PERGAMON

Progress in Neurobiology 590 (2002) 1–31

progress in Neurobiology

www.elsevier.com/locate/pneurobio

Review

Functional integration and inference in the brain

Karl Friston*

The Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK

Received 17 October 2001; accepted 16 August 2002

Abstract

Self-supervised models of how the brain represents and categorises the causes of its sensory input can be divided into two classes: those that minimise the mutual information (i.e. redundancy) among evoked responses and those that minimise the prediction error. Although these models have similar goals, the way they are attained, and the functional architectures employed, can be fundamentally different. This review describes the two classes of models and their implications for the functional anatomy of sensory cortical hierarchies in the brain. We then consider how empirical evidence can be used to disambiguate between architectures that are sufficient for perceptual learning and synthesis.

Most models of representational learning require prior assumptions about the distribution of sensory causes. Using the notion of empirical Bayes, we show that these assumptions are not necessary and that priors can be learned in a hierarchical context. Furthermore, we try to show that learning can be implemented in a biologically plausible way. The main point made in this review is that backward connections, mediating internal or generative models of how sensory inputs are caused, are essential if the process generating inputs cannot be inverted. Because these processes are dynamical in nature, sensory inputs correspond to a non-invertible nonlinear convolution of causes. This enforces an explicit parameterisation of generative models (i.e. backward connections) to enable approximate recognition and suggests that feedforward architectures, on their own, are not sufficient. Moreover, nonlinearities in generative models, that induce a dependence on backward connections, require these connections to be modulatory; so that estimated causes in higher cortical levels can interact to predict responses in lower levels. This is important in relation to functional asymmetries in forward and backward connections that have been demonstrated empirically.

To ascertain whether backward influences are expressed functionally requires measurements of functional integration among brain systems. This review summarises approaches to integration in terms of effective connectivity and proceeds to address the question posed by the theoretical considerations above. In short, it will be shown that functional neuroimaging can be used to test for interactions between bottom–up and top–down inputs to an area. The conclusion of these studies points toward the prevalence of top–down influences and the plausibility of generative models of sensory brain function.

Contents

* Tel.: +44-207-833-7454; fax: +44-207-813-1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).

0301-0082/02/$ – see front matter © 2002 Published by Elsevier Science Ltd.
PII: S0301-0082(02)00076-X

1. Introduction

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

We start by reviewing two fundamental principles of brain organisation, namely *functional specialisation* and *functional integration* and how they rest upon the anatomy and physiology of cortico-cortical connections in the brain. Section 2 deals with the nature and learning of representations from a theoretical or computational perspective. This section reviews *supervised* (e.g. connectionist) approaches, *information theoretic* approaches and those predicated on *predictive coding* and reprises their heuristics and motivation using the framework of *generative models*. The key focus of this section is on the functional architectures implied by each model of representational learning. Information theory can, in principle, proceed using only forward connections. However, it turns out that this is only possible when processes generating sensory inputs are invertible and independent. Invertibility is precluded when the cause of a percept and the context in which it is engendered interact. These interactions create a problem of contextual invariance that can only be solved using internal or generative models. Contextual invariance is necessary for categorisation of sensory input (e.g. category-specific responses) and represents a fundamental problem in perceptual synthesis. Generative models based on predictive coding solve this problem with hierarchies of backward and lateral projections that prevail in the real brain. In short, generative models of representational learning are a natural choice for understanding real functional architectures and, critically, confer a necessary role on backward connections.

Empirical evidence, from electrophysiological studies of animals and functional neuroimaging studies of human subjects, is presented in Sections 3 and 4 to illustrate the context-sensitive nature of functional specialisation and how its expression depends upon integration among remote cortical areas. Section 3 looks at extra-classical effects in electrophysiology, in terms of the predictions afforded by generative models of brain function. The theme of context-sensitive evoked responses is generalised to a cortical level and human functional neuroimaging studies in the subsequent section. The critical focus of this section is evidence for the interaction of bottom–up and top–down influences in determining regional brain responses. These interactions can be considered signatures of backward connections. The final section reviews some of the implications of the forging sections for lesion studies and neuropsychology. ‘Dynamic diaschisis’, is described, in which aberrant neuronal responses can be observed as a consequence of damage to distal brain areas providing enabling or modulatory afferents. This section uses neuroimaging in neuropsychological patients and discusses the implications for constructs based on the lesion-deficit model.

2. Functional specialisation and integration

2.1. Background

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

2.2. Functional specialisation and segregation

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

2.3. The anatomy and physiology of cortico-cortical connections

If specialisation rests upon connectivity then important organisational principles should be embodied in the neuroanatomy and physiology of extrinsic connections. Extrinsic connections couple different cortical areas whereas intrinsic connections are confined to the cortical sheet. There are certain features of cortico-cortical connections that provide strong clues about their functional role. In brief, there appears to be a hierarchical organisation that rests upon the distinction between *forward* and *backward* connections. The designation of a connection as forward or backward depends primarily on its cortical layers of origin and termination.

Table 1
Some key characteristics of extrinsic cortico-cortical connections in the brain

Hierarchical organisation	
The organisation of the visual cortices can be considered as a hierarchy (Felleman and Van Essen, 1991)	
The notion of a hierarchy depends upon a distinction between forward and backward extrinsic connections	
This distinction rests upon different laminar specificity (Rockland and Pandya, 1979; Salin and Bullier, 1995)	
Backward connections are more numerous and transcend more levels	
Backward connections are more divergent than forward connections (Zeki and Shipp, 1988)	
Forwards connections	Backwards connections
Sparse axonal bifurcations	Abundant axonal bifurcation
Topographically organised	Diffuse topography
Originate in supragranular layers	Originate in bilaminar/infragranular layers
Terminate largely in layer VI	Terminate predominantly in supragranular layers
Postsynaptic effects through fast AMPA (1.3–2.4 ms decay) and GABA$_A$ (6 ms decay) receptors	Modulatory afferents activate slow (50 ms decay) voltage-sensitive NMDA receptors

Some characteristics of cortico-cortical connections are presented below and are summarised in Table 1. The list is not exhaustive, nor properly qualified, but serves to introduce some important principles that have emerged from empirical studies of visual cortex.

- *Hierarchical organisation*

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

  Forwards connections (from a low to a high level) have sparse axonal bifurcations and are topographically organised; originating in supragranular layers and terminating largely in layer VI. Backward connections, on the other hand, show abundant axonal bifurcation and a diffuse topography. Their origins are bilaminar/infragranular and they terminate predominantly in supragranular layers (Rockland and Pandya, 1979; Salin and Bullier, 1995).
- *Forward connections are driving and backward connections are modulatory*

  Reversible inactivation (e.g. Sandell and Schiller, 1982; Girard and Bullier, 1989) and functional neuroimaging (e.g. Büchel and Friston, 1997) studies suggest that forward connections are driving, whereas backward connections can be modulatory. The notion that forward connections are concerned with the promulgation and segregation of sensory information is consistent with: (i) their sparse axonal bifurcation; (ii) patchy axonal terminations; and (iii) topographic projections. In contradistinction, backward connections are generally considered to have a role in mediating contextual effects and in the co-ordination of processing channels. This is consistent with: (i) their frequent bifurcation; (ii) diffuse axonal terminations; and (iii) non-topographic projections (Salin and Bullier, 1995; Crick and Koch, 1998).
- *Modulatory connections have slow time constants*

  Forward connections meditate their post-synaptic effects through fast AMPA (1.3–2.4 ms decay) and GABA$_A$ (6 ms decay) receptors. Modulatory afferents activate NMDA receptors. NMDA receptors are voltage-sensitive, showing nonlinear and slow dynamics (50 ms decay). They are found predominantly in supragranular layers where backward connections terminate (Salin and Bullier, 1995). These slow time-constants again point to a role in mediating contextual effects that are more enduring than phasic sensory-evoked responses.
- *Backwards connections are more divergent than forward connections*

  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 its convergence region (i.e. the region providing forward afferents to V5) is limited to the thick stripes (Zeki and Shipp, 1988). Reciprocal interactions between two levels, in conjunction with the divergence of backward connections, renders any area sensitive to the vicarious influence of other regions at the same hierarchical level even in the absence of direct lateral connections.
- *Backward connections are more numerous and transcend more levels*

  Backward connections are more abundant then forward connections. For example, the ratio of forward efferent connections to backward afferents in the lateral geniculate is about 1:10/20. Another important distinction is that backward connections will traverse a number of hierarchical levels, whereas forward connections are more restricted. For example, there are backward connections from TE and TEO to V1 but no monosynaptic connections from V1 to TE or TEO (Salin and Bullier, 1995).

In summary, the anatomy and physiology of cortico-cortical connections suggest that forward connections are driving and commit cells to a pre-specified response given the appropriate pattern of inputs. Backward connections, on the other hand, are less topographic and are in a position to modulate the responses of lower areas to driving inputs from either higher or lower areas (see Table 1). Backwards connections are abundant in the brain and are in a position to exert powerful effects on evoked responses, in lower levels, that define the specialisation of any area or neuronal population. The idea pursued below is that specialisation depends upon backwards connections and, due to the greater divergence of the latter, can embody contextual effects. Appreciating this is important for understanding how functional integration can dynamically reconfigure the specialisation of brain areas that mediate perceptual synthesis.

2.4. Functional integration and effective connectivity

Electrophysiology and imaging neuroscience have firmly established functional specialisation as a principle of brain organisation in man. The functional integration of specialised areas has proven more difficult to assess. Functional integration refers to the interactions among specialised neuronal populations and how these interactions depend upon the sensorimotor or cognitive context. Functional integration is usually assessed by examining the correlations among activity in different brain areas, or trying to explain the activity in one area in relation to activities elsewhere. *Functional connectivity* is defined as correlations between remote neurophysiological events. However, correlations can arise in a variety of ways. For example, in multi-unit electrode recordings they can result from stimulus-*locked* transients evoked by a common input or reflect stimulus-*induced* oscillations mediated by synaptic connections (Gerstein and Perkel, 1969). Integration within a distributed system is usually better understood in terms of *effective connectivity*. Effective connectivity refers explicitly to the influence that one neuronal system exerts over another, either at a synaptic (i.e. synaptic efficacy) or population level. It has been proposed that “the (electrophysiological) notion of effective connectivity should be understood as the experiment- and time-dependent, simplest possible circuit diagram that would replicate the observed timing relationships between the recorded neurons” (Aertsen and Preißl, 1991). This speaks to two important points: (i) effective connectivity is dynamic, i.e. activity- and time-dependent; and (ii) it depends upon a model of the interactions. An important distinction, among models employed in functional neuroimaging, is whether these models are linear or nonlinear. Recent characterisations of effective connectivity have focussed on nonlinear models that accommodate the modulatory or nonlinear effects mentioned above. A more detailed discussion of these models is provided in Section 5.2, after the motivation for their application is established in the next section. In this review the terms modulatory and nonlinear are used almost synonymously. Modulatory effects imply the post-synaptic response evoked by one input is modulated, or interacts, with another. By definition this interaction must depend on nonlinear synaptic mechanisms.

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

3. Representational learning

This section compares and contrasts the heuristics behind three prevalent computational approaches to representational learning and perceptual synthesis, *supervised learning*, and two forms of *self-supervised learning* based on information theory and predictive coding. These approaches will then be reconciled within the framework of *generative models*. This article restricts itself to sensory processing in cortical hierarchies. This precludes a discussion of other important ideas (e.g. reinforcement learning (Sutton and Barto, 1990; Friston et al., 1994), neuronal selection (Edelman, 1993) and dynamical systems theory (Freeman and Barrie, 1994)).

The relationship between model and real neuronal architectures is central to cognitive neuroscience. We address this relationship, in terms of *representations*, starting with an overview of representations in which the distinctions among various approaches can be seen clearly. An important focus of this section is the interaction among ‘causes’ of sensory input. These interactions posit the problem of *contextual invariance*. In brief, it will be shown that the problem of contextual invariance points to the adoption of generative models where interactions among causes of a percept are modelled explicitly. Within the class of self-supervised models, we will compare classical information theoretic approaches and predictive coding. These two schemes use different heuristics which imply distinct architectures that are sufficient for their implementation. The distinction rests on whether an explicit model, of the way sensory inputs are generated, is necessary for representational learning. If this model is instantiated in backwards connections, then theoretical distinctions may shed light on the functional role of backward and lateral connections that are so prevalent in the brain.

3.1. The nature of representations

What is a representation? Here a representation is taken to be a neuronal event that represents some ‘cause’ in the sensorium. Causes are simply the states of the process generating sensory data. It is not easy to ascribe meaning to these states without appealing to the way that we categorise things, perceptually or conceptually. High-level conceptual causes may be categorical in nature, such as the identity of a face in the visual field or the semantic category a perceived object belongs to. In a hierarchical setting, high-level causes may induce priors on lower-level causes that are more parametric in nature. For example, the perceptual cause "moving quickly" may show a one-to-many relationship with over-complete representations of different velocities in V5 (MT) units. An essential aspect of causes is their relationship to each other (e.g. 'is part of') and, in particular, their hierarchical structure. This ontology is often attended by ambiguous many-to-one and one-to-many mappings (e.g. a table has legs but so do horses; a wristwatch is a watch irrespective of the orientation of its hands). This ambiguity can render the problem of inferring causes from sensory information ill-posed (as we will see further).

Even though causes may be difficult to describe, they are easy to define operationally. Causes are the variables or states that are necessary to specify the products of a process (or model of that process) generating sensory information. In very general terms, let us frame the problem of representing real world causes $s(t)$ in terms of the system of deterministic equations

$$\begin{aligned}\dot{x} &= f(x, s)\\ u &= g(x)\end{aligned} \tag{1}$$

where s is a vector of underlying causes in the environment (e.g. the velocity of a particular object, direction of radiant light, etc.) and u represents sensory inputs. $\dot{x}$ means the rate of change of x, which here denotes some unobserved states of the world that form our sensory impression of it. The functions f and g can be highly nonlinear and allow for both the current state of the world and the causes of changes in those states to interact, when evoking responses in sensory units. Sensory input can be shown to be a function of, and only of, the causes *and their recent history*.

$$\begin{aligned}u &= G(s) = \sum_{i=1}^{\infty} \int_0^t \cdots \int_0^t \frac{\partial^i u(t)}{\partial s(t-\sigma_1)\cdots\partial s(t-\sigma_i)}\\ &\quad \times s(t-\sigma_1)\cdots s(t-\sigma_i)\,\mathrm{d}\sigma_1\cdots\mathrm{d}\sigma_i\end{aligned} \tag{2}$$

$G(s)$ is a functional (function of a function) that generates inputs from the causes. Eq. (2) is simply a functional Taylor expansion covering dynamical systems of the sort implied by Eq. (1). This expansion is called a Volterra series and can be thought of as a nonlinear convolution of the causes to give the inputs (see Box 1). Convolution is like smoothing, in this instance over time. A key aspect of this expansion is that it does not refer to the many hidden states of the world, only the causes of changes in states, that we want to represent. Furthermore, Eq. (1) does not contain any noise or error. This is because Eqs. (1) and (2) describe a real world process. There is no distinction between deterministic and stochastic behaviour until that process is observed. At the point the process is modelled, this distinction is invoked through notions of deterministic or observation noise. This section deals with how the brain might construct such models.

The importance of this formulation is that it highlights: (i) the *dynamical* aspects of sensory input; and (ii) the role of *interactions* among the causes of the sensory input. Dynamic aspects imply that the current state of the world, registered through our sensory receptors, depends not only on the extant causes but also on their history. Interactions among these causes, at any time in the past, can influence what is currently sensed. The second-order terms with $i = 2$ in Eq. (2) represent pairwise interactions among the causes. These interactions are formally identical to interaction terms in conventional statistical models of observed data and can be viewed as contextual effects, where the expression of a particular cause depends on the context induced by another. For example, the extraction of motion from the visual field depends upon there being sufficient luminance or wavelength contrast to define the surface moving. Another ubiquitous example, from early visual processing, is the occlusion of one object by another. In the absence of interactions, we would see a linear superposition of both objects, but the visual input caused by the nonlinear mixing of these two causes render one occluded by the other. At a more cognitive level, the cause associated with the word 'HAMMER' will depend on the semantic context (that determines whether the word is a verb or a noun). These contextual effects are profound and must be discounted before the representations of the underlying causes can be considered veridical.

