Brain Research Bulletin, Vol. 54, No. 3, pp. 275–285, 2001

0361-9230/01/$–see front matter

PII S0361-9230(00)00436-6

Dynamic representations and generative models of brain function

Karl J. Friston and Cathy J. Price*

The Wellcome Department of Cognitive Neurology, Institute of Neurology, London, UK

[Accepted 26 October 2000]

ABSTRACT: The main point made in this article is that the representational capacity and inherent function of any neuron, neuronal population or cortical area is dynamic and context-sensitive. This adaptive and contextual specialisation is mediated by functional integration or interactions among brain systems with a special emphasis on backwards or top-down connections. The critical notion is that neuronal responses, in any given cortical area, can represent different things at different times. Our argument is developed under the perspective of generative models of functional brain architectures, 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, driven by the resulting error in lower regions, until the mismatch is 'cancelled'. In this model 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 these other areas 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 provide incomplete accounts of real brain architectures. Generative models represent a far more plausible framework for understanding selective neurophysiological responses and how representations are constructed in the brain.

KEY WORDS: Representations, Predictive coding, Effective connectivity.

INTRODUCTION

In this article we have chosen to address the dynamic aspects of brain function in terms of representations and how they can change dynamically and in a context-sensitive fashion. With the growing interest in extra-classical receptive field effects (i.e., how the receptive fields of early sensory units 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, showing a short-term plasticity and context-sensitivity. This suggests that a cortical area or neuronal population may be specialised for one thing in one context but something else in another. This dynamical aspect of functional brain architectures depends on an interplay between functional specialisation and integration. An interplay which neuroimaging is now starting to characterise.

The article starts by reviewing the two fundamental principles of brain organisation, namely functional specialisation and functional integration and how they relate to each other. The second section discusses how functional specialisation depends on integration and interactions among neuronal populations. This discussion is motivated by basic neuroscience findings and theoretical accounts of neuronal computation based on generative models. These models emphasise the role of backwards connections and prediction in perceptual categorisation and allow for the specialisation of any cortical area to be dynamically reconfigured in a way that depends on the prevailing context. Empirical evidence from functional neuroimaging studies of human subjects is presented in the third section to illustrate the context-sensitive nature of functional specialisation and how its expression depends upon functional integration among remote cortical areas. The final section introduces 'dynamic diaschisis', in which aberrant neuronal responses can be observed as a consequence of damage to distal areas that provide enabling or modulatory afferents. This section uses neuroimaging in neuropsychological patients and discusses the implications for constructs based on classical notions, like the lesion-deficit model.

FUNCTIONAL SPECIALISATION AND INTEGRATION

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 19th century. Since the early anatomic theories of Gall, the identification of a particular brain region with a specific function has become a central theme in neuroscience. However functional localisation *per se* was not easy to demonstrate: For example, a meeting that took place on August 4, 1881 addressed the difficulties of attributing function to a cortical area, given the dependence of cerebral activity on underlying connections [27]. This meeting was entitled

* Address for correspondence: Cathy J. Price, The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London, WC1N 3BG UK. Fax: +44-207-813-1445; E-mail: c.price@fil.ion.ucl.ac.uk

"Localisation of function in the cortex cerebri". Goltz [17] 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 [2,23]) 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 allows for the integration of several cortical areas in the processing of one particular function. Adhering to the principal of functional specialisation does not necessarily imply that any function, however atomic or elemental, can be localised in a single area. The cortical infrastructure supporting a single function may 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.

Functional Specialisation and Segregation

The functional role played by any component (e.g., cortical area, subarea, neuronal population or neuron) of the brain is largely defined 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" [41]. 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., backwards) 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 one or more cortical areas 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 and challenging a subject with the appropriate sensorimotor attribute or cognitive process should lead to activity changes in these areas. This is the model upon which the search for regionally-specific effects with functional neuroimaging is based.

The Anatomy and Physiology of Cortico-Cortical Connections

If specialisation rests upon connectivity then important principles underpinning specialisation 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 extrinsic 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 'forwards' and 'backwards' connections. The anatomy and physiology of these connections suggest that forwards connections are driving and commit cells to a pre-specified response given the appropriate pattern of inputs. Backwards connections, on the other hand, are less topographically constrained and are in a position to modulate the responses of lower areas to driving inputs from either higher or lower areas. Some of this evidence is presented below. The list is not exhaustive, nor properly qualified, but serves to introduce some of the more important principles that have emerged from empirical studies of the visual cortex.

Hierarchial 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 [8]. 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. Backwards 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 [33,34].

Forward connections are driving. Backward connections are modulatory. Reversible inactivation (e.g., [16,35]) and functional neuroimaging (e.g., [4,11]) studies suggest that forward connections are driving whereas backward connections are more modulatory. The notion that forward connections are concerned with the promulgation and segregation of sensory information is consistent with (1) their sparse axonal bifurcation; (2) patch axonal terminations and (3) topographic projections. In contradistinction modulatory, backward connections are generally considered to have a role in mediating contextual effects and in the coordination of processing channels. This is consistent with (1) their frequent bifurcation; (2) diffuse axonal terminations and (3) non-topographically constrained patterns of projections [5,34].

Modulatory connections have slow time constants. Forward connections meditate their postsynaptic effects through fast AMPA (1.3–2.4-ms decay) and $GABA_A$ (6-ms decay) receptors. Modulatory afferents activate N-methyl-D-aspartate (NMDA) receptors. NMDA receptors are voltage-sensitive showing non-linear and slow dynamics (50-ms decay). They are found predominantly in supragranular layers where backward connections terminate [34]. These slow time-constants again point to a role in mediating contextual effects that are more enduring than sensory-evoked responses of a phasic nature.

