or interactions among brain systems, that employ forward (bottom-up) and backward (top-down)

connections, mediate this adaptive specialization. A critical consequence is that hierarchically organized neuronal responses, in any given cortical area, can represent different things at different times. Although most models of perceptual learning and inference require priors on the causes of sensation, 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. The arguments in this chapter were developed under empirical or hierarchal Bayes models of brain function, where higher levels provide a prediction of the inputs to lower levels. 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 explained away. From this perspective, the specialization of any region is determined both by bottom-up inputs and by top-down predictions. Specialization 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 data are generated they are in a position to modulate the selectivity of lower areas.

The theoretical neurobiology in this chapter has been used to motivate the importance of measuring effective connectivity, especially modulatory or non-linear coupling in the brain. These non-linear aspects will be a recurrent theme in subsequent chapters that discuss functional and effective connectivity from a conceptual and operational point of view.

REFERENCES

Absher JR, Benson DF (1993) Disconnection syndromes: an overview of Geschwind's contributions. *Neurology* **43**: 862–67

Aertsen A, Preil H (1991) Dynamics of activity and connectivity in physiological neuronal networks. In *Non linear dynamics and neuronal networks*, Schuster HG (ed.). VCH Publishers Inc., New York, pp 281–302

Angelucci A, Levitt JB, Walton EJ *et al.* (2002a) Circuits for local and global signal integration in primary visual cortex. *J Neurosci* **22**: 8633–46

Angelucci A, Levitt JB, Lund JS (2002b) Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area V1. *Prog Brain Res* **136**: 373–88

Ballard DH, Hinton GE, Sejnowski TJ (1983) Parallel visual computation. *Nature* **306**: 21–26

Büchel C, Friston KJ (1997) Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebr Cortex* **7**: 768–78

Crick F, Koch C (1998) Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* **391**: 245–50

Dayan P, Abbott LF (2001) *Theoretical neuroscience. Computational and mathematical modelling of neural systems.* MIT Press, [**36.3]

Dayan P, Hinton GE, Neal RM (1995) The Helmholtz machine. *Neural Comput* **7**: 889–904

Dempster AP, Laird NM, Rubin DB (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 *et al.* (1997) How the brain learns to see objects and faces in an impoverished context. *Nature* **389**: 596–98

Efron B, Morris C (1973) Stein's estimation rule and its competitors – an empirical Bayes approach. *J Am Stat Assoc* **68**: 117–30

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

Foldiak P (1990) Forming sparse representations by local anti-Hebbian learning. *Biol Cybern.* **64**: 165–70

Frey U, Morris RGM (1997) Synaptic tagging and long-term potentiation. *Nature* **385**: 533–36

Friston KJ, Price CJ, Fletcher P *et al.* (1996) The trouble with cognitive subtraction. *NeuroImage* **4**: 97–104

Friston KJ, Büchel C, Fink *et al.* (1997) Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* **6**: 218–29

Friston KJ, Büchel C (2000) Attentional modulation of V5 in human. *Proc Natl Acad Sci USA* **97**: 7591–96

Friston KJ (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* **360**: 815–36

Gerstein GL, Perkel DH (1969) Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* **164**: 828–30

Girard P, Bullier J (1989) Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *J Neurophysiol* **62**: 1287–301

Helmholtz H (1860/1962) *Handbuch der physiologischen optik* (English translation, Southall JPC, ed.), Vol. 3. Dover, New York

Hinton GE, Dayan P, Frey BJ *et al.* (1995) The 'Wake-Sleep' algorithm for unsupervised neural networks. *Science* **268**: 1158–61

Hochstein S, Ahissar M (2002) View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36**: 791–804

Jääskeläinen IP, Ahveninen J, Bonmassar G *et al.* (2004) Human posterior auditory cortex gates novel sounds to consciousness. *Proc Natl Acad Sci* **101**: 6809–14

Kass RE, Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Stat Assoc* **407**: 717–26

Kawato M, Hayakawa H, Inui T (1993) A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network* **4**: 415–22

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. *Annu Rev Psychol* **55**: 271–304

Lenzi GL, Frackowiak RSJ, Jones T (1982) Cerebral oxygen metabolism blood flow in human cerebral ischaemic infarction. *J Cereb Blood Flow Metab* **2**: 321–35

Locke J (1690/1976) *An essay concerning human understanding*. Dent, London

MacKay DM (1956) The epistemological problem for automata. In *Automata studies*, Shannon CE, McCarthy J (eds). Princeton University Press, Princeton, pp 235–51

McIntosh AR (2000) Towards a network theory of cognition. *Neural Networks* **13**: 861–70

Mesulam MM (1998) From sensation to cognition. *Brain* **121**: 1013–52

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

Murphy PC, Sillito AM (1987) Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature* **329**: 727–29

Näätänen R (2003) Mismatch negativity: clinical research and possible applications. *Int J Psychophysiol* **48**: 179–88

Neisser U (1967) *Cognitive psychology*. Appleton-Century-Crofts, New York

Pearl J (2000) *Causality, models, reasoning and inference*. Cambridge University Press, Cambridge

Phillips CG, Zeki S, Barlow HB (1984) Localisation of function in the cerebral cortex: past present and future. *Brain* **107**: 327–61

Poggio T, Torre V, Koch C (1985) Computational vision and regularisation theory. *Nature* **317**: 314–19

Pollen DA (1999) On the neural correlates of visual perception. *Cereb Cortex* **9**: 4–19

Price CJ (1998) The functional anatomy of word comprehension and production. *Trends Cogn Sci* **2**: 281–88

Price CJ, Warburton EA, Moore CJ *et al.* (2001) Dynamic diaschisis: anatomically remote and context-sensitive human brain lesions. *J Cogn Neurosci* **13**: 419–29

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

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

Rivadulla C, Martinez LM, Varela C *et al.* (2002) Completing the corticofugal loop: a visual role for the corticogeniculate type 1 metabotropic glutamate receptor. *J Neurosci* **22**: 2956–62

Rockland KS, Pandya DN (1979) Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res* **179**: 3–20

Salin P-A, Bullier J (1995) Corticocortical connections in the visual system: structure and function. *Psychol Bull* **75**: 107–54

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

Sherman SM, Guillery RW (1998) On the actions that one nerve cell can have on another: distinguishing 'drivers' from 'modulators'. *Proc Natl Acad Sci USA* **95**: 7121–26

Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annu Rev Neurosci* **24**: 1193–216

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

Zeki S (1990) The motion pathways of the visual cortex. In *Vision: coding and efficiency*, Blakemore C (ed.). Cambridge University Press, Cambridge, pp 321–45

Zeki S, Shipp S (1988) The functional logic of cortical connections. *Nature* **335**: 311–17