The problem the brain has to contend with is to find a function of the input $u(t)$ that recognises or represents the underlying causes. To do this, the brain must effectively undo the convolution and interactions to expose contextually invariant causes. In other words, the brain must perform some form of nonlinear unmixing of 'causes' and 'context' without knowing either. The key point here is that this nonlinear mixing may not be invertible and that the estimation of causes from input may be fundamentally ill posed. For example, no amount of unmixing can discern the parts of an object that are occluded by another. The mapping $u = s^2$ provides a trivial example of this non-invertibility. Knowing u does not uniquely determine s.

Nonlinearities are not the only source of non-invertibility. Because sensory inputs are convolutions of causes, there is a potential loss of information during the convolution or smoothing that may have been critical for a unique determination of the causes. The convolution implied by Eq. (2) means the brain has to de-convolve the inputs to obtain these causes. In estimation theory this problem is sometimes called 'blind de-convolution' because the estimation is blind to the underlying causes that are convolved to give the observed variables. To simplify the presentation of the ideas below we will assume that the vectors of causes s, and their estimates v, include a sufficient history to accommodate the dynamics implied by Eq. (1).

Box 1 Dynamical systems and Volterra kernels.

Input-state–output systems and Volterra series

Neuronal systems are inherently nonlinear and lend themselves to modelling by nonlinear dynamical systems. However, due to the complexity of biological systems it is difficult to find analytic equations that describe them adequately. Even if these equations were known the state variables are often not observable. An alternative approach to identification is to adopt a very general model (Wray and Green, 1994) and focus on the inputs and outputs. Consider the single input–single output (SISO) system

$$\begin{aligned}\dot{x}(t) &= f(x(t), u(t))\\ y(t) &= g(x(t))\end{aligned}$$

The Fliess fundamental formula (Fliess et al., 1983) describes the causal relationship between the outputs and the recent history of the inputs. This relationship can be expressed as a Volterra series, in which the output *y*(*t*) conforms to a nonlinear convolution of the inputs *u*(*t*), critically without reference to the state variables *x*(*t*). This series is simply a functional Taylor expansion of *y*(*t*).

$$\begin{aligned}y(t) &= \textstyle\sum_{i=1}^{\infty}\int_0^t \cdots \int_0^t \kappa_i(\sigma_1, \cdots \sigma_i)u(t-\sigma_1)\cdots u(t-\sigma_i)\,\mathrm{d}\sigma_1\cdots \mathrm{d}\sigma_i\\ \kappa_i(\sigma_1, \cdots, \sigma_i) &= \frac{\partial^i y(t)}{\partial u(t-\sigma_1)\cdots \partial u(t-\sigma_i)}\end{aligned}$$

where $\kappa_i(\sigma_1, \ldots, \sigma_i)$ is the *i*th-order kernel. Volterra series have been described as a ‘power series with memory’ and are generally thought of as a high-order or ‘nonlinear convolution’ of the inputs to provide an output. See Bendat (1990) for a fuller discussion. This expansion is used in a number of places in the main text. When the inputs and outputs are measured neuronal activity the Volterra kernels have a special interpretation.

Volterra kernels and effective connectivity

Volterra kernels are useful for characterising the effective connectivity or influences that one neuronal system exerts over another because they represent the causal characteristics of the system in question. Neurobiologically they have a simple and compelling interpretation—they are synonymous with effective connectivity.

$$\kappa_1(\sigma_1) = \frac{\partial y(t)}{\partial u(t-\sigma_1)}, \qquad \kappa_2(\sigma_1, \sigma_2) = \frac{\partial^2 y(t)}{\partial u(t-\sigma_1)\partial u(t-\sigma_2)}, \qquad \ldots$$

It is evident that the first-order kernel embodies the response evoked by a change in input at $t - \sigma_1$. In other words it is a time-dependant measure of *driving* efficacy. Similarly the second-order kernel reflects the *modulatory* influence of the input at $t - \sigma_1$ on the response evoked at $t - \sigma_2$. And so on for higher orders.

All the schemas considered below can be construed as trying to effect a blind de-convolution of sensory inputs to estimate the causes with a recognition function.

$$v = R(u, \phi, \theta) \tag{3}$$

Here v represents an estimate of the causes and could correspond to the activity of neuronal units (i.e. neurons or populations of neurons) in the brain. The parameters ϕ and θ determine the transformations that sensory input is subject to and can be regarded as specifying the connection strengths and architecture of a neuronal network model or effective connectivity (see Box 1). For reasons that will become apparent later, we make a distinction between parameters for forward connections ϕ and backward connections θ.

The problem of recognising causes reduces to finding the right parameters such that the activity of the representational units v have some clearly defined relationship to the causes *s*. More formally, one wants to find the parameters that maximise the mutual information or statistical dependence between the dynamics of the representations and their causes. Models of neuronal computation try to solve this problem in the hope that the ensuing parameters can be interpreted in relation to real neuronal infrastructures. The greater the biological validity of the constraints under which these solutions are obtained, the more plausible this relationship becomes. In what follows, we will consider three modelling approaches: (i) supervised models; (ii) models based on information theory; and (iii) those based on predictive coding. The focus will be on the sometimes hidden constraints imposed on the parameters and the ensuing implications for connectivity architectures and the representational properties of the units. In particular, we will ask whether backward connections, corresponding to the parameters θ, are necessary. And if so what is their role? The three approaches are reprised at the end of this section by treating them as special cases of generative models. Each subsection below provides the background and heuristics for each approach and describes its implementation using the formal-

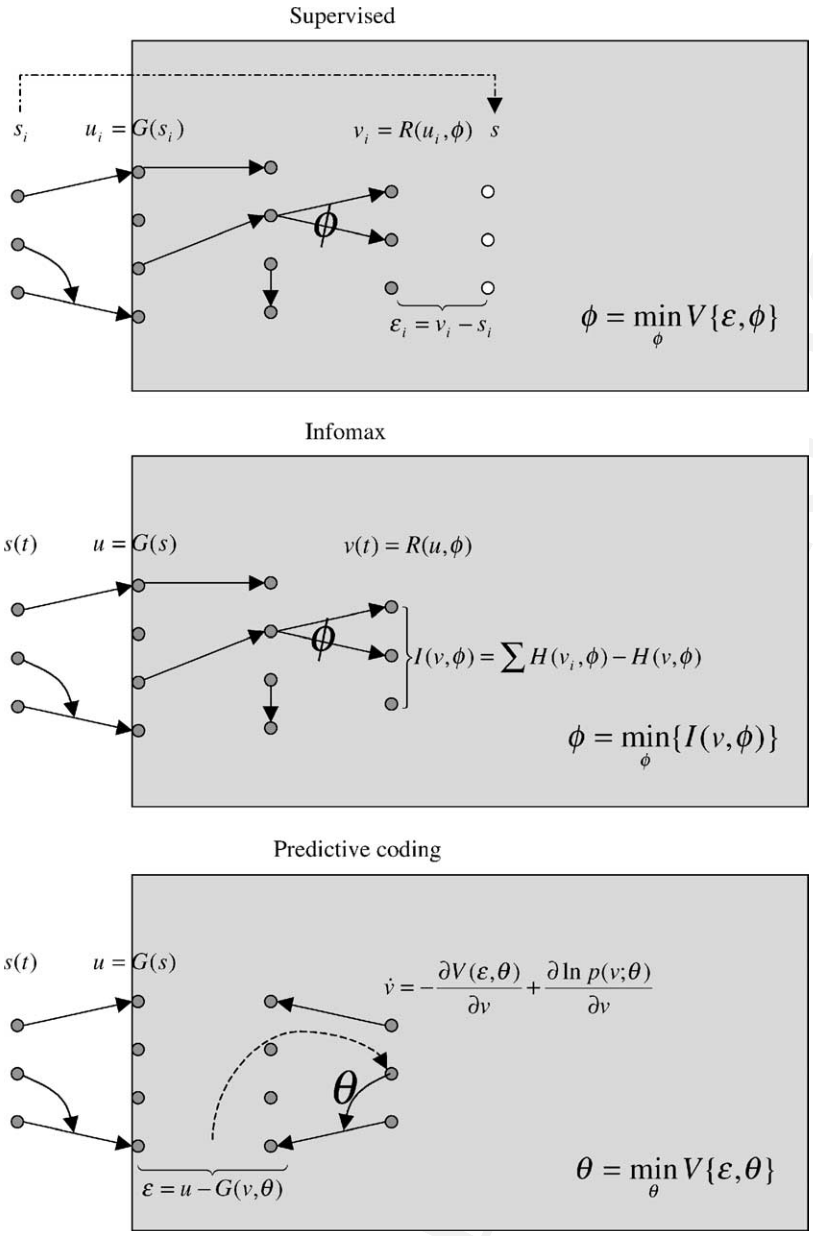
Supervised
s_i
$u_i = G(s_i)$
$v_i = R(u_i,\phi)$
s
ϕ
$\varepsilon_i = v_i - s_i$
$\phi = \min_{\phi} V\{\varepsilon,\phi\}$
Infomax
$s(t)$
$u = G(s)$
$v(t) = R(u,\phi)$
ϕ
$I(v,\phi) = \sum H(v_i,\phi) - H(v,\phi)$
$\phi = \min_{\phi}\{I(v,\phi)\}$
Predictive coding
$s(t)$
$u = G(s)$
$\dot{v} = -\frac{\partial V(\varepsilon,\theta)}{\partial v} + \frac{\partial \ln p(v;\theta)}{\partial v}$
θ
$\varepsilon = u - G(v,\theta)$
$\theta = \min_{\theta} V\{\varepsilon,\theta\}$

Fig. 1. Schematic illustrating the architectures implied by supervised, information theory-based approaches and predictive coding. The circles represent nodes in a network and the arrows represent a few of the connections. See the main text for an explanation of the equations and designation of the variables each set of nodes represents. The light grey boxes encompass connections and nodes within the model. Connection strengths are determined by the free parameters of the model ϕ (forward connections) and θ (backward connections). Nonlinear effects are implied when one arrow connects with another. Nonlinearities can be construed as the modulation of responsiveness to one input by another (see Box 1 for a more formal account). The broken arrow in the lower panel denotes connections that convey an error signal to the higher level from the input level.

ism above. Fig. 1 provides a graphical overview of the three schemes.

3.2. *Supervised models*

Connectionism is an approach that has proved very useful in relating putative cognitive architectures to neuronal ones and, in particular, modelling the impact of brain lesions on cognitive performance. Connectionism is used here as a well-known example of supervised learning in cognitive neuroscience. We start by reviewing the role played by connectionist models in the characterisation of brain systems underlying cognitive functions.

3.2.1. *Category specificity and connectionism*

Semantic memory impairments can result from a variety of pathophysiological insults, including Alzheimer's disease, encephalitis and cerebrovascular accidents (e.g. Nebes, 1989; Warrington and Shallice, 1984). The concept of category specificity stems from the work of Warrington and colleagues (Warrington and McCarthy, 1983; Warrington and Shallice, 1984) and is based on the observation that patients with focal brain lesions have difficulties in recognising or naming specific categories of objects. Patients can exhibit double dissociations in terms of their residual semantic capacity. For example, some patients can name artifacts but have difficulty with animals, whereas others can name animals with more competence than artifacts. These findings have engendered a large number of studies, all pointing to impairments in perceptual synthesis, phonological or lexico-semantic analysis that is specific for certain categories of stimuli. There are several theories that have been posited to account for category specificity. Connectionist models have been used to adjudicate among some of them.

Connectionist (e.g. parallel distributed processing or PDP) techniques use model neuronal architectures that can be lesioned to emulate neuropsychological deficits. This involves modelling semantic networks using connected units or nodes and suitable learning algorithms to determine a set of connection strengths (Rumelhart and McClelland, 1986). Semantic memory impairments are then simulated by lesioning the model to establish the nature of the interaction between neuropathology and cognitive deficit (e.g. Hinton and Shallice, 1991; Plaut and Shallice, 1993). A compelling example of this sort of approach is the connectionist model of Farah and McClelland (1991): patterns of category-specific deficits led Warrington and McCarthy (1987) to suggest that an animate/inanimate distinction could be understood in terms of a differential dependence on functional and structural (perceptual) features for recognition. For example, tools have associated motor acts whereas animals do not, or tools are easier to discriminate based upon their structural descriptions than four-legged animals. Farah and McClelland (1991) incorporated this difference in terms of the proportion of the two types of semantic featural representations encoding a particular object, with perceptual features dominating for animate objects and both represented equally for artifacts. Damage to visual features led to impairment for natural kinds and conversely damage to functional features impaired the output for artifacts. Critically the model exhibited category-specific deficits in the absence of any category-specific organisation. The implication here is that an anatomical segregation of structural and functional representations is sufficient to produce category-specific deficits following focal brain damage. This example serves to illustrate how the connectionist paradigm can be used to relate neuronal and cognitive domains. In this example, connectionist models were able to posit a plausible anatomical infrastructure wherein the specificity of deficits, induced by lesions, is mediated by differential dependence on either the functional or structural attributes of an object and not by any (less plausible) category-specific anatomical organisation per se.

3.2.2. Implementation

In connectionist models causes or ‘concepts’ like “TABLE” are induced by patterns of activation over units encoding semantic primitives (e.g. structural—“has four legs” or functional—“can put things on it”). These primitives are simple localist representations “that are assumed to be encoded by larger pools of neurons in the brain” (Devlin et al., 1998). Irrespective of their theoretical bias, connectionist models assume the existence of fixed representations (i.e. units that represent a structural, phonological or lexico-semantic primitive) that are activated by some input. These representational attributions are immutable where each unit has its ‘label’. The representation of a concept, object or ‘cause’ in the sensorium is defined in terms of which primitives are active.

Connectionist models employ some form of *supervised learning* where the model parameters (connection strengths or biases) change to minimise the difference between the observed and required output. This output is framed in terms of a distributed profile or pattern of activity over the (output) units $v = R(u, \phi)$ which arises from sensory input u corresponding to activity in (input) primitives associated with the stimulus being simulated. There are often hidden units interposed between the input and output units. The initial input (sometimes held constant or ‘clamped’ for a while) is determined by a generative function of the ith stimulus or cause $u_i = G(s_i)$. Connectionist models try to find the free parameters ϕ that minimise some function or potential V of the error or difference between the output obtained and that desired

$$\begin{aligned}\phi &= \min_{\phi} V(\varepsilon, \phi)\\ \varepsilon_i &= R(u_i, \phi) - s_i\end{aligned} \tag{4}$$

The potential is usually the (expected) sum of squared differences. Although the connectionist paradigm has been very useful in relating cognitive science and neuropsychology, it has a few limitations in the context of understanding how the brain learns to represent things:

- First, one has to know the underlying cause s_{i} and the generative function, whereas the brain does not. This is the conventional criticism of supervised algorithms as a model of neuronal computation. Neural networks, of the sort used in connectionism, are well known to be flexible nonlinear function approximators. In this sense they can be used to approximate the inverse of any generative function $u_i = G(s_i)$ to give model architectures that can lesioned. However, representational learning in the brain has to proceed without any information about the processes generating inputs and the ensuing architectures cannot be ascribed to connectionist mechanisms.
- Secondly, the generative mapping $u_i = G(s_i)$ precludes nonlinear interactions among stimuli or causes, dynamic or static. This is a fundamental issue because one of the main objectives of neuronal modelling is to see how representations emerge with the nonlinear mixing and contextual effects prevalent in real sensory input. Omitting interactions among the causes circumvents one of the most important questions that could have been asked; namely

how does the brain unmix sensory inputs to discount contextual effects and other aspects of nonlinear mixing? In short, the same inputs are activated by a given cause, irrespective of the context. This compromises the plausibility of connectionist models when addressing the emergence of representations.