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 patch in another area that has a diameter of approximately 5–8 mm. An important distinction between forward and backward connections is that backward connections are more divergent and transcend more levels. For example, the divergence region of a point in V5 (i.e., the region receiving backwards 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 [40]. An example of backward connections traversing hierarchical levels are those that connect TE and TEO to V1 although there are no mono-synaptic connections from V1 to TE or TEO [34]. Thus, reciprocal interaction between two levels, in conjunction with the divergence of backwards 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. Forward connections by contrast are more restricted and less numerous. For example, the ratio of forward efferent connections to backwards afferents in the lateral geniculate is about 1:10/20.

In short, backwards connections are abundant and are in a position to exert powerful effects on evoked responses in lower levels where these responses define the specialisation of any area or neuronal population. The idea promoted in this article 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 the role of functional integration in dynamically reshaping the specialisation of brain areas that mediate perceptual synthesis and adaptive behavioral responses.

The aspects of connectivity above constrain the infrastructure of neuronal architectures. However, they do not provide for any direct way of characterising the influences that one neuron, or population, exerts over another. These influences are assessed in terms of neurophysiological measurements using the concept of effective connectivity.

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. From the perspective of neuroimaging, functional specialisation calls for the identification of regionally specific effects that can be attributed to changing stimuli or task conditions. Functional integration, on the other hand, 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 [10]. 'Functional connectivity' is defined as correlations between remote neuro-physiological 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 [15]. Integration within a distributed system is usually better understood in terms of effective connectivity. Effective connectivity refers explicitly to the influence that one neural 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" [3]. This speaks to two important points: (1) Effective connectivity is dynamic, i.e., activity-and time-dependent and (2) it depends upon a model of the interactions. The models employed in functional neuroimaging can be classified as those based on regression models [11,12] or structural equation modelling [25]. A more important distinction is whether these models are linear or non-linear. Recent characterisations of effective connectivity, in neuroimaging, have focussed on non-linear models that accommodate the modulatory effects described above.

Non-linear Coupling Among Brain Areas

Linear models of effective connectivity assume that the multiple inputs to a region are linearly separable. This assumption precludes activity-dependent connections that are expressed in one context and not in another. The resolution of this problem lies in adopting non-linear models 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, where that context is instantiated by activity in further brain regions exerting modulatory effects. These non-linearities can be introduced into structural equation modelling using, so-called 'moderator' variables that represent the interaction between two regions when causing activity in a third [4]. From the point of view of regression models modulatory effects can be modelled with non-linear input-output models (e.g., a Volterra series formulation). Within these models the influence of one region on another has two components; (1) the direct or driving influence of input from the first (e.g., lower) region, irrespective of the activities elsewhere and (2) an activity-dependent, modulatory component that represents an interaction with inputs from the remaining (e.g., higher) regions. The example provided in Fig. 1 addresses the modulation of visual cortical responses by attentional mechanisms (e.g., [38]) and the mediating role of activity-dependent changes in effective connectivity. Figure 1 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). This is a nice example of how a higher level region (the parietal area) is modulating responses in a lower level area (V5). The result suggests that backwards parietal inputs may be a sufficient explanation for attentional modulation of visually evoked extrastriate responses (see Fig. 1 legend for more details).

In summary the brain can be considered as an ensemble of functionally specialised areas that are coupled in a non-linear fashion by effective connections. 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. Non-linear coupling means that the responses of any cortical region, to inputs from another, depends upon activity in all regions that provide [modulatory] afferents. These are generally higher-level regions. This dependency represents interactions among inputs that cause the response. These sorts of influences can now be measured with functional neuroimaging and there is a reasonable understanding of their physiological basis. In the next section we describe a theoretical perspective, provided by 'generative models', that highlights the functional importance of backwards connections and modulatory or non-linear interactions.

GENERATIVE MODELS

The relationship between functional and neuronal architectures is central to the cognitive neuroscience endeavour. This section addresses this relationship by considering generative models. In brief we will suggest that the role of backwards 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 causes changes in the higher area until there is a reconciliation. There is no more error, when, and only when, the bottom-up driving inputs to an area are in harmony with the top-down prediction and a consensus between the prediction and the actual input is established. In other words, when there is no more error, there is no further change to the higher order representation because the driving inputs are quiescent. Given this conceptual model a change in activity corresponds to some transient error signal that induces the appropriate change in higher areas until an appropriate higher-level representation emerges and the error is 'cancelled' by backwards connections. Clearly the prediction error will depend on the context and consequently the backwards con-