In summary, connectionist models specify distributed profiles of activity over (semantic) primitives that are induced by (conceptual) causes and try to find connectivity parameters that emulate the inverse of these mappings. They have been used to understand how the performance (storage and generalisation) of a network responds to simulated damage, after learning is complete. However, connectionism has a limited role in understanding representational learning per se. In the next subsection we will look at self-supervised approaches that do not require the causes for learning.

3.3. Information theoretic approaches

There have been many compelling developments in theoretical neurobiology that have used information theory (e.g. Barlow, 1961; Optican and Richmond, 1987; Linsker, 1988; Oja, 1989; Foldiak, 1990; Tovee et al., 1993; Tononi et al., 1994). Many appeal to the principle of maximum information transfer (e.g. Linsker, 1988; Atick and Redlich, 1990; Bell and Sejnowski, 1995). This principle has proven extremely powerful in predicting some of the basic receptive field properties of cells involved in early visual processing (e.g. Atick and Redlich, 1990; Olshausen and Field, 1996). This principle represents a formal statement of the common sense notion that neuronal dynamics in sensory systems should reflect, efficiently, what is going on in the environment (Barlow, 1961). In the present context, the principle of maximum information transfer (infomax; Linsker, 1988) suggests that a model's parameters should maximise the mutual information between the sensory input u and the evoked responses or outputs $v = R(u, \phi)$. This maximisation is usually considered in the light of some sensible constraints, e.g. the presence of noise in sensory input (Atick and Redlich, 1990) or dimension reduction (Oja, 1989) given the smaller number of divergent outputs from a neuronal population than convergent inputs (Friston et al., 1992).

Intuitively, mutual information is like the covariance or correlation between two variables but extended to cover multivariate observations. It is a measure of statistical dependence. In a similar way, entropy can be regarded as the uncertainty or variability of an observation (cf. variance of a univariate observation). The mutual information between inputs and outputs under ϕ is given by

$$\begin{aligned} I(u, v; \phi) &= H(u) + H(v; \phi) - H(u, v; \phi) \\ &= H(v; \phi) - H(v|u) \end{aligned} \tag{5}$$

where $H(v|u)$ is the conditional entropy or uncertainty in the output, given the input. For a deterministic system there is no such uncertainty and this term can be discounted (see Bell and Sejnowski, 1995). More generally

$$\frac{\partial}{\partial\phi} I(u, v; \phi) = \frac{\partial}{\partial\phi} H(v; \phi) \tag{6}$$

It follows that maximising the mutual information is the same as maximising the entropy of the responses. The infomax principle (maximum information transfer) is closely related to the idea of efficient coding. Generally speaking, redundancy minimisation and efficient coding are all variations on the same theme and can be considered as the infomax principle operating under some appropriate constraints or bounds. Clearly it would be trivial to conform to the infomax principle by simply multiplying the inputs by a very large number. What we would like to do is to capture the information in the inputs using a small number of output channels operating in some bounded way. The key thing that distinguishes among the various information theoretic schemas is the nature of the constraints under which entropy is maximised. These constraints render infomax a viable approach to recovering the original causes of data, if one can enforce the outputs to conform to the same distribution of the causes (see Section 3.3.1). One useful way of looking at constraints is in terms of efficiency.

3.3.1. Efficiency, redundancy and information

The efficiency of a system can be considered as the complement of redundancy (Barlow, 1961), the less redundant, the more efficient a system will be. Redundancy is reflected in the dependencies or mutual information among the outputs. (cf. Gawne and Richmond, 1993).

$$I(v; \phi) = \sum H(v_i; \phi) - H(v; \phi) \tag{7}$$

Here $H(v_i; \phi)$ is the entropy of the ith output. Eq. (7) implies that redundancy is the difference between the joint entropy and the sum of the entropies of the individual units (component entropies). Intuitively this expression makes sense if one considers that the variability in activity of any single unit corresponds to its entropy. Therefore, an efficient neuronal system represents its inputs with the minimal excursions from baseline firing rates. Another way of thinking about Eq. (7) is to note that maximising efficiency is equivalent to minimising the mutual information among the outputs. This is the basis of approaches that seek to de-correlate or orthogonalise the outputs. To minimise redundancy one can either minimise the entropy of the output units or maximise their joint entropy, while ensuring the other is bounded in some way. Olshausen and Field (1996) present a very nice analysis based on sparse coding. Sparse coding minimises redundancy using single units with low entropy. Sparse coding implies coding by units that fire very sparsely and will, generally, not be firing. Therefore, one can be relatively certain about their (quiescent) state, conferring low entropy on them.

Approaches that seek to maximise the joint entropy of the units include principal component analysis (PCA) learning algorithms (that sample the subspace of the inputs that have the highest entropy) (e.g. Foldiak, 1990) and independent component analysis (ICA). In PCA the component entropies are bounded by scaling the connection strengths of a simple recognition model $v = R(u, \phi) = \phi u$ so that the sum of the variances of v_i is constant. ICA finds nonlinear functions of the inputs that maximise the joint entropy (Common, 1994; Bell and Sejnowski, 1995). The component entropies are constrained by the passing the outputs through a sigmoid squashing function $v = R(u, \phi) = \sigma(\phi u)$ so that the outputs lie in a bounded interval (hypercube). See Section 3.6.1 for a different perspective on ICA in which the outputs are not bounded but forced to have cumulative density functions that conform to the squashing function.

An important aspect of the infomax principle is that it goes a long way to explaining functional segregation in the cortex. One perspective on functional segregation is that each cortical area is segregating its inputs into relatively independent functional outputs. This is exactly what infomax predicts. See Friston et al. (2001 and references therein) for an example of how infomax can be used to predict the segregation of processing streams from V2 to specialised motion, colour and form areas in extrastriate cortex.

3.3.2. Implementation

In terms of the above formulation, information theoretic approaches can be construed as finding the parameters of a forward recognition function that maximise the efficiency or minimise the redundancy

$$\begin{aligned} \phi &= \min\nolimits_{\phi} I(v;\phi) \\ v &= R(u,\phi) \end{aligned} \tag{8}$$

But when are the outputs of an infomax model veridical estimates of the causes of its inputs? This is assured when: (i) the generating process is invertible; and (ii) the real world causes are independent such that $H(s) = \sum H(s_i)$. This can be seen by noting

$$\begin{aligned} I(v;\phi) &= \sum H(v_i;\phi) - H(v;\phi) \\ &= \sum H(R_i(G(s),\phi)) - \sum H(s_i) \\ &\quad - \left\langle \ln \left| \frac{\partial R(G(s),\phi)}{\partial v} \right| \right\rangle \geq 0 \end{aligned} \tag{9}$$

with equality when $v = R(u, \phi) = G^{-1}(u) = s$. Compared to the connectionist scheme this has the fundamental advantage that the algorithm is unsupervised by virtue of the fact that the causes and generating process are not needed by Eq. (8). Note that the architectures in Fig. 1, depicting connectionist and infomax schemes, are identical apart from the nodes representing desired output (unfilled circles in the upper panel). However, there are some outstanding problems:

- First, infomax recovers causes only when the generating process is invertible. However, as we have seen above the nonlinear convolution of causes generating inputs may not be invertible. This means that the recognition enacted by forward connections may not be defined in relation to the generation of inputs.
- Second, we have to assume that the causes are independent. While this may be sensible for simple systems it is certainly not appropriate for more realistic hierarchical processes that generate sensory inputs (see Section 3.5.1). This is because correlations among causes at any level are induced by, possibly independent, casual changes at supraordinate levels.

Finally, the dynamical nature of evoked neuronal transients is lost in many information theoretic formulations which treat the inputs as a stationary stochastic process, not as the products of a dynamical system. This is because the mutual information and entropy measures, that govern learning, pertain to probability distributions. These densities do not embody information about the temporal evolution of states, if they simply describe the probability the system will be found in a particular state when sampled over time. Indeed, in many instances, the connection strengths are identifiable given just the densities of the inputs, without any reference to the fact that they were generated dynamically or constituted a time-series (cf. principal component learning algorithms that need only the covariances of the inputs). Discounting dynamics is not a fundament of infomax schemas. For example, my own work using ICA referred to above (Friston et al., 2000) expanded inputs using temporal basis functions to model the functional segregation of motion, colour and form in V2. This segregation emerged as a consequence of maximising the information transfer between spatio-temporal patterns of visual inputs and V2 outputs.

In summary ICA and like-minded approaches, that try to find some deterministic function of the inputs that maximises information transfer, impose some simplistic and strong constraints on the generating process that must be met before veridical representations emerge. In the final approach, considered here, we discuss predictive coding models that do not require invertibility or independence and, consequently, suggest a more natural form for representational learning.

3.4. Predictive coding and the inverse problem

Over the past years predictive coding and generative models have supervened over other modelling approaches to brain function and represent one of the most promising avenues, offered by computational neuroscience, to understanding neuronal dynamics in relation to perceptual categorisation. In predictive coding the dynamics of units in a network are trying to predict the inputs. As with infomax schemas, the representational aspects of any unit emerge spontaneously as the capacity to predict improves with learning. There is no a priori ‘labelling’ of the units or any supervision in terms of what a correct response should be (cf. connectionist approaches). The only correct response is one in which the implicit internal model of the causes and their nonlinear mixing is sufficient to predict the input with minimal error.

Conceptually, predictive coding and generative models (see further) 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 emphasise the role of backward connections in mediating the prediction, at lower or input levels, based on the activity of units in higher levels. The connection strengths of the model are changed so as to minimise the error between the predicted and observed inputs at any level. This is in direct contrast to connectionist approaches were connection strengths change to minimise the error between the observed and *desired* output. In predictive coding there is no 'output' because the representational meaning of the units is not pre-specified but emerges during learning.

Predictive coding schemes can also be regarded as arising from the distinction between forward and inverse models adopted in machine vision (Ballard et al., 1983; Kawato et al., 1993). Forward models generate inputs from causes, whereas inverse models approximate the reverse transformation of inputs to causes. This distinction embraces the non-invertibility of generating processes and the ill-posed nature of inverse problems. As with all underdetermined inverse problems the role of constraints becomes central. In the inverse literature a priori constraints usually enter in terms of regularised 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 data. Computational vision aims to understand how such descriptions can be obtained from inherently ambiguous and noisy data. A recent development in this field sees early vision as a set of ill-posed problems, which can be solved by the use of regularisation 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 provides an approximated inverse model of the imaging process (optics), while the backprojection connection from the higher area to the lower area provides a forward model of the optics" (Kawato et al., 1993).

3.4.1. Implementation

Predictive, or more generally, generative, models turn the inverse problem on its head. Instead of trying to find functions of the inputs that predict the causes they find functions of causal estimates that predict the inputs. As in approaches based on information theory, the causes do not enter into the learning rules, which are therefore unsupervised. Furthermore, they do not require the convolution of causes, engendering the inputs, to be invertible. This is because generative or forward model is instantiated explicitly. Here the forward model is the nonlinear mixing of causes that, by definition must exist. The estimation of the causes still rests upon constraints, but these are now framed in terms of the forward model and have a much more direct relationship to casual processes in the real world. The ensuing mirror symmetry between the real generative process and its forward model is illustrated in the architecture in Fig. 1. Notice that the connections within the model are now going backwards. In the predictive coding scheme these backward connections, parameterised by θ form predictions from some estimate of the causes v to provide a prediction error. The parameters now change to minimise some function of the prediction error cf. Eq. (4).

$$\begin{aligned} \theta &= \min{}_{\theta} V(\varepsilon, \theta) \\ \varepsilon &= u - G(v, \theta) \end{aligned} \qquad (10)$$

The differences between Eqs. (10) and (4) are that the errors are at the input level, as opposed to the output level and the parameters now pertain to a forward model instantiated in backward connections. This minimisation scheme eschews the real causes *s* but where do their estimates come from? These casual estimates or representations change in the same way as the other free parameters of the model. They change to minimise prediction error subject to some a priori constraint, modelled by a regularisation term $\lambda(v, \theta)$, usually through gradient ascent.[1]

$$\dot{v} = -\frac{\partial V(\varepsilon, \theta)}{\partial v} + \frac{\partial \lambda(v, \theta)}{\partial v} \qquad (11)$$

The error is conveyed from the input layer to the output layer by forward connections that are rendered as a broken line in the lower panel of Fig. 1. This component of the predictive coding scheme has a principled (Bayesian) motivation that is described in the next subsection. For the moment, consider what would transpire after training and prediction error is largely eliminated. This implies the brain's nonlinear convolution of the estimated causes recapitulates the real convolution of the real causes. In short, there is a veridical (or at least sufficient) representation of both the causes and the dynamical structure of their mixing through the backward connections θ.

The dynamics of representational units or populations implied by Eq. (11) represents the essential difference between this class of approaches and those considered above. Only in predictive coding are the dynamics changing to minimise the same objective function as the parameters. In both the connectionist and infomax schemes the representations of a given cause can only be changed vicariously through the connection parameters. Predictive coding is a strategy that has some compelling (Bayesian) underpinnings (see further) and is not simply using a connectionist architecture in auto-associative mode or using error minimisation

[1] For simplicity, time constants have been omitted from expressions describing the ascent of states or parameters on objective functions.

to maximise information transfer. It is a real time, dynamical scheme that embeds two concurrent processes. (i) The parameters of the generative or forward model change to emulate the real world mixing of causes, using the current estimates; and (ii) these estimates change to best explain the observed inputs, using the current forward model. Both the parameters and the states change in an identical fashion to minimise prediction error. The predictive coding scheme eschews the problems associated with earlier schemes. It can easily accommodate nonlinear mixing of causes in the real world. It does not require this mixing to be invertible and needs only the sensory inputs. However, there is an outstanding problem:

- To finesse the inverse problem, posed by non-invertible generative models, regularisation constraints are required. These resolve the problem of non-invertibility that confounds simple infomax schemes but introduce a new problem. Namely one needs to know the prior distribution of the causes. This is because, as shown next, the regularisation constraints are based on these priors.

In summary, predictive coding treats representational learning as an ill-posed inverse problem and uses an explicit parameterisation of a forward model to generate predictions of the observed input. The ensuing error is then used to refine the forward model. This component of representational learning is dealt with below (Section 3.6). The predictions are based on estimated causes that also minimise predictive error, under some constraints that resolve the generally ill-posed estimation problem. We now consider these constraints from a Bayesian point of view.

3.4.2. *Predictive coding and Bayesian inference*

One important aspect of predictive coding and generative models (see further) is that they portray the brain as an inferential machine (Dayan et al., 1995). From this perspective, functional architectures exist, not to filter the input to obtain the causes, but to estimate causes and test the predictions against the observed input. A compelling aspect of predictive coding schemas is that they lend themselves to Bayesian treatment. This is important because it can be extended using empirical Bayes and hierarchical models. In what follows we shall first describe the Bayesian view of regularisation in terms of priors on the causes. We then consider hierarchical models in which priors can be derived empirically. The key implication, for neuronal implementations of predictive coding, is that empirical priors eschew assumptions about the independence of causes (cf. infomax schemes) or the form of constraints in regularised inverse solutions.