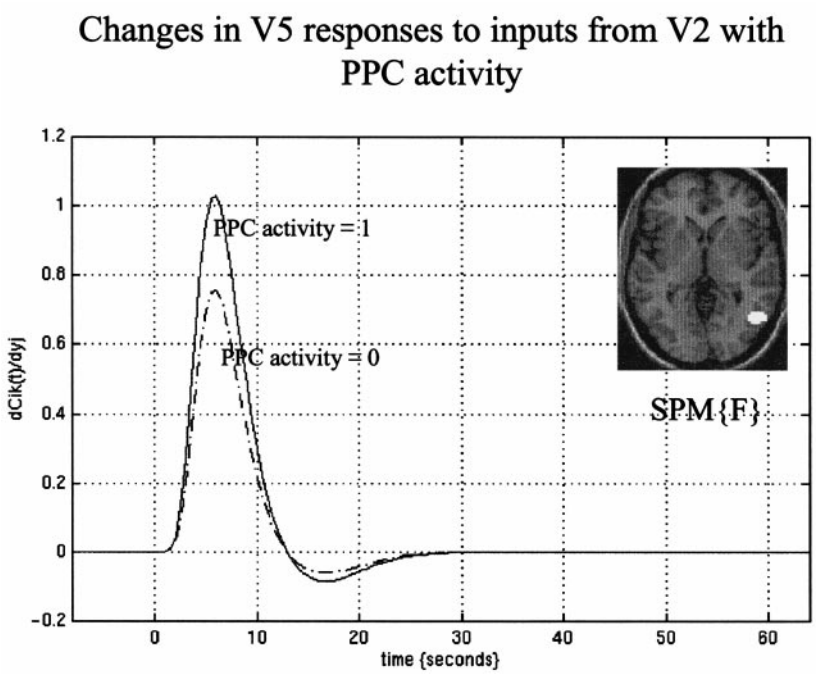

FIG. 1. Characterization of responses in V5 to inputs from V2 and their modulation by posterior parietal cortex (PPC) using simulated inputs at different levels of PPC activity. The broken lines represent estimates of V5 responses when PPC activity = 0 according to a second order Volterra model of effective connectivity with inputs to V5 based on the activity inV2, PPC, and the pulvinar. The simulated input, from V2, corresponded to a square wave of 500 ms duration convolved with a hemodynamic response function. The solid curves represent the same response when PPC activity is one. 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 [37] that evidenced a modulatory effect ($p < 0.05$ uncorrected). These voxels were identified by thresholding statistical parametric maps (SPMs; see [12]) of the F statistic testing for the contribution of second order kernels involving V2 and PPC while treating all other components as nuisance variables. Subjects were studied with functional magnetic resonance imaging under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes).

nections confer a context-sensitivity on the functional specificity of the lower area. In short the neuronal responses do not just depend on bottom-up input but on the difference between bottom-up input and top-down predictions.

This model may appear specious in that higher areas are simply trying to predict what lower areas already 'know'. However, from a bottom-up perspective the convergence and divergence of cortico-cortical connections mean that higher-order representations are dynamically assembled from many lower order representations to engender a perceptual synthesis. From a top-down perspective the prediction is much more informed, than the input being predicted, because convergence from multiple high-level representations provides constraints on what is currently being perceived (i.e., provides a prediction that is conditional on the wider context).

The prevalence of non-linear or modulatory bottom-down effects can be inferred from the fact that, almost by definition, context interacts with the content of any representation. Backwards connections from higher areas, that do not receive forward connections from the area in question, can be considered as providing contextual modulation of the prediction from higher areas that do. This is because these contextual inputs are not subject to control by the prediction error (there are no forward connections to implement this control) and are therefore unlikely to elicit a response by themselves.

In this section we review briefly generative models and representations in a way that is motivated by empirical and theoretical advances in basic neuroscience. This section concludes with an example of dynamic representation in infero-temporal cortex based on unit recordings in monkeys before looking at similar phenomena, in humans, in the next section.

Generative Models and Predictive Coding

Over the past years 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 generative models the dynamics of units in a network are trying to predict the inputs. The representational aspects of any unit emerges 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 sensory input is sufficient to predict it with minimal error. 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., [6,19]) to biologically plausible models of visual processing (e.g., [32]). The goal of generative models is "to learn representations that are economical to describe but allow the input to be reconstructed accurately" [19]. These models emphasise the role of backwards 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 the connection strengths change to minimise the error between the observed and *desired output*. In generative models there is no 'output' because the representational meaning of the units is not pre-specified but emerges during learning. The representation can therefore be described as labile and depends upon the context in which activity is evoked (e.g., the extra-classical receptive field effects modelled in [32]). The latter is important and results from the fact that the responses of low level units are a strong function of activity at higher levels.

In summary, previous (e.g., connectionist [18]) models have assumed the existence of fixed representations at a neuronal level. In contradistinction generative models offer an alternative approach that does not enforce a fixed relationship between the activity of any unit and what is being represented. In the next section we consider the nature of real neuronal representations and whether they are consistent with a generative perspective.

Neuronal Representations

Here a representation is taken to be a neuronal event that represents some 'cause' in the sensorium. It can be defined operationally as the neuronal responses evoked by the cause being represented. Using this definition, one practical way of getting at which representations a given unit (neuron) participates in can be based on those causes that elicit a response. Clearly, some causes will elicit a response and others will not, and this is the basis of "selectivity". Selective responses therefore define the unit's receptive field and indeed electrophysiologically, receptive fields are mapped using these selective responses. The selectivity of the receptive field (see below) depends upon the synaptic connection strengths of the inputs to any neuron. In other words, the receptive field is a function of the strength of the synaptic connections engendering the responses. In short, at some fundamental level, there is an intimate relationship between the selectivity of a neuron's responses, its receptive field and implicitly its specialisation and the representation the neuron participates in.

Classical models (e.g., classical receptive fields and connectionism) assume that evoked responses will be invariably expressed in the same units or neuronal populations irrespective of the context. The problem is that real neuronal representations are not invariant but depend upon the context in which responses are evoked: If the representation is a function of the strength of the synaptic connections mediating it and non-linear coupling among neuronal populations modulates the efficacy of these connections in an activity-dependent way, then the representation is itself activity-dependent and dynamic. Put simply, for a given sensory cause, the context can change which units represent that cause or the nature of that representation over a given set of units. For example, visual cortical units have dynamic receptive fields that can change from moment to moment (cf. the non-classical receptive field effects in generative models [32] or attentional modulation of evoked responses [38]). Given this, the activity evoked in any unit can be partitioned into two components; (1) a latent component that is insensitive to the context and does not depend upon non-linear interactions with other synaptic inputs and (2) a context-sensitive component that is a function of activity in modulatory presynaptic inputs. For a given cause this means that the evoked responses have two components, corresponding to 'latent' and 'contextual' representations.

The evidence for contextual representations 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 the change in connection strength or synaptic efficacy, either potentiation or depression, following pre-synaptic inputs (e.g., [1]). In brief, the underlying connection strengths, that define what that 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 [38] and human functional magnetic resonance imagine (fMRI) studies [4] that attention to specific visual attributes can profoundly alter the receptive fields or event-related responses to the same stimuli. In our own fMRI studies, these influences can be characterised in terms of a modulation of the connection strengths between early visual processing areas and the motion-sensitive area V5/MT by activity in posterior parietal regions (see Fig. 1). 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. As noted above, in section on "Functional Specialisation and Segregation", forward connections from sensory areas are generally considered to be driving, eliciting obligatory responses in the neurons that they target, whereas backwards connections are more modulatory in nature [5,16,33,35] interacting with the driving inputs to change their effective connection strengths (i.e., the representation of a cause). There is clear evidence that lateral connections in visual cortex are modulatory in nature [20], again speaking to an interaction between the functional segregation implicit in the column architecture of V1 and the neuronal dynamics in distal populations.

The picture that emerges from anatomical and electrophysiological studies of the brain is of a skeleton of reciprocal extrinsic connections, where these connections are driving (and excitatory). This skeleton is encompassed by lateral and backwards connections that can exert a modulatory influence on lower or equivalent stages of cortical transformations and define a hierarchy of cortical areas. The modulatory effects change the effective strength of driving connections and implicity change what each unit or population will respond to (i.e., what is represented). These modulatory effects may be expressed directly in terms of voltage-sensitive mechanisms (e.g., NMDA receptors) or may emerge through non-linear interactions involving intrinsic interneurons. Theoretical work, based on these observations, suggests that lateral and backwards interactions may convey contextual information that shapes the responses of any neuron to its inputs (e.g., [21,28]). One perspective, on this dynamic re-modelling of receptive fields by contextual input, is that it informs the extraction of causes in the external world, using conditional probabilities derived from higher levels of processing.

From the point of view of generative models, the way in which a particular sensory cause will present itself will depend on many other contextual cognitive and sensorial attributes. These attributes will be represented at higher levels and, will modulate the prediction, conveyed by backwards connections, in a way that conforms to its most probable expression. This is consistent with the more diffuse patterns of backward projection and their potential to exert modulatory effects.

An Example From Electrophysiology

In the next section we will illustrate the contextual nature of representations, and implicit specialisation, in the infero-temporal lobe using neuroimaging. Here we consider the evidence for contextual representations in terms of single cell responses, to visual stimuli, in the infero-temporal cortex of awake behaving monkeys. If specialisation for the high-order attributes of a stimulus are conferred by top-down influences then one might expect to see the emergence of selectivity, for these 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 backwards afferents can dynamically reshape the response profile and selectivity of lower areas. This temporal delay in the emergence of selectivity is precisely what one sees empirically. Indeed the late components of event-related potentials in the electroencephalogram are sometimes referred to as 'endogenous' to reflect the dependency on top-down processing. Here we focus on a more direct example: Sugase et al. [36] 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 neurons 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 may rely on backwards connections which imbue the neurons with a selectivity that is not intrinsic to the area but depends on social and other associational (cf. semantic in humans) processing at higher levels. The amygdala is involved in social behaviour and emotional learning and is interconnected with inferior temporal regions. As pointed out by the authors "amygdala neurons may interact with the face-response neurons in the inferior temporal cortex" [36].

These results present a difficulty for classical models: Do inferior temporal units represent faces generically or do they represent 'happy' faces? In one context, activity in these units reflects a representation of any face (before elaboration of processing in higher areas) and in another they represent a specific facial emotion (after their selectivity has been dynamically modulated by top-down influences). A generative perspective resolves this ambiguity.

Summary

By virtue of the non-linear and modulatory effect of backwards connections in the brain, that can dynamically change the response properties of neurons, neuronal representations become a function of, and dependent upon, input from distal cortical areas. The existence of long-range modulatory effects leads directly to the notion of two sorts of functional specialisation in the brain: (1) Latent specialisation that depends only on 'driving' connections and that is context-insensitive. (2) Contextual specialisation that is conferred by 'modulatory'interactions with other areas at the same level, or higher, in a cortical hierarchy. The latter is context-sensitive and explicitly dependant upon non-linear coupling among brain regions. In the next section we look at some empirical evidence from functional neuroimaging that confirms the idea that functional specialisation is both context-sensitive and depends on interactions with higher brain areas.

FUNCTIONAL SPECIALISATION AND BRAIN IMAGING

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 the activity in higher areas. The second example in this section shows that this is indeed possible. Both of the empirical examples given below depend on eliciting regionally specific responses, in different contexts, with factorial experimental designs.

Context-Sensitive Specialisation

Categorical designs, such as cognitive subtraction, have been the mainstay of functional neuroimaging over the past decade (e.g., [24,26]). Cognitive subtraction involves elaborating two tasks that differ in a separable component. Ensuing differences in brain activity are then attributed to this component. For example, consider the difference between simply saying "yes" when are cognisable 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 (e.g., 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 (e.g., the integration of object recognition and response). 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, which eschew the assumption of pure insertion.

Factorial designs involve combining two or more factors within a task or tasks. Consider repeating the above implicit object recognition experiment in another context, for example phonological retrieval (of the object's name or the non-object's colour). The factors in this example are implicit object recognition with two levels (objects *vs.* non-objects) and phonological retrieval (naming vs. saying "yes"). The idea here is to look at the interaction between these factors, or the effect that one factor has on the responses due to the other. Generally, interactions can be thought of as a difference in activations brought about by another processing demand. In other words, in changing the context of a particular task one can modulate the activation and examine the interaction between the activation and the context employed. Dual task interference paradigms are a clear example of this approach (e.g., [9]).