Suppose we knew the a priori distribution of the causes $p(v)$, but wanted the best estimate given the input. This maximum a posteriori (MAP) estimate maximises the posterior $p(v|u)$. The two probabilities are related through Bayes rule which states that the probability of the cause and input occurring together is the probability of the cause given the input times the probability of the input. This, in turn, is the same as the probability of the input given the causes times the prior probability of the causes.

$$p(u, v) = p(v|u)p(u) = p(u|v)p(v) \tag{12}$$

The MAP estimator of the causes is the most likely given the data.

$$v_m = \max_v \ln p(v|u) = \max_v[\ln p(u|v) + \ln p(v)] \tag{13}$$

The first term on the right is known as the log likelihood or likelihood potential and the second is the prior potential. A gradient ascent to find v_m would take the form

$$\begin{aligned} \dot{v} &= \frac{\partial \ell}{\partial v} \\ \ell(u) &= \ln p(u|v;\theta) + \ln p(v;\theta) \end{aligned} \tag{14}$$

where the dependence of the likelihood and priors on the model parameters has been made explicit. The likelihood is defined by the forward model $u = G(v, \theta) + \varepsilon$ where $p(u|v;\theta) \propto \exp(-V(\varepsilon, \theta))$. V now plays the role of a Gibb's potential that specifies ones distributional assumptions about the prediction error. Now we have

$$\dot{v} = -\frac{\partial V(\varepsilon, \theta)}{\partial v} + \frac{\partial \ln p(v;\theta)}{\partial v} \tag{15}$$

This is formally identical to the predictive coding scheme Eq. (11), in which the regularisation term $\lambda(v, \theta) = \ln p(v;\theta)$ becomes a log prior that renders the ensuing estimation Bayesian. In this formulation the state of the brain changes, not to minimise error per se, but to attain an estimate of the causes that maximises both the likelihood of the input given that estimate and the prior probability of the estimate being true. The implicit Bayesian estimation can be formalised from a number of different perspectives. Rao and Ballard (1998) give a very nice example using the Kalman filter that goes some way to dealing with the dynamical aspect of real sensory inputs.

3.5. *Cortical hierarchies and empirical Bayes*

The problem with Eq. (15) is that the brain cannot construct priors de novo. They have to be learned along with the forward model. In Bayesian estimation priors are estimated from data using *empirical* Bayes. Empirical Bayes harnesses the hierarchical structure of a forward model, treating the estimates of causes at one level as prior expectations for the subordinate level (Efron and Morris, 1973). This provides a natural framework within which to treat cortical hierarchies in the brain, each providing constraints on the level below. Fig. 2 depicts a hierarchical architecture that is described in more detail below. This extension models the world as a hierarchy of (dynamical) systems where supraordinate causes induce, and moderate, changes in subordinate causes. For example, the presence of a particular object in the visual field changes the incident light falling on a particular part of the retina. A more abstract example, that illustrates the brain's inferential capacities, is presented in Fig. 3. On reading the

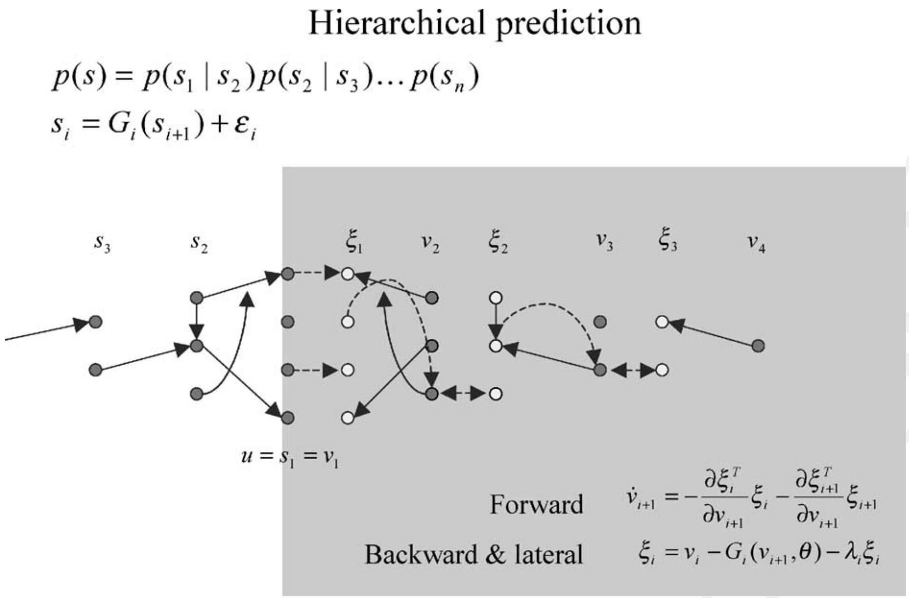
Hierarchical prediction
$p(s) = p(s_1 | s_2)p(s_2 | s_3)\ldots p(s_n)$
$s_i = G_i(s_{i+1}) + \varepsilon_i$
s_3
s_2
ξ_1
v_2
ξ_2
v_3
ξ_3
v_4
$u = s_1 = v_1$
Forward
$\dot{v}_{i+1} = -\frac{\partial \xi_i^T}{\partial v_{i+1}}\xi_i - \frac{\partial \xi_{i+1}^T}{\partial v_{i+1}}\xi_{i+1}$
Backward & lateral
$\xi_i = v_i - G_i(v_{i+1},\theta) - \lambda_i \xi_i$

Fig. 2. Schematic depicting a hierarchical extension to the predictive coding architecture, using the same format as Fig. 1. Here hierarchical arrangements within the model serve to provide predictions or priors to representations in the level below. The open circles are the error units and the filled circles are the representations of causes in the environment. These representations change to minimise both the discrepancy between their predicted value and the mismatch incurred by their own prediction of the representations in the level below. These two constraints correspond to prior and likelihood potentials, respectively (see main text).

first sentence ‘Jack and Jill went up the hill’ we perceive the word ‘event’ as ‘went’. In the absence of any hierarchical inference the best explanation for the pattern of visual stimulation incurred by the text is ‘event’. This would correspond to the maximum likelihood estimate of the word and would be the most appropriate in the absence of prior information about which is the most likely word. However, within hierarchical inference the semantic context provides top–down predictions to which the posterior estimate is accountable. When this prior biases in favour of ‘went’ we tolerate a small error as a lower level of visual analysis to minimise the overall prediction error at the visual and lexical level. This illustrates the role of higher level estimates in providing predictions or priors for subordinate levels. These priors offer contextual guidance towards the most likely cause of the input. Note that predictions at higher levels are subject to the same constraints, only the highest level, if there is one in the brain, is free to be directed solely by bottom–up influences (although there are always implicit priors). If the brain has evolved to recapitulate the casual structure of its environment, in terms of its sensory infrastructures, it is interesting to reflect on the possibility that our visual cortices reflect the hierarchical casual structure of our environment.

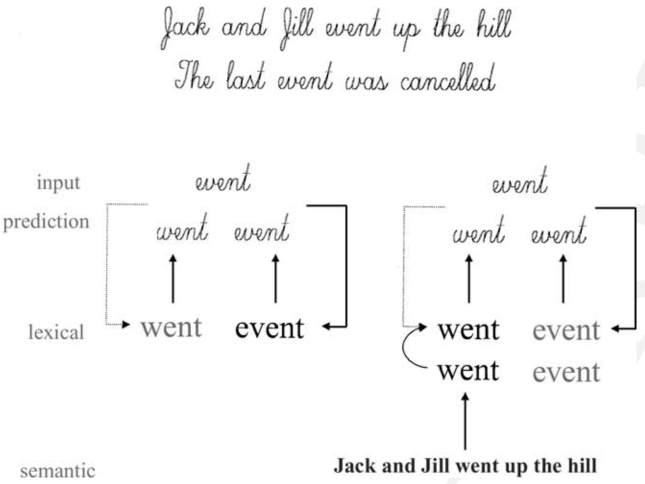
Jack and Jill went up the hill
The last event was cancelled
input
prediction
lexical
semantic
event
went
event
went
event
event
went
event
went
event
went
event
Jack and Jill went up the hill

Fig. 3. Schematic illustrating the role of priors in biasing towards one representation of an input or another. *Left*: The word ‘event’ is selected as the most likely cause of the visual input. *Right*: The word ‘went’ is selected as the most likely word that is: (i) a reasonable explanation for the sensory input; and (ii) conforms to prior expectations induced by semantic context.

The hierarchical structure of the real world is literally reflected by the hierarchical architectures trying to minimise prediction error, not just at the level of sensory input but at all levels of the hierarchy (notice the deliberate mirror symmetry in Fig. 2). The nice thing about this architecture is that the dynamics of casual representations at the ith level v_i require only the error for the current level and the immediately preceding level. This follows from the Markov property of hierarchical systems where one only needs to know the immediately supraordinate causes to determine the density of causes at any level in question, i.e. $p(v_i|v_{i+1}, \ldots, v_n) = p(v_i|v_{i+1})$. The fact that only error from the current and lower level is required to drive the dynamics of v_i is important because it permits a biologically plausible implementation, where the connections driving the error minimisation have only to run forward from one level to the next (see Section 3.5.1 and Fig. 2).

3.5.1. *Empirical Bayes in the brain*

The biological plausibility of the scheme depicted in Fig. 2 can be established fairly simply. To do this a hierarchical predictive scheme is described in some detail. A more thorough account of this scheme, including simulations of various neurobiological and psychophysical phenomena, will appear in future publications. For the moment, we will review neuronal implementation at a purely theoretical level, using the framework developed above.

Consider any level i in a cortical hierarchy containing units (neurons or neuronal populations) whose activity v_i is predicted by corresponding units in the level above v_{i+1}. The hierarchical form of the implicit generative model is

$$\begin{aligned} u &= G_1(v_2, \theta_1) + \varepsilon_1 \\ v_2 &= G_2(v_3, \theta_2) + \varepsilon_2 \\ v_3 &= \cdots \end{aligned} \tag{16}$$

with $v_1 = u$. Technically, these models fall into the class of conditionally independent hierarchical models when the error terms are independent at each level (Kass and Steffey, 1989). These models are also called *parametric empirical Bayes* (PEB) models because the obvious interpretation of the higher-level densities as priors led to the development of PEB methodology (Efron and Morris, 1973). We require units in all levels to jointly maximise the posterior probabilities of v_{i+1} given v_i. We will assume the errors are Gaussian with covariance $\sum_i = \sum(\lambda_i)$. Therefore, θ_i and λ_i parameterise the means and covariances of the likelihood at each level.

$$\begin{aligned} &p(v_i\,|v_{i+1}) = N(v_i : G(v_{i+1}, \theta_i), \textstyle\sum_i) \\ &\propto \left|\textstyle\sum_i\right|^{-1/2} \exp\left(-\tfrac{1}{2}\varepsilon_i^T \textstyle\sum_i^{-1} \varepsilon_i\right) \end{aligned} \tag{17}$$

This is also the prior density for the level below. Although θ_i and λ_i are both parameters of the forward model λ_i are sometimes referred to as hyperparameters and in classical statistics correspond to variance components. We will preserve the distinction between parameters and hyperparameters because minimising the prediction error with respect to the estimated causes and parameters is sufficient to maximise the likelihood of neuronal states at all levels. This is the essence of predictive coding. For the hyperparameters there is an additional term that depends on the hyperparameters themselves (see further).

In this hierarchical setting, the objective function comprises a series of log likelihoods

$$\begin{aligned} &\ell(u) = \ln p(u|v_1) + \ln(v_1|v_2) + \cdots + \ell(u) = -\tfrac{1}{2}\xi_1^T\xi_1 - \tfrac{1}{2}\xi_2^T\xi_2 - \cdots - \tfrac{1}{2}\ln\left|\textstyle\sum_1\right| - \tfrac{1}{2}\ln\left|\textstyle\sum_2\right| - \cdots \\ &\xi_i = v_i - G_i(v_{i+1}, \theta) - \lambda_i\xi_i = (1 + \lambda_i)^{-1}\varepsilon_i \end{aligned} \tag{18}$$

Here $\sum(\lambda_i)^{1/2} = 1 + \lambda_i$. The likelihood at each level corresponds to $p(v_i|v_{i+1})$ *which also plays the role of a prior on* v_i that is jointly maximised with the likelihood of the level below $p(v_{i-1}|v_i)$. In a neuronal setting the (whitened) prediction error is encoded by the activities of units denoted by ξ_i. These error units receive a prediction from units in the level above[2] and connections from the principal units v_i being predicted. Horizontal interactions among the error units serve to de-correlate them (cf. Foldiak, 1990), where the symmetric lateral connection strengths λ_i hyper-parameterise the covariances of the errors $\sum_i$ which are the prior covariances for level $i-1$.

The estimators v_{i+1} and the connection strength parameters perform a gradient ascent on the compound log probability.

$$\begin{aligned} \dot{v}_{i+1} &= \frac{\partial\ell}{\partial v_{i+1}} = -\frac{\partial\xi_i^T}{\partial v_{i+1}}\xi_i - \frac{\partial\xi_{i+1}^T}{\partial v_{i+1}}\xi_{i+1} \\ \dot{\theta}_i &= \frac{\partial\ell}{\partial\theta_i} = -\frac{\partial\xi_i^T}{\partial\theta_i}\xi \\ \dot{\lambda}_i &= \frac{\partial\ell}{\partial\lambda_i} = -\frac{\partial\xi_i^T}{\partial\lambda_i}\xi - (1 + \lambda_i)^{-1} \end{aligned} \tag{19}$$

When $G_i(v_{i+1}, \theta)$ models dynamical processes (i.e. is effectively a convolution operator) this gradient ascent is more complicated. In a subsequent paper we will show that, with dynamical models, it is necessary to maximise both ℓ and its temporal derivatives (e.g. $\dot{\ell}$). An alternative is to assume a simple hidden Markov model for the dynamics and use Kalman filtering (cf. Rao and Ballard, 1998). For the moment, we will assume the inputs change sufficiently slowly for gradient ascent not to be confounded.

Despite the complicated nature of the hierarchical model and the abstract theorising, three simple and biologically plausible things emerge:

- *Reciprocal connections*

  The dynamics of representational units v_{i+1} are subject to two, locally available, influences. A likelihood term mediated by forward afferents from the error units in the level below and an empirical prior term conveyed by error units in the same level. This follows from the conditional independence conferred by the hierarchical structure of the model. Critically, the influences of the error units in both levels are meditated by linear connections with a strength that is exactly the same as the (negative) effective connectivity of the reciprocal connection from v_{i+1} to ξ_i and ξ_{i+1} (see Box 1 for definition of effective connectivity). In short, the lateral, forwards and backward connections are all reciprocal, consistent with anatomical observations. Lateral connections, within each level

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

decorrelate the error units allowing competition between prior expectations with different precisions (precision is the inverse of variance).

- *Functionally asymmetric forward and backward connections*

  The forward connections are the reciprocal (negative transpose) of the backward effective connectivity $\partial\xi_i/\partial v_{i+1}$ from the higher level to the lower level, extant at that time. However, the functional attributes of the forward and backward influences are different. The influences of units on error units in the lower level mediate the forward model $\xi_i = -G_i(v_{i+1}, \theta) + \dots$. These can be nonlinear, where each unit in the higher level *may modulate or interact with the influence of others* (according to the nonlinearities in G). In contradistinction, *the influences of units in lower levels do not interact* when producing changes in the higher level because their effects are linearly separable $\dot{v}_{i+1} = -\partial\xi_i/\partial v_{i+1}\xi_i - \cdots$. This is a key observation because the empirical evidence, reviewed in the previous section, suggests that backward connections are in a position to interact (e.g. though NMDA receptors expressed predominantly in the supragranular layers receiving backward connections) whereas forward connections are not. It should be noted that, although the implied forward connections $\partial\xi_i/\partial v_{i+1}$ mediate linearly separable effects of ξ_i on v_{i+1}, these connections might be activity- and time-dependent because of their dependence on v_{i+1}.

- *Associative plasticity*

  Changes in the parameters correspond to plasticity in the sense that the parameters control the strength of backward and lateral connections. The backward connections parameterise the prior expectations of the forward model and the lateral connections hyper-parameterise the prior covariances. Together they parameterise the Gaussian densities that constitute the priors (and likelihoods) of the model. The motivation for these parameters maximising the same objective function ℓ as the neuronal states is discussed in the next subsection. For the moment, we are concerned with the biological plausibility of these changes. The plasticity implied is seen more clearly with an explicit parameterisation of the connections. For example, let $G_i(v_{i+1}, \theta_i) = \theta_i v_{i+1}$. In this instance

$$\begin{aligned}\dot{\theta}_i &= (1+\lambda_i)^{-1}\xi_i v_{i+1}^T \\ \dot{\lambda}_i &= (1+\lambda_i)^{-1}(\xi_i\xi_i^T - 1)\end{aligned} \tag{20}$$

This is just Hebbian or associative plasticity where the connection strengths change in proportion to the product of pre and post-synaptic activity. An intuition about Eq. (20) 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 casual estimates at the higher level $\langle\xi_i v_{i+1}^T\rangle = 0$. The lateral connections stop changing when the prediction error has been whitened $\langle\xi_i\xi_i^T\rangle = 1$.

Non-diagonal forms for λ_i complicate the biological interpretation because changes at any one connection depend on changes elsewhere. The problem can be finessed slightly by rewriting the equations as

$$\begin{aligned}\dot{\theta}_i &= \xi_i v_{i+1}^T - \lambda_i\dot{\theta}_i \\ \dot{\lambda}_i &= \xi_i\xi_i^T - \lambda_i\dot{\lambda}_i - 1\end{aligned} \tag{21}$$

where the decay terms are mediated by integration at the cell body in a fashion similar to that described in Friston et al. (1993).

The overall scheme implied by Eq. (19) sits comfortably the hypothesis (Mumford, 1992). "On the role of the reciprocal, topographic pathways between two cortical areas, one often a 'higher' area dealing with more abstract information about the world, the other 'lower', dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top–down, bottom–up loops".

In summary, the predictive coding approach 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 minimise error at all levels of the cortical hierarchy and implicitly render themselves posterior estimates of the causes given the data. In contradistinction to connectionist schemas, hierarchical prediction does not require any desired output. Indeed predictions of intermediate outputs at each hierarchical level emerge spontaneously. Unlike information theoretic approaches they do not assume independent causes and invertible generative processes. In contrast to regularised inverse solutions (e.g. in machine vision) they do not depend on a priori constraints. These emerge spontaneously as empirical priors from higher levels. The Bayesian considerations above pertain largely to the estimates of the causes. In the final subsection we consider the estimation of model parameters using the framework provided by density learning with generative models.

3.6. Generative models and representational learning

In this section we bring together the various schemes considered above using the framework provided by density estimation as a way of fitting generative models. This section follows Dayan and Abbott (2001) to which the reader is referred for a fuller discussion. Generative models represent a generic formulation of representational leaning in a self-supervised context. There are many forms of generative models that range from conventional statistical models (e.g. factor and cluster analysis) and those motivated by Bayesian inference and learning (e.g. Dayan et al., 1995; Hinton et al., 1995). Indeed many of the algorithms discussed under the heading of information theory can be formulated as generative models. 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). In current treatments, representational learning is framed in terms of estimating probability densities of the inputs and outputs. Although density learning is formulated at a level of abstraction that eschews many issues of neuronal implementation (e.g. the dynamics of real-time learning), it does provide a unifying framework that connects the various schemes considered so far.

The goal of generative models is to make the density of the inputs, implied by the generative model $p(u;\theta)$, as close as possible to those observed $p(u)$. The generative model is specified in terms of the prior distribution over the causes $p(u;\ \theta)$ and the conditional *generative* distribution of the inputs given the causes $p(u|v;\theta)$ which together define the marginal distribution that has to be matched to the input distribution

$$p(u;\theta)=\int p(u|v;\theta)\,p(v;\theta)\,\mathrm{d}v \tag{22}$$

Once the parameters of the generative model have been estimated, through this matching, the posterior density of the causes, given the inputs are given by the recognition model defined in terms of the *recognition* distribution

$$p(v|u;\theta)=\frac{p(u|v;\theta)\,p(v;\theta)}{p(u;\theta)} \tag{23}$$

However, as considered in depth above, the generative model may not be invertible and it may not be possible to compute the recognition distribution from Eq. (23). In this instance, an approximate recognition distribution can be used $q(v;u,\phi)$ that we try to approximate to the true one. The distribution has some parameters ϕ that need to be learned, for example, the strength of forward connections. The question addressed in this review is whether forward connections are sufficient for representational leaning. For a moment, consider deterministic models that discount probabilistic or stochastic aspects. We have been asking whether we can find the parameters of a deterministic recognition model that renders it the inverse of a generating process

$$v(u,\phi)=G^{-1}(u,\theta) \tag{24}$$

The problem is that $G(v,\theta)$ is a nonlinear convolution and is generally not invertible. The generative model approach posits that it is sufficient to find the parameters of an (approximate) recognition model ϕ and the generative model θ that predict the inputs

$$G(v(u,\phi),\theta)=u \tag{25}$$

under the constraint that the recognition model is (approximately) the inverse of the generative model. Eq. (25) is the same as Eq. (24) after applying G to both sides. The implication is that one needs an explicit parameterisation of the (approximate) recognition (inverse) model and generative (forward) models that induces the need for both forward and backward influences. Separate recognition and generative models resolve the problem caused by the non-invertibility of generating processes. The corresponding motivation, in probabilistic learning, rests on finessing the combinatorial explosion of ways in which stochastic generative models can generate input patterns (Dayan et al., 1995). The combinatorial explosion represents another perspective on the uninvertible 'many to one' relationship between causes and inputs.

In the general density learning framework, representational learning has two components that can be seen in terms of expectation maximisation (EM, Dempster et al., 1977). In the **E**-Step the approximate recognition distribution is modified to match the density implied by the generative model parameters, so that $q(v;u,\phi)\approx p(v|u;\theta)$ and in the **M**-Step these parameters are changed to render $p(u;\theta)\approx p(u)$. In other words, the **E**-Step ensures the recognition model approximates the generative model and the **M**-Step ensures that the generative model can predict the observed inputs. If the model is invertible the **E**-Step reduces to setting $q(v;u,\phi)=p(v|u;\theta)$ using Eq. (23). Probabilistic recognition proceeds by using $q(v;u,\phi)$ to determine the probability that v caused the observed sensory inputs. This recognition becomes deterministic when $q(v;u,\phi)$ is a Dirac δ-function over the MAP estimator of the causes v_m. The distinction between probabilistic and deterministic recognition is important because we have restricted ourselves to deterministic models thus far but these are special cases of density estimation in generative modelling.

3.6.1. Density estimation and EM

EM provides a useful procedure for density estimation that helps relate many different models within a framework that has direct connections with statistical mechanics. Both steps of the EM algorithm involve maximising a function of the densities that corresponds to the negative free energy in physics.

$$\begin{aligned}F(\phi,\theta)&=\left\langle\int q(v;u,\phi)\ln\frac{p(v,u;\theta)}{q(v;u,\phi)}\,\mathrm{d}v\right\rangle_u\\&=\langle\ln p(u;\theta)\rangle_u-\langle KL(q(v;u,\phi),\,p(v|u;\theta))\rangle_u\end{aligned} \tag{26}$$

This objective function comprises two terms. The first is the expected log likelihood of the inputs, under the generative model, over the observed inputs. Maximising this term implicitly minimises the Kullback–Leibler (KL) divergence[3] between the actual input density and that implied by the generative model. This is equivalent to maximising the log like-

[3] A measure of the discrepancy between two densities.

lihood of the inputs. The second term is the KL divergence between the approximating and true recognition densities. In short, maximising F encompasses two components of representational learning: (i) it increases the likelihood that the generative model could have produced the inputs; and (ii) minimises the discrepancy between the approximate recognition model and that implied by the generative model. The **E**-Step increases F with respect to the recognition parameters ϕ through minimising the KL term, ensuring a veridical approximation to the recognition distribution implied by θ. The **M**-Step increases F by changing θ, enabling the generative model to reproduce the inputs.

$$\begin{aligned} \mathbf{E}: &\quad \phi = \min_{\phi} F(\phi, \theta) \\ \mathbf{M}: &\quad \theta = \min_{\theta} F(\phi, \theta) \end{aligned} \tag{27}$$

This formulation of representational leaning is critical for the thesis of this review because it shows that backward connections, parameterising a generative model, are essential when the model is not invertible. If the generative model is invertible then the KL term can be discounted and learning reduces to the **M**-Step (i.e. maximising the likelihood). In principle, this could be done using a feedforward architecture corresponding to the inverse of the generative model. However, when processes generating inputs are non-invertible (due to nonlinear interactions among, and temporal convolutions of, the causes) a parameterisation of the generative model (backward connections) and approximate recognition model (forward connections) is required that can be updated in **M**- and **E**-Steps, respectively. In short, non-invertibility enforces an explicit parameterisation of the generative model in representational learning. In the brain this parameterisation may be embodied in backward and lateral connections.

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

$$\begin{aligned} &p(v; \theta) = N(v : 0, 1) \\ &p(u \,|v \,; \theta) = N(u : \theta v, \textstyle\sum) \end{aligned} \tag{28}$$

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

$$\begin{aligned} &p(v; \theta) = \textstyle\prod p(v_i, \theta) \\ &p(u \,|v \,; \theta) = N(u : \theta v, \textstyle\sum) \end{aligned} \tag{29}$$

where $p(v_i\theta)$ are chosen to be suitably sparse (i.e. heavy-tailed) with a cumulative density function that corresponds to the squashing function in Section 3.3.1. The deterministic equivalent of sparse coding is ICA that obtains when $\sum \to 0$. The relationships among different models are rendered apparent under the perspective of generative models. It is useful to revisit the schemes above to examine their implicit generative and recognition models.

3.6.2. *Supervised representational learning*

In supervised schemes the generative model is already known and only the recognition model needs to be estimated. The generative model is known in the sense that the desired output determines the input either deterministically or stochastically (e.g. the input primitives are completely specified by their cause, which is the desired output). In this case only the **E**-Step is required in which the parameters ϕ that specify $q(v; u, \phi)$ change to maximise F. The only term in Eq. (26) that depends on ϕ is the divergence term, such that learning reduces to minimising the expected difference between the approximate recognition density and that required by the generative model. This can proceed probabilistically (e.g. Contrastive Hebbian learning in stochastic networks (Abbott and Dayan, 2001, p. 322)) or deterministically. In the deterministic mode $q(v; u, \phi)$ corresponds to a δ-function over the point estimator $v_m = R(u, \phi)$. The connection strengths ϕ are changed, typically using the delta rule, such that the distance between the modes of the approximate and desired recognition distributions are minimised over all inputs. This is equivalent to nonlinear function approximation; a perspective that can be adopted on all supervised learning of deterministic mappings with neural nets.

Note, again, that any scheme, based on supervised learning, requires the processes generating inputs to be known a priori and as such cannot be used by the brain.

3.6.3. *Information theory*

In section on information theory we had considered whether infomax principles were sufficient to specify deterministic recognition architectures, in the absence of backward connections. They were introduced in terms of finding some function of the inputs that produces an output density with maximum entropy. Maximisation of F attains the same thing through minimising the discrepancy between the observed input distribution $p(u)$ and that implied by a generative model with maximum entropy priors. Although the infomax and density learning approaches have the same objective their heuristics are complementary. Infomax is motivated by maximising the mutual information between u and v under some constraints. The generative model approach takes its heuristics from the assumption that the causes of inputs are independent and possibly non-Gaussian. This results in a prior with maximum entropy $p(v; \theta) = \prod p(v_i; \theta)$. The reason for adopting non-Gaussian priors (e.g. sparse coding and ICA) is that the central limit theorem implies mixtures of causes will have Gaussian distributions and therefore something that is not Gaussian is unlikely to be a mixture.

For invertible deterministic models $v = R(u, \phi) = G^{-1}(u, \theta)$ the KL component of F disappears leaving only the likelihood term.

$$\begin{aligned} F &= \langle \ln p(u;\theta)\rangle_u = \langle \ln p(v;\theta)\rangle_u + \langle \ln p(u|v;\theta)\rangle_u \\ &= \left\langle \ln \prod p(v_i;\theta)\right\rangle_u + \left\langle \ln \left|\frac{\partial R(u,\phi)}{\partial u}\right|\right\rangle_u \\ &= -\sum H(v_i;\theta) + H(v;\phi) - H(u) \qquad (30) \end{aligned}$$

This has exactly the same dependence on the parameters as the objective function employed by infomax in Eq. (7). In this context, the free energy and the information differ only by the entropy of the inputs $-F = I + H(u)$. This equivalence rests on uses maximum entropy priors of the sort assumed for sparse coding.

Notice again that, in the context of invertible deterministic generative models, the parameters of the recognition model specify the generative model and only the recognition model (i.e. forward connections meditating $v = R(u, \phi)$) needs to be instantiated. If the generative modal cannot be inverted the recognition model is not defined and the scheme above is precluded. In this instance one has to parameterise both an approximate recognition and generative model as required by EM. This enables the use of nonlinear generative models, such as nonlinear PCA (e.g. Kramer, 1991; Karhunen and Joutsensalo, 1994; Dong and McAvoy, 1996; Taleb and Jutten, 1997). These schemes typically employ a 'bottleneck' architecture that forces the inputs through a small number of nodes. The output from these nodes then diverges to produce the predicted inputs. The approximate recognition model is implemented, deterministically in connections to the bottleneck nodes and the generative model by connection from these nodes to the outputs. Nonlinear transformations, from the bottleneck nodes to the output layer, recapitulate the nonlinear mixing of the real causes of the inputs. After learning, the activity of the bottleneck nodes can be treated as estimates of the causes. These representations obtain by projection of the input onto a low-dimensional curvilinear manifold (encompassing the activity of the bottleneck nodes) by an approximate recognition model.

3.6.4. Predictive coding

In the forgoing, density learning is based on the expectations of probability distributions over the inputs. Clearly the brain does not have direct access to these expectations but sees only one input at any instant. In this instance representational learning has to proceed on-line, by sampling inputs over time.

For deterministic recognition models, $q(v; u, \phi)$ is parameterised by its input-specific mode $v(u)$, where $q(v(u); u) = 1$ and

$$\begin{aligned} \ell(u) &= \int q(v;u,\phi) \ln \frac{p(v,u;\theta)}{q(v;u,\phi)}\, dv = \ln p(v(u), u;\theta) \\ &= \ln p(u|v(u);\theta) + \ln p(v(u);\theta) \qquad (31) \\ F &= \langle \ell(u)\rangle_u \end{aligned}$$

$\ell(u)$ is simply the log of the joint probability, under the generative model, of the observed inputs and their cause, implied by approximate recognition. This log probability can be decomposed into a log likelihood and log prior and is exactly the same objective function used to find the MAP estimator in predictive coding cf. Eq. (14).

On-line representational learning can be thought of as comprising two components, corresponding to the **E** and **M**-Steps. The expectation (**E**) component updates the recognition density, whose mode is encoded by the neuronal activity v, by maximising $\ell(u)$. Maximising $\ell(u)$ is sufficient to maximise its expectation F over inputs because it is maximised for each input separately. The maximisation (**M**) component corresponds to an ascent of these parameters, encoded by the connection strengths, on the same log probability

$$\begin{aligned} E:&\quad \dot{\phi} = \dot{v} = \frac{\partial \ell}{\partial v} \\ M:&\quad \dot{\theta} = \frac{\partial \ell}{\partial \theta} \end{aligned} \qquad (32)$$

such that the expected change approximates[4] an ascent on F; $\langle\dot{\theta}\rangle \approx \langle \partial\ell/\partial\theta\rangle_u = \partial F/\partial\theta$. Eq. (32) is formally identical to Eq. (19), the hierarchical prediction scheme, where the hyperparameters have been absorbed into the parameters. In short, predictive coding can be regarded as an on-line or dynamic form of density estimation using a deterministic recognition model and a stochastic generative model. Conjoint changes in neuronal states and connection strengths map to the expectation maximisation of the approximate recognition and generative models, respectively. Note that there is no explicit parameterisation of the recognition model; the recognition density is simply represented by its mode for the input u at a particular time. This affords a very unconstrained recognition model that can, in principle, approximate the inverse of highly nonlinear generative models.

3.7. Summary

In summary, the formulation of representational learning in terms of generative models embodies a number of key distinctions: (i) the distinction between invertible versus non-invertible models; (ii) deterministic versus probabilistic representations; and (iii) dynamic versus density learning.

Non-invertible generative models require their explicit parameterisation and suggest an important role for backward connections in the brain. Invertible models can, in principle be implemented using only forward connections because the recognition model completely specifies the generative model and vice versa. However, nonlinear and dynamic aspects of the sensorium render invertibility highly unlikely.