Consider the above object recognition experiment again. The factorial nature of this experiment can be seen by noting that object-specific responses are elicited (by asking subjects to view objects relative to meaningless shapes) with and without phonological retrieval. This 'two by two' design allows one to look

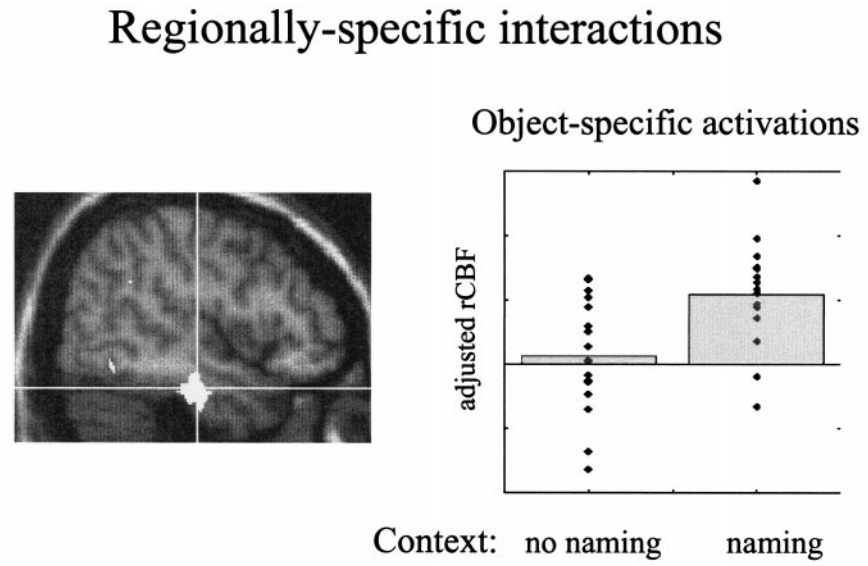

FIG. 2. This example of regionally specific interactions comes from an experiment where subjects were asked to either view coloured non-object shapes or coloured objects and say "yes", or to name either the coloured object or the colour of the shape. A regionally specific interaction in the left infero-temporal cortex is shown (left). The statistical parametric map threshold is $p < 0.05$ (uncorrected). The corresponding activities in the maxima of this region are portrayed in terms of object recognition-dependent responses with and without naming (right). It is seen that this region shows object recognition responses when and only when the phonology of that object has to be retrieved. The 'extra' activation with naming corresponds to the interaction. These data were acquired from six subjects scanned 12 times using positron emission tomography. Abbreviation: rCBF, regional cerebral blood flow.

specifically at the interaction between phonological retrieval and object recognition. This analysis identifies not regionally specific activations but regionally specific *interactions*. When we actually performed this experiment these interactions were evident in the left inferior temporal region and can be associated with the integration of phonology and object recognition (see Fig. 2 left panel and [13] 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 (see Fig. 2, right panel). In relation to the distinction between latent and contextual specialisation 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.

Psychophysiological Interactions

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 [14]. The supposition here is that a remote region is the source of backwards or lateral modulatory afferents that confer functional specificity on the index region. For example, by combining information about activity in the posterior parietal cortex, mediating attention to a particular stimulus attribute, and information about the stimulus, can we identify regions that respond to that stimulus when, and only when, activity in the parietal region 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 (see Fig. 1). This has clear ramifications in terms of the top-down modulation of specialised cortical areas by higher brain regions. This approach is interesting from two points of view. Firstly, the explanatory variables used to predict activity in any brain region (i.e., the response variable) comprises a standard predictor variable based on the experimental design (e.g., the presence or absence of a particular stimulus attribute) and a response variable from another part of the brain. The second reason that this analysis is interesting is that it uses techniques usually used to make inferences about functional specialisation to infer something about integration and vice versa. The statistical model employed in testing for psychophysiological interactions is a simple regression model of effective connectivity that embodies non-linear (second-order or modulatory effects). As such this class of model speaks directly to functional specialisation of a non-linear and contextual sort. Figure 3 illustrates a specific example (see [7] 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 most significantly expressed in the right infero-temporal region. Changes in parietal activity were introduced experimentally by pre-exposure of the stimuli before some scans but not others. The data in the lower right panel of Fig. 3 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 instantiation of attentional, memory or learning differences in face-specific responses, in infero-temporal regions that are medi-