[4] This approximation can be finessed by using traces, to approximate the expectation explicitly, and changing the connections in proportion with the trace.

This section has focused on the conditions under which forward connections are sufficient to parameterise a generative model. In short, these conditions rest on invertibility and speak to the need for backward connections in the context of nonlinear and noninvertible generative models.

Most of the examples in this section have focussed on deterministic recognition models where neuronal dynamics encode the most likely causes of the current sensory input. This is largely because we have been concerned with how the brain represents things. The distinction between deterministic and probabilistic representation addresses a deeper question about whether neuronal dynamics represent the state of the world or the probability densities of those states. From the point of view of hierarchical models the state of the neuronal units encodes the mode of the posterior density at any given level. This can be considered a point recognition density. However, the states of units at any level also induce a prior density in the level below. This is because the prior mode is specified by dynamic top–down influences and the prior covariance by the strength of lateral connections. These covariances render the generative model a probabilistic one.

By encoding densities in terms of their modes, using neuronal activity, the posterior and prior densities can change quickly with sensory inputs. However, this does entail unimodal densities. From the point of view of a statistician this may be an impoverished representation of the world that compromises any proper inference, especially when the posterior distribution is multimodal. However, it is exactly this approximate nature of recognition that pre-occupies psychophysicists and psychologists; The emergence of unitary, deterministic perceptual representations in the brain is commonplace and is of special interest when the causes are ambiguous (e.g. illusions and perceptual transitions induced by binocular rivalry and ambiguous figures).

The brain is a dynamical system that samples inputs dynamically over time. It does not have instantaneous access to the statistics of its inputs that are required for distinct **E**- and **M**-Steps. Representational learning therefore has to proceed under this constraint. In this review, hierarchical predictive coding has been portrayed as a variant of density leaning that conforms to these constraints.

We have seen that supervised, infomax and generative models require prior assumptions about the distribution of causes. This section introduced empirical Bayes to show that these assumptions are not necessary and that priors can be learned in a hierarchical context. Furthermore, we have tried to show that hierarchical prediction can be implemented in brain-like architectures using mechanisms that are biologically plausible.

4. Generative models and the brain

The arguments in the preceding section clearly favour predictive coding, over supervised or information theoretic frameworks, as a more plausible account of functional brain architectures. However, it should be noted that the differences among them have been deliberately emphasised. For example, predictive coding and the implicit error minimisation results in the maximisation of information transfer. In other words, predictive coding conforms to the principle of maximum information transfer, but in a distinct way. Predictive coding is entirely consistent with the principle of maximum information. The infomax principle is a principle, whereas predictive coding represents a particular scheme that serves that principle. There are examples of infomax that do not employ predictive coding (e.g. transformations of stimulus energy in early visual processing; Atick and Redlich, 1990) that may be specified genetically or epigenetically. However, predictive coding is likely to play a much more prominent role at higher levels of processing for the reasons detailed in the previous section.

In a similar way predictive coding, especially in its hierarchical formulation, conforms to the same PDP principles that underpin connectionist schemes. The representation of any cause depends upon the internally consistent representations of subordinate and supraordinate causes in lower and higher levels. These representations mutually induce and maintain themselves, across and within all levels of the sensory hierarchy, through dynamic and reentrant interactions (Edelman, 1993). The same PDP phenomena (e.g. lateral interactions leading to competition among representations) can be observed. For example, the lateral connection strengths embody what has been learnt empirically about the prior covariances among causes. A prior that transpires to be very precise (i.e. low variance) will receive correspondingly low strength inhibitory connections from its competing error units (recall $\sum(\lambda_i)^{1/2} = 1 + \lambda_i$). It will therefore supervene over other error units and have a greater corrective impact on the estimate causing the prediction error. Conversely, top–down expectations that are less informative will induce errors that are more easily suppressed and have less effect on the representations. In predictive coding, these dynamics are driven explicitly by error minimisation, whereas in connectionist simulations the activity is determined solely by the connection strengths established during training.

In addition to the theoretical bias toward generative models and predictive coding, the clear emphasis on backward and reentrant (Edelman, 1993) dynamics make it a more natural framework for understanding neuronal infrastructures. Fig. 1 shows the fundamental difference between infomax and generative schemes. In the infomax schemes the connections are universally forward. In the predictive coding scheme the forward connections (broken line) drive the prediction so as to minimise error whereas backwards connections (solid lines) use these representations of causes to emulate mixing enacted by the real world. The nonlinear aspects of this mixing imply that only backward influences interact in the predictive coding scheme whereas the nonlinear *un*mixing, in classical infomax schemas, is mediated by forward connections. Section 2 assembled some of the anatomical and physiological evidence suggesting that backward connections are prevalent in the real brain and could support nonlinear mixing through their modulatory characteristics. It is pleasing that purely theoretical considerations and neurobiological empiricism converge on the same architecture. Before turning to electrophysiological and functional neuroimaging evidence for backward connections we consider the implications for classical views of receptive fields and the representational capacity of neuronal units.

4.1. Context, causes and representations

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

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

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

4.2. Neuronal responses and representations

Classical models (e.g. classical receptive fields) assume that evoked responses will be expressed invariably in the same units or neuronal populations irrespective of the context. However, real neuronal responses are not invariant but depend upon the context in which they are evoked. For example, visual cortical units have dynamic receptive fields that can change from moment to moment (cf. the non-classical receptive field effects modelled in (Rao and Ballard, 1998)). Another example is attentional modulation of evoked responses that can change the sensitivity of neurons to different perceptual attributes (e.g. Treue and Maunsell, 1996). The evidence for contextual responses comes from neuroanatomical and electrophysiological studies. There are numerous examples of context-sensitive neuronal responses. Perhaps the simplest is short-term plasticity. Short-term plasticity refers to changes in connection strength, either potentiation or depression, following pre-synaptic inputs (e.g. Abbot et al., 1997). In brief, the underlying connection strengths, that define what a unit represents, are a strong function of the immediately preceding neuronal transient (i.e. preceding representation). A second, and possibly richer, example is that of attentional modulation. It has been shown, both in single unit recordings in primates (Treue and Maunsell, 1996) and human functional fMRI studies (Büchel and Friston, 1997), that attention to specific visual attributes can profoundly alter the receptive fields or event-related responses to the same stimuli.

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

Instead, the *neural context* in which an area is active may define the cognitive function.” His argument is predicated on careful characterisations of effective connectivity using neuroimaging.

4.2.1. Examples from electrophysiology

In the next section we will illustrate the context-sensitive nature of cortical activations, and implicit specialisation, in the inferior temporal lobe using neuroimaging. Here we consider the evidence for contextual representations in terms of single cell responses, to visual stimuli, in the temporal cortex of awake behaving monkeys. If the representation of a stimulus depends on establishing representations of subordinate and supraordinate causes at all levels of the visual hierarchy, then information about the high-order attributes of a stimulus, must be conferred by top–down influences. Consequently, one might expect to see the emergence of selectivity, for high-level attributes, *after* the initial visually evoked response (it typically takes about 10 ms for volleys of spikes to be propagated from one cortical area to another and about a 100 ms to reach prefrontal areas). This is because the representations at higher levels must emerge before backward afferents can reshape the response profile of neurons in lower areas. This temporal delay, in the emergence of selectivity, is precisely what one sees empirically: Sugase et al. (1999) recorded neurons in macaque temporal cortex during the presentation of faces and objects. The faces were either human or monkey faces and were categorised in terms of identity (whose face it was) and expression (happy, angry, etc.). “Single neurones conveyed two different scales of facial information in their firing patterns, starting at different latencies. Global information, categorising stimuli as monkey faces, human faces or shapes, was conveyed in the earliest part of the responses. Fine information about identity or expression was conveyed later”, starting on average about 50 ms after face-selective responses. These observations demonstrate representations for facial identity or expression that emerge dynamically in a way that might rely on backward connections. These influences imbue neurons with a selectivity that is not intrinsic to the area but depends on interactions across levels of a processing hierarchy.

A similar late emergence of selectivity is seen in motion processing. A critical aspect of visual processing is the integration of local motion signals generated by moving objects. This process is complicated by the fact that local velocity measurements can differ depending on contour orientation and spatial position. Specifically, any local motion detector can measure only the component of motion perpendicular to a contour that extends beyond its field of view (Pack and Born, 2001). This “aperture problem” is particularly relevant to direction-selective neurons early in the visual pathways, where small receptive fields permit only a limited view of a moving object. Pack and Born (2001) have shown “that neurons in the middle temporal visual area (known as MT or V5) of the macaque brain reveal a dynamic solution to the aperture problem. MT neurons initially respond primarily to the component of motion perpendicular to a contour’s orientation, but over a period of approximately 60 ms the responses gradually shift to encode the true stimulus direction, regardless of orientation”.

The preceding examples were taken from electrophysiology. Similar predictions can be made, albeit at a less refined level, about population responses elicited in functional neuroimaging where functional specialisation (cf. selectivity in unit recordings) is established by showing regionally-specific responses to some sensorimotor attribute or cognitive component. At the level of cortical responses in neuroimaging the dynamic and contextual nature of evoked responses means that regionally-specific responses to a particular cognitive component may be expressed in one context but not another. In the next section we look at some empirical evidence from functional neuroimaging that confirms the idea that functional specialisation is conferred in a context-sensitive fashion by backwards connections from higher brain areas.

5. Functional architectures assessed with brain imaging

Information theory and predictive coding schemas suggest alternative architectures that are sufficient for representational learning. Forward connections are sufficient for the former, whereas the latter posits that most of the brain’s infrastructure is used to predict sensory input through a hierarchy of top–down projections. Clearly to adjudicate between these alternatives the existence of backward influences must be established. This is a slightly deeper problem for functional neuroimaging than might be envisaged. This is because making causal inferences about effective connectivity is not straightforward (see Pearl, 2000). It might be thought that showing regional activity was partially predicted by activity in a higher level would be sufficient to confirm the existence of backward influences, at least at a population level. The problem is that this statistical dependency does not permit any causal inference. Statistical dependencies could easily arise in a purely forward architecture because the higher level activity is predicated on activity in the lower level. One resolution of this problem is to perturb the higher level directly using transmagnetic stimulation or pathological disruptions (see Section 6). However, discounting these interventions, one is left with the difficult problem of inferring backward influences, based on measures that could be correlated because of forward connections. Although there are causal modelling techniques that can address this problem we will take a simpler approach and note that interactions between bottom–up and top–down influences cannot be explained by a purely feedforward architecture. This is because the top–down influences have no access to the bottom–up inputs. An interaction, in this context, can be construed as an effect of backward connections on the driving efficacy of forward connections. In other words, the response evoked by the same driving bottom–up inputs depends upon the context established by top–down inputs. This interaction is used below simply as evidence for the existence of backward influences. However, there are some instances of predictive coding that emphasises this phenomenon. For example, the "Kalman filter model demonstrates how certain forms of attention can be viewed as an emergent property of the interaction between top–down expectations and bottom–up signals" (Rao, 1999).

The remainder of this article focuses on the evidence for these interactions. From the point of view of functionally specialised responses these interactions manifest as context-sensitive or contextual specialisation, where modality-, category- or exemplar-specific responses, driven by bottom up inputs are modulated by top–down influences induced by perceptual set. The first half of this section adopts this perceptive. The second part of this section uses measurements of effective connectivity to establish interactions between bottom–up and top–down influences. All the examples presented below rely on attempts to establish interactions by trying to change sensory-evoked neuronal responses through putative manipulations of top–down influences. These include inducing independent changes in perceptual set, cognitive (attentional) set and, in the last section through the study of patients with brain lesions.

5.1. Context-sensitive specialisation

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

5.1.1. Categorical designs

Categorical designs, such as cognitive subtraction, have been the mainstay of functional neuroimaging over the past decade. Cognitive subtraction involves elaborating two tasks that differ in a separable component. Ensuing differences in brain activity are then attributed to this component. The tenet of cognitive subtraction is that the difference between two tasks can be formulated as a separable cognitive or sensorimotor component and that the regionally specific differences in hemodynamic responses identify the corresponding functionally specialised area. Early applications of subtraction range from the functional anatomy of word processing (Petersen et al., 1989) to functional specialisation in extrastriate cortex (Lueck et al., 1989). The latter studies involved presenting visual stimuli with and without some sensory attribute (e.g. colour, motion etc.). The areas highlighted by subtraction were identified with homologous areas in monkeys that showed selective electrophysiological responses to equivalent visual stimuli.

Consider a specific example; namely the difference between simply saying "yes" when a recognisable object is seen, and saying "yes" when an unrecognisable non-object is seen. Regionally specific differences in brain activity that distinguish between these two tasks could be implicated in implicit object recognition. Although its simplicity is appealing this approach embodies some strong assumptions about the way that the brain implements cognitive processes. A key assumption is 'pure insertion'. Pure insertion asserts that one can insert a new component into a task without effecting the implementation of pre-existing components (for example, how do we know that object recognition is not itself affected by saying "yes"?). The fallibility of this assumption has been acknowledged for decades, perhaps most explicitly by Sternberg's revision of Donder's subtractive method. The problem for subtraction is as follows: if one develops a task by adding a component then the new task comprises not only the previous components and the new component but the integration of the new and old components (for example, the integration of phonology and object recognition). This integration or *interaction* can itself be considered as a new component. The difference between two tasks therefore includes the new component and the interactions between the new component and those of the original task. Pure insertion requires that all these interaction terms are negligible. Clearly in many instances they are not. We next consider factorial designs that eschew the assumption of pure insertion.

5.1.2. 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 repeating the above implicit object recognition experiment in another context, for example naming (of the object's name or the non-object's colour). 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. Generally, interactions can be thought of as a difference in activations brought about by another processing demand. Dual task interference paradigms are a clear example of this approach (e.g. Fletcher et al., 1995).

Consider the above object recognition experiment again. Noting that object-specific responses are elicited (by asking subjects to view objects relative to meaningless shapes), with and without phonological retrieval, reveals the factorial nature of this experiment. This 'two by two' design allows one to look specifically at the interaction between phono-

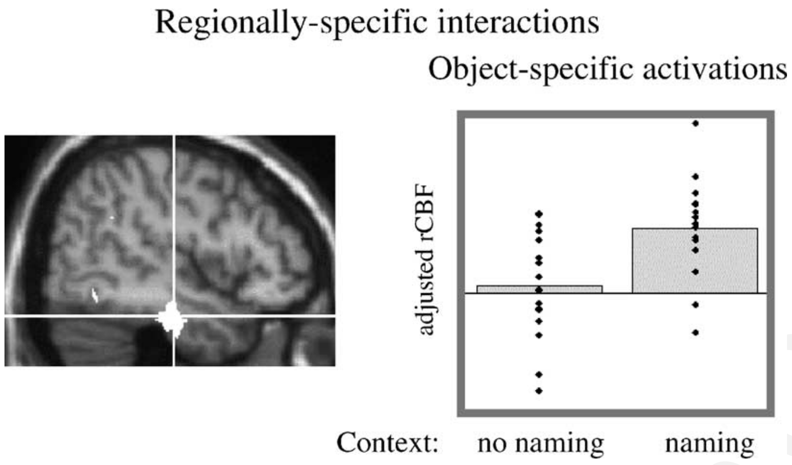

Fig. 4. This example of regionally specific interactions comes from an experiment where subjects were asked to view coloured non-object shapes or coloured objects and say "yes", or to name either the coloured object or the colour of the shape. *Left*: A regionally specific interaction in the left infero-temporal cortex. The SPM threshold is $P < 0.05$ (uncorrected) (Friston et al., 1995b). *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 6 subjects scanned 12 times using PET.

logical retrieval and object recognition. This analysis identifies not regionally specific activations but regionally specific *interactions*. When we actually performed this experiment these interactions were evident in the left posterior, inferior temporal region and can be associated with the integration of phonology and object recognition (see Fig. 4 and Friston et al., 1996 for details). Alternatively this region can be thought of as expressing recognition-dependent responses that are realised in, and only in, the context of having to name the object seen. These results can be construed as evidence of contextual specialisation for object-recognition that depends upon modulatory afferents (possibly from temporal and parietal regions) that are implicated in naming a visually perceived object. There is no empirical evidence in these results to suggest that the temporal or parietal regions are the source of this top–down influence but in the next example the source of modulation is addressed explicitly using psychophysiological interactions.