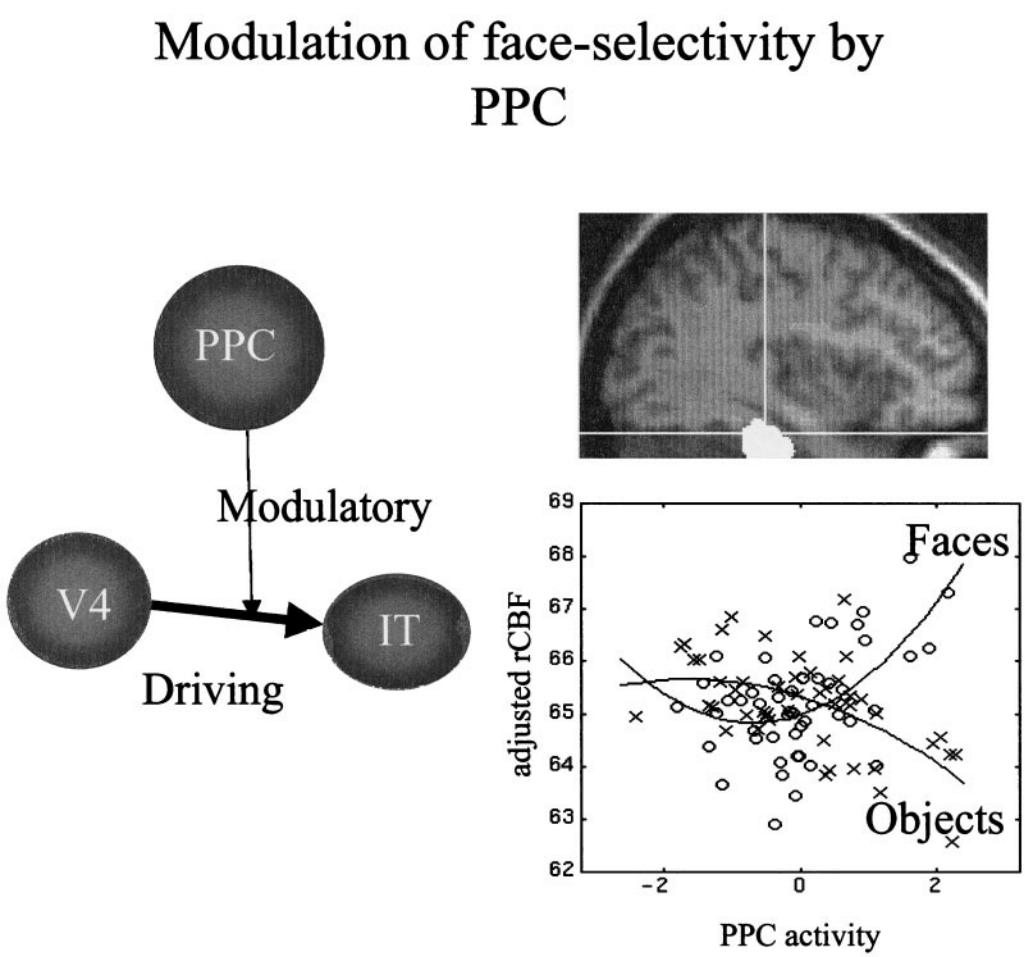

FIG. 3. Top right: A statistical parametric map (SPM) that identifies areas whose activity can be explained on the basis of an interaction between the presence of faces in visually presented stimuli and activity in a reference location in the posterior [medial] parietal cortex (PPC). This analysis can be thought of as finding those areas that are subject to top-down modulation of face-specific responses by medial parietal activity. The largest effect was observed in the right infero-temporal region. Lower right: The corresponding activity is displayed as a function of [mean corrected] PPC 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 positron emission tomography. Left: Schematic depicting the underlying conceptual model in which driving afferents from ventral form areas (here designated as V4) excite responses in infero-temporal (IT) subject to permissive modulation by PPC afferents. Abbreviation: rCBF, regional cerebral blood flow.

ated by interactions with medial parietal cortex. Note that we could have included the priming effect explicitly in the statistical model but chose to substitute parietal activity in its place, enabling us to make a more mechanistic inference: Namely, not only do infero-temporal, face-specific responses show priming but this priming is mediated by modulatory influences from a higher (parietal) area. This is a clear example of contextual specialisation that depends on top-down non-linear effects.

THE LESION-DEFICIT MODEL REVISITED

If it is the case that functional specialisation depends on modulatory interactions among cortical areas then one would predict changes in functionally-specific responses in cortical regions that receive 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 regionally specific dysfunction, caused by, but remote from, a lesion that is itself context-dependent (i.e., elicited by some tasks but not others). We have referred to this phenomenon as 'dynamic diaschisis' [31].

Dynamic Diaschisis

In this section, we describe the pathophysiological phenomenon of '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

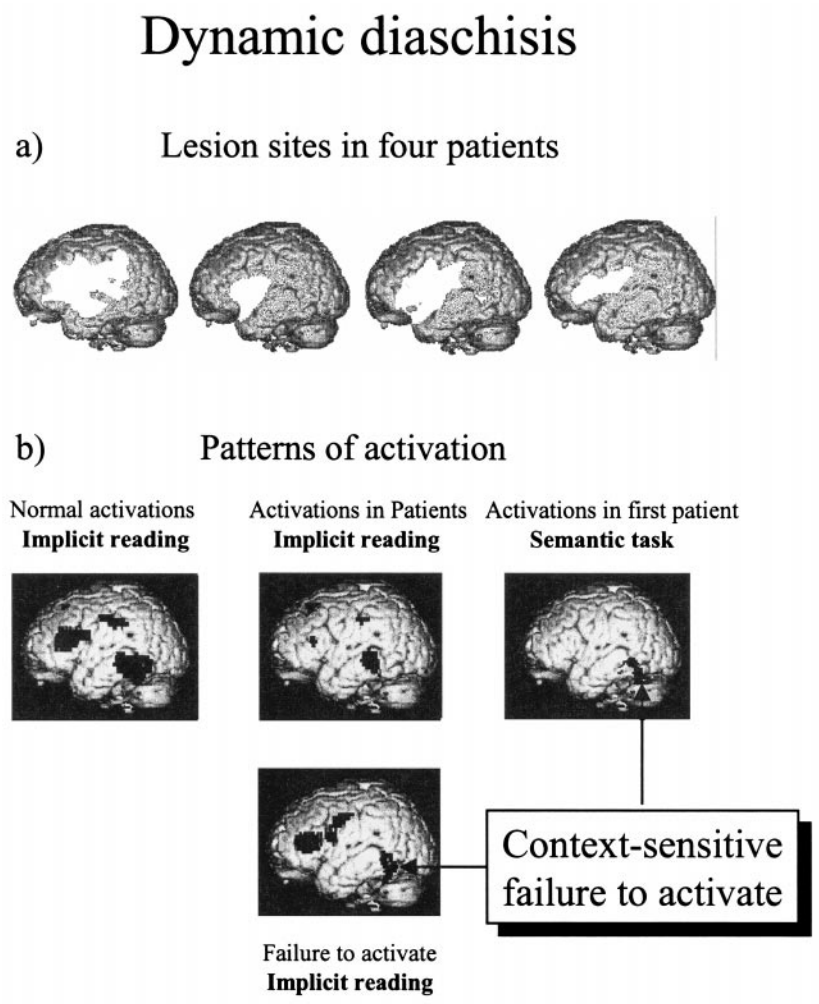

FIG. 4. (a) These rendering illustrate the extent of the cerebral infarcts, as identified by voxel-based morphometry. Regions of reduced grey matter (relative to neurologically normal controls) are shown in white on the left hemisphere. The statistical parametric maps (SPMs) were thresholded at $p < 0.001$ uncorrected. All patients had damage to Broca's area. The first (upper left) patient's left middle cerebral artery infarct was most extensive encompassing temporal and parietal regions as well as frontal and motor cortex. (b) SPMs illustrating the functional imaging results with regions of significant activation shown in black on the left hemisphere. Results are shown for: Normal subjects reading words; activations common to normal subjects and patients reading words; the first patient activating normally for a semantic task but abnormally for the implicit reading task; and (below) areas where normal subjects activated significantly more than patients during implicit reading.

to, damaged regions. The clearest example is 'crossed cerebellar diaschisis' [22] in which abnormalities of cerebellar metabolism are characteristically seen 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 driving and modulatory inputs from, and reciprocal interactions with, many 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 neural dynamics depend only upon integration with undamaged regions. If the region can respond normally in some situations then forward driving components must be intact. This suggests that dynamic diaschisis will only present itself when the lesion involves a hierarchically equivalent or higher area.

An 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. 4a). These patients had speech output deficit but relatively pre-

served comprehension. Generally functional imaging studies can only make inferences about abnormal neuronal responses when changes in cognitive strategy can be excluded. We ensured this by engaging the patients in an explicit task that they were able to perform normally. This involved a key press response when a visually presented letter string contained a letter with an ascending visual feature (e.g., h, k, l, or t). While the task remained constant, the stimuli presented were either words or consonant letter strings. Activations detected for words, relative to letters, were attributed to implicit word processing. Each patient showed normal activation of the left posterior middle temporal cortex, that has been associated with semantic processing [39]. 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. 4b). These two regions are crucial for word production [29]. Examination of individual responses in this area revealed that all 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 (1) lies adjacent and posterior to a region of the left middle temporal cortex that activated normally (see middle column of Fig. 4b); and (2) 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 [30], the left posterior inferior temporal lobe responded normally. The abnormal 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 specific form of functional disconnection where regional dysfunction can be attributed to the loss of modulatory or 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 profound implications for neuropsychological inference.

CONCLUSION

The central idea that we have presented is that functionally specific responses may be constructed by interactions among neuronal systems where the specificity does not rest upon the intrinsic response profiles of the neurons themselves but on the context in which these responses are elicited. This context is established by inputs from higher brain areas. If this is correct then there are two approaches that are clear candidates for characterising functional specificity in the human brain. Both rely on neuroimaging and involve (1) an analysis of the effective connectivity among cortical regions and (2) an explicit analysis of the context-sensitivity of the responses elicited. In terms of analyses of effective connectivity we are already in a position, using fMRI, to make inferences about modulatory or context-dependant changes in effective connection strengths as illustrated by the results on attentional modulation in Fig. 1. One can envisage similar experiments addressing the modulation of connections from V2 to, for example, the fusiform region, by putative naming areas in the basal infero-temporal region. These experiments would point to the role of top-down modulation in configuring word-specific responses in early components of the ventral visual pathway. The second experimental approach depends upon measuring context-sensitive responses associated with contextual representations. This effectively reduces to looking for interactions and calls for multifactorial designs as discussed elsewhere [13]. As noted in the section on "Functional Specialisation and Brain Imaging", this general approach has been refined to incorporate the influence of remote areas on regionally specific responses using psychophysiological interactions. An example of this sort of approach could test for interactions between activity in the basal infero-temporal region and the presence of words in visually presented letter strings. This multifactorial approach is motivated, quite simply by the presence of possible modulatory or non-linear effects mediated by functional integration and would be impossible to implement using lesion studies. This line of argument suggests that it may no longer be sufficient to demonstrate, say, a face-specific area by simply presenting face and house stimuli to subjects. A multifactorial approach that emphasised the context-sensitive nature of this specificity would involve the presentation of houses and faces, both under two task conditions that emphasised house and face processing, respectively.

ACKNOWLEDGEMENT

This work was funded by the Wellcome Trust.