5.1.3. *Psychophysiological interactions*

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

The statistical model employed in testing for psychophysiological interactions is a simple regression model of effective connectivity that embodies nonlinear (second-order or modulatory effects). As such, this class of model speaks directly to functional specialisation of a nonlinear and contextual sort. Fig. 5 illustrates a specific example (see Dolan et al., 1997 for details). Subjects were asked to view (degraded) faces and non-face (object) controls. The interaction between activity in the parietal region and the presence of faces was expressed most significantly in the right infero-temporal region not far from the homologous left infero-temporal region implicated in the object naming experiment above. Changes in parietal activity were induced experimentally by pre-exposure of the (un-degraded) stimuli before some scans but not others to prime them. The data in the right panel of Fig. 5 suggests that the infero-temporal region shows face-specific responses, relative to non-face objects, when, and only when, parietal activity is high. These results can be interpreted as a priming-dependent face-specific response, in infero-temporal regions that are mediated by interactions with medial parietal cortex. This is a clear example of contextual specialisation that depends on top–down effects.

5.2. *Effective connectivity*

The previous examples demonstrating contextual specialisation are consistent with functional architectures implied by predictive coding. However, they do not provide defini-

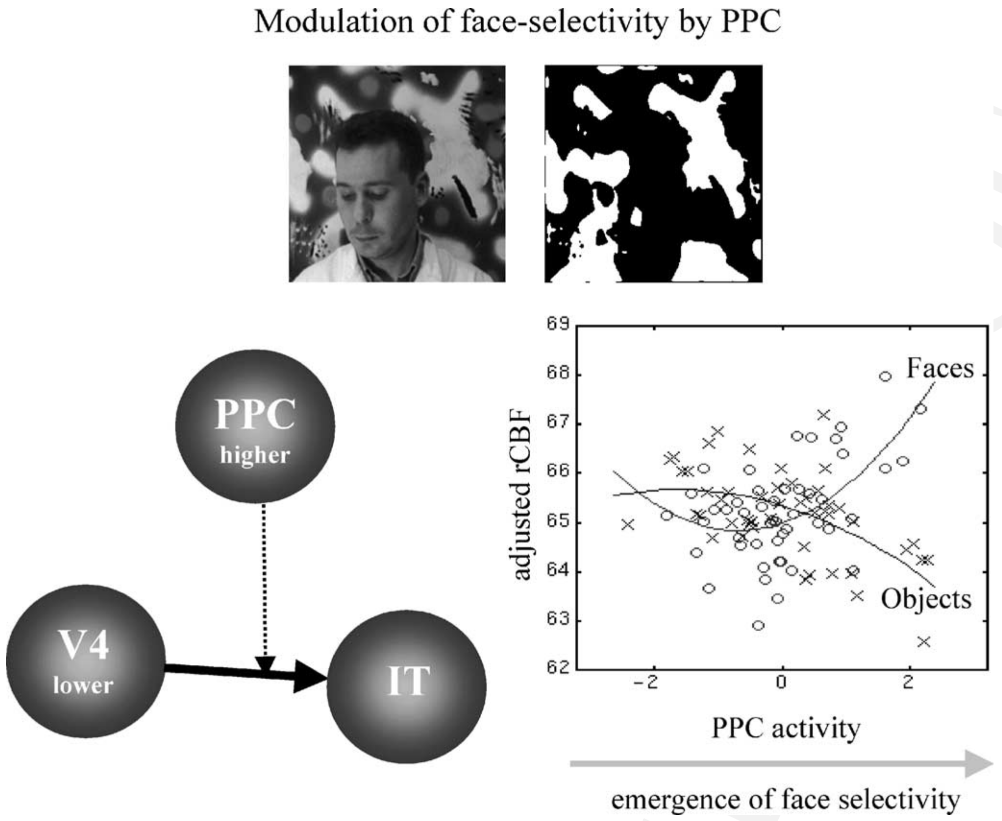

Fig. 5. *Top*: Examples of the stimuli presented to subjects. During the measurement of brain responses only degraded stimuli where shown (e.g. the right hand picture). In half the scans the subject was given the underlying cause of these stimuli, through presentation of the original picture (e.g. left) before scanning. This priming induced a profound difference in perceptual set for the primed, relative to non-primed, stimuli, *Right*: Activity observed in a right infero-temporal region, as a function of (mean corrected) PPC activity. This region showed the most significant interaction between the presence of faces in visually presented stimuli and activity in a reference location in the posterior medial parietal cortex (PPC). This analysis can be thought of as finding those areas that are subject to top–down modulation of face-specific responses by medial parietal activity. The crosses correspond to activity whilst viewing non-face stimuli and the circles to faces. The essence of this effect can be seen by noting that this region differentiates between faces and non-faces when, and only when, medial parietal activity is high. The lines correspond to the best second-order polynomial fit. These data were acquired from six subjects using PET. *Left*: Schematic depicting the underlying conceptual model in which driving afferents from ventral form areas (here designated as V4) excite infero-temporal (IT) responses, subject to permissive modulation by PPC projections.

tive evidence for an interaction between top–down and bottom–up influences. In this subsection we look for direct evidence of these interactions using functional imaging. This rests upon being able to measure effective connectivity in a way that is sensitive to interactions among inputs. This requires a plausible model of coupling among brain regions that accommodates nonlinear and dynamical effects. We have used a model that is based on the Volterra expansion introduced in Section 3. Before turning to empirical evidence for interactions between bottom–up and top–down inputs the motivation for this particular model of effective connectivity is presented briefly.

5.2.1. *Effective connectivity and Volterra kernels*

The problem faced, when trying to measure effective connectivity, is that measurements of brain responses are usually very limited, either in terms of their resolution (in space or time) or in terms of the neurophysiological or biophysical variable that is measured. Given the complicated nature of neuronal interactions, involving a huge number of microscopic variables, it may seem an impossible task to make meaningful measurements of coupling among brain systems, especially with measurements afforded by techniques like fMRI. However, the problem is not as intractable as one might think.

Suppose that the variables x represented a complete and self-consistent description of the state variables of a brain region. In other words, everything needed to determine the evolution of that region's state, at a particular place and time, was embodied in these measurements. If such a set of variables existed they would satisfy some immensely complicated nonlinear equations (cf. Eq. (1))

$$\begin{aligned}\dot{x} &= f(s,u)\\ y &= g(x)\end{aligned} \tag{33}$$

u represents a set of inputs conveyed by projections from other regions and x is a large vector of state variables which range from depolarisation at every point in the dendritic tree to the phosphorylation status of every relevant enzyme; from the biochemical status of every glial cell compartment to every aspect of gene expression. The vast majority of these variables are hidden and not measurable directly. However, there are measurements y that can be made, that, as we have seen in Section 3, are simply a nonlinear convolution of the inputs with some Volterra kernels. These measures usually reflect the activity of whole cells or populations and are measured in many ways, for example firing at the initial segment of an axon or local field potentials. The critical thing here is that the output is casually related to the inputs, *which are the outputs of other regions*. This means that that we never need to know the underlying and 'hidden' variables that describe the details of each region's electrochemical status. We only need to know the history of its inputs, which obtain from the measurable outputs of other regions. In principle, a complete description of regional responses could be framed in terms of inputs and the Volterra kernels required to produce the outputs. The nice thing about the kernels is that they can be interpreted directly as effective connectivity (see Box 1).

Because the inputs (and outputs) are measurable one can estimate the kernels empirically. The first-order kernel is simply the change in response induced by a change in input in the recent past. The second-order kernels are the change in the first–order effective connectivity induced by changes in a second (modulatory) input and so on for higher orders.

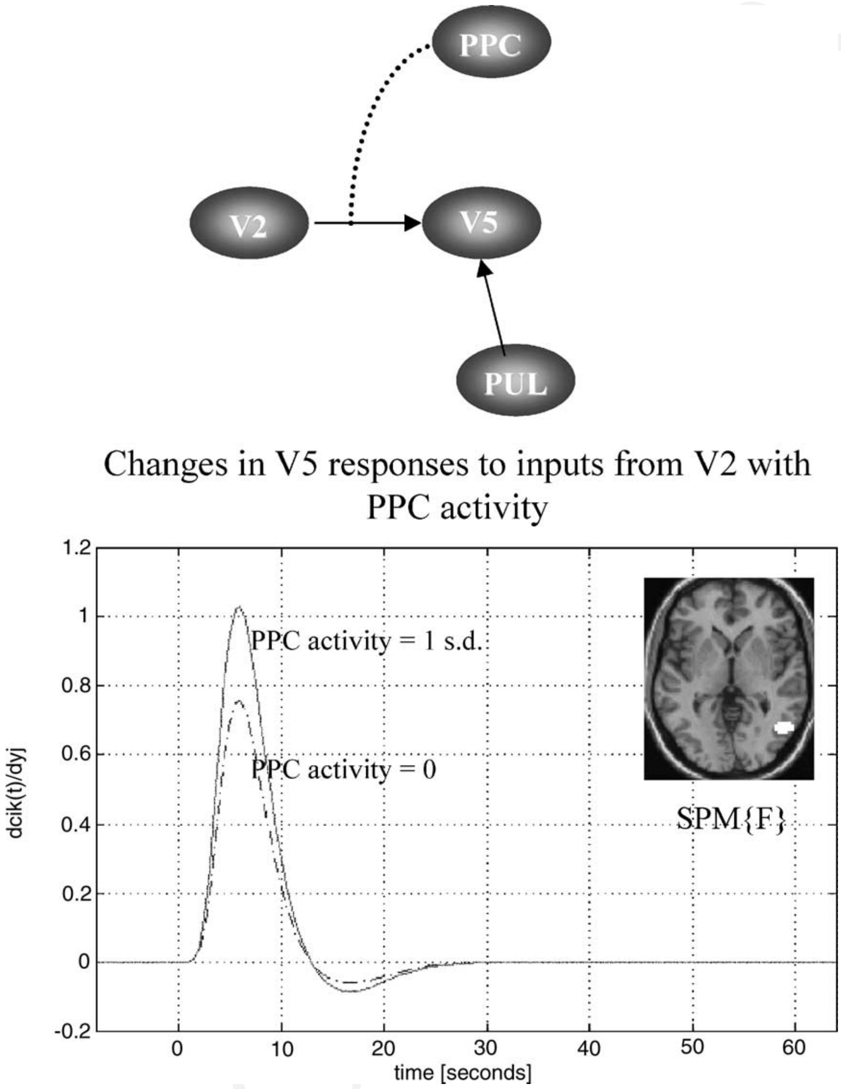

Fig. 6. *Left*: Brain regions and connections comprising the model. *Right*: Characterisation of the effects of V2 inputs on V5 and their modulation by posterior parietal cortex (PPC). The broken lines represent estimates of V5 responses when PPC activity is zero, according to a second-order Volterra model of effective connectivity with inputs to V5 from V2, PPC and the pulvinar (PUL). The solid curves represent the same response when PPC activity is one standard deviation of its variation over conditions. It is evident that V2 has an activating effect on V5 and that PPC increases the responsiveness of V5 to these inputs. The insert shows all the voxels in V5 that evidenced a modulatory effect ($P < 0.05$ uncorrected). These voxels were identified by thresholding a SPM (Friston et al., 1995b) of the F statistic testing for the contribution of second-order kernels involving V2 and PPC (treating all other terms as nuisance variables). The data were obtained with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes).

Another nice thing about the Volterra formulation is that the response is linear in the unknowns, which can be estimated using standard least square procedures. In short, Volterra kernels are synonymous with effective connectivity because they characterise the measurable effect that an input has on its target.

5.2.2. *Nonlinear coupling among brain areas*

Linear models of effective connectivity assume that the multiple inputs to a brain region are linearly separable. This assumption precludes activity-dependent connections that are expressed in one context and not in another. The resolution of this problem lies in adopting nonlinear models like the Volterra formulation that include interactions among inputs. These interactions can be construed as a context- or activity-dependent modulation of the influence that one region exerts over another (Büchel and Friston, 1997). In the Volterra model, second-order kernels model modulatory effects. Within these models the influence of one region on another has two components: (i) the direct or *driving* influence of input from the first (e.g. hierarchically lower) region, irrespective of the activities elsewhere; and (ii) an activity-dependent, *modulatory* component that represents an interaction with inputs from the remaining (e.g. hierarchically higher) regions. These are mediated by the first and second-order kernels, respectively. The example provided in Fig. 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 right panel in Fig. 6 shows a characterisation of this modulatory effect in terms of the increase in V5 responses, to a simulated V2 input, when posterior parietal activity is zero (broken line) and when it is high (solid lines). In this study subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes). The brain regions and connections comprising the model are shown in the upper panel. The lower panel shows a characterisation of the effects of V2 inputs on V5 and their modulation by posterior parietal cortex (PPC) using simulated inputs at different levels of PPC activity. It is evident that V2 has an activating effect on V5 and that PPC increases the responsiveness of V5 to these inputs. The insert shows all the voxels in V5 that evidenced a modulatory effect ($P < 0.05$ uncorrected). These voxels were identified by thresholding statistical parametric maps of the F statistic (Friston et al., 1995b) testing for the contribution of second-order kernels involving V2 and PPC while treating all other components as nuisance variables. The estimation of the Volterra kernels and statistical inference procedure is described in Friston and Büchel (2000).

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

6. Functional integration and neuropsychology

If functional specialisation depends on interactions among cortical areas then one might predict changes in functional specificity in cortical regions that receive enabling or modulatory afferents from a damaged area. A simple consequence is that aberrant responses will be elicited in regions hierarchically below the lesion if, and only if, these responses depend upon inputs from the lesion site. However, there may be other contexts in which the region's responses are perfectly normal (relying on other, intact, afferents). This leads to the notion of a context-dependent regionally-specific abnormality, caused by, but remote from, a lesion (i.e. an abnormal response that is elicited by some tasks but not others). We have referred to this phenomenon as 'dynamic diaschisis' (Price et al., 2000).

6.1. *Dynamic diaschisis*

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

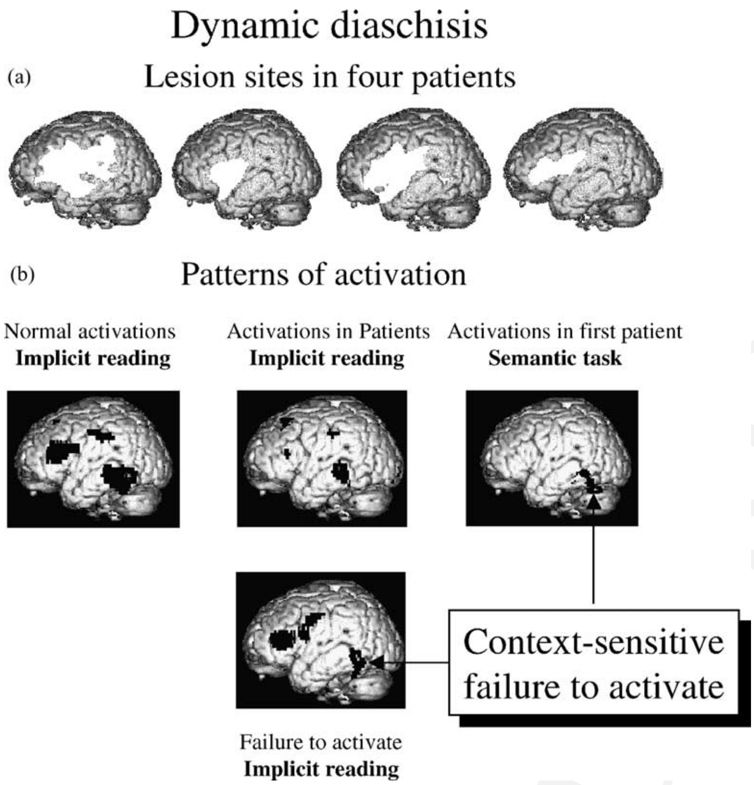

Fig. 7. (a) *Top*: These renderings illustrate the extent of cerebral infarcts in four patients, as identified by voxel-based morphometry. Regions of reduced grey matter (relative to neurologically normal controls) are shown in white on the left hemisphere. The SPMs (Friston et al., 1995b) were thresholded at $P < 0.001$ uncorrected. All patients had damage to Broca's area. The first (upper left) patient's left middle cerebral artery infarct was most extensive encompassing temporal and parietal regions as well as frontal and motor cortex. (b) *Bottom*: SPMs illustrating the functional imaging results with regions of significant activation shown in black on the left hemisphere. Results are shown for: (i) normal subjects reading words (left); (ii) activations common to normal subjects and patients reading words using a conjunction analysis (middle-top); (iii) areas where normal subjects activate significantly more than patients reading words, using the group times condition interaction (middle lower); and (iv) the first patient activating normally for a semantic task. Context-sensitive failures to activate are implied by the abnormal activations in the first patient, for the implicit reading task, despite a normal activation during a semantic task.