REFERENCES

1. Abbot, L. F.; Varela, J. A.; Sen, K.; Nelson, S. B. Synaptic depression and cortical gain control. Science 275:220–223; 1997.
2. Absher, J. R.; Benson, D. F. Disconnection syndromes: An overview of Geschwind's contributions. Neurology 43:862–867; 1993.
3. Aertsen, A.; Preißl, H. Dynamics of activity and connectivity in physiological neuronal networks. In: Schuster, H. G., ed. Non linear dynamics and neuronal networks. New York: VCH Publishers Inc.; 1991:281–302.
4. Bÿchel, C.; Friston, K. J. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. Cereb. Cortex 7:768–778; 1997.
5. Crick, F.; Koch, C. Constraints on cortical and thalamic projections: The no-strong-loops hypothesis. Nature 391:245–250; 1998.
6. Dayan, P.; Hinton, G. E.; Neal, R. M. The Helmholtz machine. Neural Comput. 7:889–904; 1995.
7. Dolan, R. J.; Fink, G. R.; Rolls, E.; Booth, M.; Holmes, A.; Frackowiak, R. S. J.; Friston, K. J. How the brain learns to see objects and faces in an impoverished context Nature 389:596–598; 1997.
8. Felleman, D. J.; Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1:1–47; 1991.
9. Fletcher, P. C.; Frith, C. D.; Grasby, P. M.; Shallice, T.; Frackowiak, R. S. J.; Dolan, R. J. Brain systems for encoding and retrieval of auditory-verbal memory. Brain 118:401–416; 1995.
10. Friston, K. J. Functional and effective connectivity inneuroimaging: A synthesis Hum. Brain Mapp. 2:56–78; 1995.
11. Friston, K. J.; Ungerleider, L. G.; Jezzard, P.; Turner, R. Characterizing modulatory interactions between V1 and V2 in human cortex with fMRI. Hum. Brain Mapp. 2:211–224; 1995.
12. Friston, K. J.; Holmes, A. P.; Worsley, K. J.; Poline, J. B.; Frith, C. D.; Frackowiak, R. S. J. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2:189–210; 1995.
13. Friston, K. J.; Price, C. J.; Fletcher, P.; Moore, C.; Frackowiak, R. S. J.; Dolan, R. J. The trouble with cognitive subtraction. Neuroimage 4:97–104; 1996.
14. Friston, K. J.; Bÿchel, C.; Fink, G. R.; Morris, J.; Rolls, E.; Dolan, R. J. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229; 1997.
15. Gerstein, G. L.; Perkel, D. H. Simultaneously recorded trains of action potentials: Analysis and functional interpretation. Science 164:828–830; 1969.
16. Girard, P.; Bullier, J. Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. J. Neurophysiol. 62:1287–1301; 1989.

17. Goltz, F. In: MacCormac, W., ed. Transactions of the 7th International Medical Congress, vol. I. London: J. W. Kolkmann; 1881:218–228.
18. Hinton, G. T.; Shallice, T. Lesioning an attractor network: Investigations of acquired dyslexia. Psychol. Rev. 98:74–95; 1991.
19. Hinton, G. E.; Dayan, P.; Frey, B. J.; Neal, R. M. The "wake-sleep" algorithm for unsupervised neural networks. Science 268:1158–1161; 1995.
20. Hirsch, J. A.; Gilbert, C. D. Synaptic physiology of horizontal connections in the cat's visual cortex. J. Neurosci. 11:1800–1809; 1991.
21. Kay, J.; Phillips, W. A. Activation functions, computational goals and learning rules for local processors with contextual guidance. Neural Comput. 9:895–910; 1996.
22. Lenzi, G. L.; Frackowiak, R. S. J.; Jones, T. J. Cerebral oxygen metabolism and blood flow in human cerebral ischemic infarction. Cereb. Blood Flow Metab. 2:321–335; 1982.
23. Lichtheim, L. On aphasia. Brain 7:422–484; 1885.
24. 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. The colour centre in the cerebral cortex of man. Nature 340:386–389; 1989.
25. McIntosh, A. R.; Gonzalez-Lima, F. Structural equation modelling and its application to network analysis in functional brain imaging. Hum. Brain Mapp. 2:2–22; 1994.
26. Petersen, S. E.; Fox, P. T.; Posner, M. I.; Mintun, M.; Raichle, M. E. Positron emission tomographic studies of the processing of single words. J. Cogn. Neurosci. 1:153–170; 1989.
27. Phillips, C. G.; Zeki, S.; Barlow, H. B. Localization of function in the cerebral cortex: Past present and future. Brain 107:327–361; 1984.
28. Phillips, W. A.; Singer, W. In search of common foundations forcortical computation. Behav. Brain Sci. 20:57–83; 1997.
29. Price, C. J. The functional anatomy of word comprehension and production. Trends. Cogn. Sci. 2:281–288; 1998.
30. Price, C. J.; Mummery, C. J.; Moore, C. J.; Frackowiak, R. S. J.; Friston, K. J. Delineating necessary and sufficient neural systems with functional imaging studies of neuropsychological patients. J. Cogn. Neurosci.11:371–382; 1999.
31. Price, C. J.; Warburton, E. A.; Moore, C. J.; Frackowiak, R. S. J.; Friston, K. J. Dynamic diaschisis: Anatomically remote and context specific human brain lesions. J. Cogn. Neurosci.; in press.
32. Rao, R. P. N.; Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. Nat. Neurosci. 2:79–87; 1999.
33. Rockland, K. S.; Pandya, D. N. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res. 179:3–20; 1979.
34. Salin, P.-A.; Bullier, J. Corticocortical connections in the visual system: Structure and function. Psychol. Bull. 75:107–154; 1995.
35. Sandell, J. H.; Schiller, P. H. Effect of cooling area 18 on striate cortex cells in the squirrel monkey. J. Neurophysiol. 48:38–48; 1982.
36. Sugase, Y.; Yamane, S.; Ueno, S.; Kawano, K. Global and fine information coded by single neurons in the temporal visual cortex. Nature 400:869–873; 1999.
37. Talairach, P.; Tournoux, J. A stereotactic coplanar atlas of the human brain. Stuttgart: Thieme; 1988.
38. Treue, S.; Maunsell, H. R. Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382:539–541; 1996.
39. Vandenberghe, R.; Price, C.; Wise, R.; Josephs, O.; Frackowiak, R. S. J. Functional anatomy of a common semantic system for words and pictures. Nature 383:254–256; 1996.
40. Zeki, S.; Shipp, S. The functional logic of cortical connections. Nature 335:311–317; 1988.
41. Zeki, S. The motion pathways of the visual cortex. In: Blakemore, C., ed. Vision: Coding and efficiency. Cambridge, UK: Cambridge University Press; 1990:321–345.