6.1.1. An empirical demonstration

We investigated this possibility in a functional imaging study of four aphasic patients, all with damage to the left posterior inferior frontal cortex, classically known as Broca's area (see Fig. 7, upper panels). These patients had speech output deficits but relatively preserved comprehension. Generally functional imaging studies can only make inferences about abnormal neuronal responses when changes in cognitive strategy can be excluded. We ensured this by engaging the patients in an explicit task that they were able to perform normally. This involved a keypress response when a visually presented letter string contained a letter with an ascending visual feature (e.g.: h, k, l, or t). While the task remained constant, the stimuli presented were either words or consonant letter strings. Activations detected for words, relative to letters, were attributed to implicit word processing. Each patient showed normal activation of the left posterior middle temporal cortex that has been associated with semantic processing (Price, 1998). However, none of the patients activated the left posterior inferior frontal cortex (damaged by the stroke), or the left posterior inferior temporal region (undamaged by the stroke) (see Fig. 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 while all four patients showed the reverse effect. The abnormal responses in the left posterior inferior temporal lobe occurred even though this undamaged region: (i) lies adjacent and posterior to a region of the left middle temporal cortex that activated normally (see middle column of Fig. 7b); and (ii) is thought to be involved in an earlier stage of word processing than the damaged left inferior frontal cortex (i.e. is hierarchically lower than the lesion). From these results we can conclude that, during the reading task, responses in the left basal temporal language area rely on afferent inputs from the left posterior inferior frontal cortex. When the first patient was scanned again, during an explicit semantic task, the left posterior inferior temporal lobe responded normally. The abnormal implicit reading related responses were therefore task-specific.

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

7. Conclusion

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

The main point made in this review is that backward connections, mediating internal or generative models of how sensory inputs are caused, are essential if the processes generating inputs are non-invertible. Because these generating processes are dynamical in nature, sensory input corresponds to a non-invertible nonlinear convolution of causes. This non-invertibility demands an explicit parameterisation of generative models (backward connections) to enable approximate recognition and suggests that feedforward architectures, are not sufficient for representational learning. Moreover, nonlinearities in generative models, that induce dependence on backward connections, require these connections to be modulatory; so that estimated causes in higher cortical levels can interact to predict responses in lower levels. This is important in relation to asymmetries in forward and backward connections that have been characterised empirically.

The arguments in this article were developed under prediction models of brain function, where higher-level systems provide a prediction of the inputs to lower-level regions. Conflict between the two is resolved by changes in the higher-level representations, which are driven by the ensuing error in lower regions, until the mismatch is ‘cancelled’. From this perspective the specialisation of any region is determined both by bottom–up driving inputs and by top–down predictions. Specialisation is therefore not an intrinsic property of any region but depends on both forward and backward connections with other areas. Because the latter have access to the context in which the inputs are generated they are in a position to modulate the selectivity or specialisation of lower areas. The implications for classical models (e.g. classical receptive fields in electrophysiology, classical specialisation in neuroimaging and connectionism in cognitive models) are severe and suggest these models may provide incomplete accounts of real brain architectures. On the other hand, predictive coding in the context of hierarchical generative models not only accounts for many extra-classical phenomena seen empirically but also enforces a view of the brain as an inferential machine through its empirical Bayesian motivation.

Uncited references

Friston (1995a), Friston (2000), Fukushima (1986), Harth et al. (1987), McIntosh and Gonzalez-Lima (1994).

Acknowledgements

The Wellcome Trust funded this work. I would like to thank my colleagues for help in writing this review and developing these ideas, especially Cathy Price for the psychological components and Peter Dayan for the theoretical neurobiology.

References

Abbot, L.F., Varela, J.A., Nelson, S.B., 1997. Synaptic depression and cortical gain control. Science 275, 220–223.

Absher, J.R., Benson, D.F., 1993. Disconnection syndromes: an overview of Geschwind’s contributions. Neurology 43, 862–867.

Aertsen, A., Preißl, H., 1991. Dynamics of activity and connectivity in physiological neuronal Networks. In: Schuster, H.G. (Ed.), Nonlinear Dynamics and Neuronal Networks. VCH publishers, New York, NY, USA, pp. 281–302.

Atick, J.J., Redlich, A.N., 1990. Towards a theory of early visual processing. Neural Comput. 2, 308–320.

Ballard, D.H., Hinton, G.E., Sejnowski, T.J., 1983. Parallel visual computation. Nature 306, 21–26.

Barlow, H.B., 1961. Possible principles underlying the transformation of sensory messages. In: Rosenblith, W.A. (Ed.), Sensory Communication. MIT Press, Cambridge, MA.

Bell, A.J., Sejnowski, T.J., 1995. An information maximisation approach to blind separation and blind de-convolution. Neural Comput. 7, 1129–1159.

Bendat, J.S., 1990. Nonlinear System Analysis and Identification from Random Data. Wiley, New York, USA.

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

Common, P., 1994. Independent component analysis, a new concept? Signal Processing 36, 287–314.

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

Dayan, P., Hinton, G.E., Neal, R.M., 1995. The Helmholtz machine. Neural Comput. 7, 889–904.

Dempster, A.P., Laird, N.M., Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

Devlin, J.T., Gunnerman, L.M., Andersen, E.S., Seidenberg, M.S., 1998. Category-specific semantic deficits in focal and widespread brain damage: a computational account. J. Cog. Neurosci. 10, 77–84.

Dolan, R.J., Fink, G.R., Rolls, E., Booth, M., Holmes, A., Frackowiak, R.S.J., Friston, K.J., 1997. How the brain learns to see objects and faces in an impoverished context. Nature 389, 596–598.

Dong, D., McAvoy, T.J., 1996. Nonlinear principal component analysis—based on principal curves and neural networks. Comput. Chem. Eng. 20, 65–78.

Edelman, G.M., 1993. Neural Darwinism: selection and reentrant signalling in higher brain function. Neuron 10, 115–125.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. J. Am. Stat. Assoc. 68, 117–130.

Farah, M., McClelland, J., 1991. A computational model of semantic memory impairments: modality specificity and emergent category specificity. J. Exp. Psychol. Gen. 120, 339–357.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Fletcher, P.C., Frith, C.D., Grasby, P.M., Shallice, T., Frackowiak, R.S.J., Dolan, R.J., 1995. Brain systems for encoding and retrieval of auditory–verbal memory. Brain 118, 401–416.

Fliess, M., Lamnabhi, M., Lamnabhi-Lagarrigue, F., 1983. An algebraic approach to nonlinear functional expansions. IEEE Trans. Circuits Syst. 30, 554–570.

Foldiak, P., 1990. Forming sparse representations by local anti-Hebbian learning. Biol. Cybern. 64, 165–170.

Freeman, W., Barrie, J., 1994. Chaotic oscillations and the genesis of meaning in cerebral cortex. In: Buzsaki, Llinas, R., Singer, W., Berthoz, A., Christen, T. (Eds.), Temporal Coding in the Brain. Springer, Berlin, pp. 13–38.

Friston, K.J., 2000. The labile brain. III. Transients and spatio-temporal receptive fields. Phil. Trans. R. Soc. Lond. B 355, 253–265.

Friston, K.J., Büchel, C., 2000. Attentional modulation of V5 in human. Proc. Natl. Acad. Sci. U.S.A. 97, 7591–7596.

Friston, K.J., Frith, C., Passingham, R.E., Dolan, R., Liddle, P., Frackowiak, R.S.J., 1992. Entropy and cortical activity: information theory and PET findings. Cereb. Cortex 3, 259–267.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J., 1993. Principal component analysis learning algorithms: a neurobiological analysis. Proc. R. Soc. B 254, 47–54.

Friston, K.J., Tononi, G., Reeke, G.H., Sporns, O., Edelman, G.E., 1994. Value-dependent selection in the brain: simulation in synthetic neural model. Neuroscience 39, 229–243.

Friston, K.J., 1995a. Functional and effective connectivity in neuroimaging: a synthesis. Hum. Brain Mapp. 2, 56–78.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J., 1995b. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Friston, K.J., Price, C.J., Fletcher, P., Moore, C., Frackowiak, R.S.J., Dolan, R.J., 1996. The trouble with cognitive subtraction. NeuroImage 4, 97–104.

Friston, K.J., Büchel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6, 218–229.

Gawne, T.J., Richmond, B.J., 1993. How independent are the messages carried by adjacent inferior temporal cortical neurons? J. Neurosci. 13, 2758–2771.

Fukushima, K., 1986. A neural network model for selective attention in visual pattern recognition. Biol. Cybern. 55, 5–15.

Gerstein, G.L., Perkel, D.H., 1969. Simultaneously recorded trains of action potentials: analysis and functional interpretation. Science 164, 828–830.

Girard, P., Bullier, J., 1989. Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. J. Neurophysiol. 62, 1287–1301.

Harth, E., Unnikrishnan, K.P., Pandya, A.S., 1987. The inversion of sensory processing by feedback pathways: a model of visual cognitive functions. Science 237, 184–187.

Hinton, G.T., Shallice, T., 1991. Lesioning an attractor network: investigations of acquired dyslexia. Psychol. Rev. 98, 74–95.

Hinton, G.E., Dayan, P., Frey, B.J., Neal, R.M., 1995. The wake–sleep algorithm for unsupervised neural networks. Science 268, 1158–1161.

Hirsch, J.A., Gilbert, C.D., 1991. Synaptic physiology of horizontal connections in the cat's visual cortex. J. Neurosci. 11, 1800–1809.

Karhunen, J., Joutsensalo, J., 1994. Representation and separation of signals using nonlinear PCA type learning. Neural Netw. 7, 113–127.

Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407, 717–726.

Kay, J., Phillips, W.A., 1996. Activation functions, computational goals and learning rules for local processors with contextual guidance. Neural Comput. 9, 895–910.

Kramer, M.A., 1991. Nonlinear principal component analysis using auto-associative neural networks. AIChE J. 37, 233–243.

Lenzi, G.L., Frackowiak, R.S.J., Jones, T., 1982. Cerebral oxygen metabolism and blood flow in human cerebral ischaemic infarction. J. Cereb. Blood Flow Metab. 2, 321–335.

Linsker, R., 1988. Self-organisation in a perceptual network. Computer 105–117.

Lueck, C.J., Zeki, S., Friston, K.J., Deiber, M.P., Cope, N.O., Cunningham, V.J., Lammertsma, A.A., Kennard, C., Frackowiak, R.S.J., 1989. The colour centre in the cerebral cortex of man. Nature 340, 386–389.

Kawato, M., Hayakawa, H., Inui, T., 1993. A forward-inverse optics model of reciprocal connections between visual cortical areas. Network 4, 415–422.

McIntosh, A.R., 2000. Towards a network theory of cognition. Neural Netw. 13, 861–870.

McIntosh, A.R., Gonzalez-Lima, F., 1994. Structural equation modelling and its application to network analysis in functional brain imaging. Hum. Brain Mapp. 2, 2–22.

Mumford, D., 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern 66, 241–251.

Nebes, R., 1989. Semantic memory in Alzheimer's disease. Psychol. Bull. 106, 377–394.

Oja, E., 1989. Neural networks, principal components, and subspaces. Int. J. Neural Syst. 1, 61–68.

Olshausen, B.A., Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609.

Optican, L., Richmond, B.J., 1987. Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II. Information theoretic analysis. J. Neurophysiol. 57, 132–146.

Pack, C.C., Born, R.T., 2001. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. Nature 409, 1040–1042.

Pearl, J., 2000. Causality, Models, Reasoning and Inference. Cambridge University Press, Cambridge, UK.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1989. Positron emission tomographic studies of the processing of single words. J. Cog. Neurosci. 1, 153–170.

Phillips, W.A., Singer, W., 1997. In search of common foundations for cortical computation. Behav. Brain Sci. 20, 57–83.

Phillips, C.G., Zeki, S., Barlow, H.B., 1984. Localisation of function in the cerebral cortex: past present and future. Brain 107, 327–361.

Plaut, D., Shallice, T., 1993. Deep dyslexia—a case study of connectionist neuropsychology. Cog. Neuropsychol. 10, 377–500.

Poggio, T., Torre, V., Koch, C., 1985. Computational vision and regularisation theory. Nature 317, 314–319.

Price, C.J., 1998. The functional anatomy of word comprehension and production. Trends Cog. Sci. 2, 281–288.

Price, C.J., Warburton, E.A., Moore, C.J., Frackowiak, R.S.J., Friston, K.J., 2000. Dynamic diaschisis: anatomically remote and context-specific human brain lesions. J. Cog. Neurosci. 00, 00–00.

Rao, R.P., 1999. An optimal estimation approach to visual perception and learning. Vision Res. 39, 1963–1989.

Rao, R.P., Ballard, D.H., 1998. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. Nat. Neurosci. 2, 79–87.

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

Rumelhart, D., McClelland, J., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.

Salin, P.-A., Bullier, J., 1995. Corticocortical connections in the visual system: structure and function. Psychol. Bull. 75, 107–154.

Sandell, J.H., Schiller, P.H., 1982. Effect of cooling area 18 on striate cortex cells in the squirrel monkey. J. Neurophysiol. 48, 38–48.

Sugase, Y., Yamane, S., Ueno, S., Kawano, K., 1999. Global and fine information coded by single neurons in the temporal visual cortex. Nature 400, 869–873.

Sutton, R.S., Barto, A.G., 1990. Time derivative models of Pavlovian reinforcement. In: Gabriel, M., Moore, J. (Eds.), Learning and Computational Neuroscience: Foundations of Adaptive Networks. MIT Press, Cambridge, MA, pp. 497–538.

Taleb, A., Jutten, C., 1997. Nonlinear source separation: the post-nonlinear mixtures. In: Proceedings of the ESANN'97, Bruges, April 1997, pp. 279–284.

Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. U.S.A. 91, 5033–5037.

Tovee, M.J., Rolls, E.T., Treves, A., Bellis, R.P., 1993. Information encoding and the response of single neurons in the primate temporal visual cortex. J. Neurophysiol. 70, 640–654.

Treue, S., Maunsell, H.R., 1996. Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382, 539–541.

Warrington, E.K., McCarthy, R., 1983. Category specific access dysphasia. Brain 106, 859–878.

Warrington, E.K., McCarthy, R., 1987. Categories of knowledge: further fractionations and an attempted integration. Brain 110, 1273–1296.

Warrington, E.K., Shallice, T., 1984. Category specific semantic impairments. Brain 107, 829–853.

Wray, J., Green, G.G.R., 1994. Calculation of the Volterra kernels of nonlinear dynamic systems using an artificial neuronal network. Biol. Cybern. 71, 187–195.

Zeki, S., 1990. The motion pathways of the visual cortex. In: Blakemore, C. (Ed.), Vision: Coding and Efficiency. Cambridge University Press, Cambridge, UK, pp. 321–345.

Zeki, S., Shipp, S., 1988. The functional logic of cortical connections. Nature 335, 311–317.