Annu. Rev. Neurosci. 2002. 25:221–50
doi: 10.1146/annurev.neuro.25.112701.142846

BEYOND PHRENOLOGY: What Can Neuroimaging Tell Us About Distributed Circuitry?

Karl Friston
The Wellcome Department of Cognitive Neurology, University College London, Queen Square, London, WC1N 3BG United Kingdom; email: k.friston@fil.ion.ucl.ac.uk

Key Words neuroimaging, predictive coding, generative model, information theory, effective connectivity

■ **Abstract** Unsupervised models of how the brain identifies and categorizes the causes of its sensory input can be divided into two classes: those that minimize the mutual information (i.e., redundancy) among evoked responses and those that minimize the prediction error. Although these models have the same goal, the way that goal is attained, and the functional architectures required, are fundamentally different. This review describes the differences, in the functional anatomy of sensory cortical hierarchies, implied by the two models. We then consider how neuroimaging can be used to disambiguate between them. The key distinction reduces to whether backward connections are employed by the brain to generate a prediction of sensory inputs. To ascertain whether backward influences are evident empirically requires a characterization of functional integration among brain systems. This review summarizes the approaches to measuring functional integration in terms of effective connectivity and proceeds to address the question posed by the theoretical considerations. In short, it will be shown that the conjoint manipulation of bottom-up and top-down inputs to an area can be used to test for interactions between them, in elaborating cortical responses. The conclusion, from these sorts of neuroimaging studies, points to the prevalence of top-down influences and the plausibility of generative models of sensory brain function.

INTRODUCTION

Functional neuroimaging, or human brain mapping, has enjoyed an enormous amount of success in systems and cognitive neuroscience over the past decade. Much of this success rests on being able to identify functionally specialized areas. Implicit in the term mapping is cartography, which some have referred to as neo-phrenology. Functional cartography, by itself, is clearly not going to reveal the principles that underlie the brain's functional architectures. However, it is an important prelude. This review is about using neuroimaging to answer questions about organizational principles, in terms of distributed and coupled interactions among specialized brain systems. The question, chosen to illustrate this sort of application, concerns the respective roles of forward and backward connections

among cortical areas and how this coupling mediates perceptual synthesis and categorization. The first half of this review establishes the potential importance of backward connections using ideas from theoretical neurobiology and machine learning. The ensuing predictions are then addressed using empirical examples from neuroimaging.

The article starts by reviewing two fundamental principles of brain organization, namely functional specialization and functional integration, and how they rest on the anatomy and physiology of cortico-cortical connections in the brain. The second section deals with the nature of representations from a theoretical or computational perspective. This section contrasts information theoretic approaches and those predicated on predictive coding. This section concludes that predictive coding architectures are more plausible because they lend themselves to a Bayesian formulation, in which constraints from higher levels of a cortical hierarchy provide contextual guidance to lower levels of processing. This confers a context-sensitivity on evoked responses.

Empirical evidence from electrophysiological studies of animals and functional neuroimaging studies of human subjects is presented in the third and fourth sections to illustrate the context-sensitive nature of functional specialization and how its expression depends on functional integration among remote cortical areas. The third section (on generative models and the brain) 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 pursued at a cortical level in human functional neuroimaging studies in the subsequent section (on functional architectures and brain imaging). 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 a predictive coding strategy.

FUNCTIONAL SPECIALIZATION AND INTEGRATION

The brain appears to adhere to two fundamental principles of functional organization, functional integration and functional specialization, where the integration within and among specialized areas is mediated by effective connectivity. The distinction relates to that between "localizationism" and "[dis]connectionism," which dominated thinking about cortical function in the nineteenth century. Since the early anatomic theories of Gall, the identification of a particular brain region with a specific function has become a central theme in neuroscience. However, functional localization per se was not easy to demonstrate: For example, a meeting that took place on August 4, 1881 addressed the difficulties of attributing function to a cortical area, given the dependence of cerebral activity on underlying connections (Phillips et al. 1984). This meeting was entitled "Localization of Function in the Cortex Cerebri." Goltz, although accepting the results of electrical stimulation in dog and monkey cortex, considered that the excitation method was inconclusive in that the behaviors elicited might have originated in related pathways, or current

could have spread to distant centers. In short the excitation method could not be used to infer functional localization because localizationism discounted interactions, or functional integration among different brain areas. It was proposed that lesion studies could supplement excitation experiments. Ironically, it was observations on patients with brain lesions some years later (see Absher & Benson 1993) that led to the concept of disconnection syndromes and the refutation of localizationism as a complete or sufficient explanation of cortical organization. Functional localization implies that a function can be localized in a cortical area, whereas specialization suggests that a cortical area is specialized for some aspects of perceptual or motor processing, in which this specialization can be anatomically segregated within the cortex. The cortical infrastructure supporting a single function may then involve many specialized areas whose union is mediated by the functional integration among them. Functional specialization and integration are not exclusive; they are complementary. Functional specialization is only meaningful in the context of functional integration and vice versa.

Functional Specialization

The functional role, played by any component (e.g., cortical area, subarea, 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 interstripes. When recordings are made in V2, directionally selective (but not wavelength or color selective) cells are found exclusively in the thick stripes. Retrograde (i.e., backward) labeling of cells in V5 is limited to these thick stripes. All the available physiological evidence suggests that V5 is a functionally homogeneous area that is specialized for visual motion. Evidence of this nature supports the notion that patchy connectivity is the anatomical infrastructure that underpins functional segregation and specialization.

THE ANATOMY AND PHYSIOLOGY OF CORTICO-CORTICAL CONNECTIONS If specialization depends on connectivity, then important principles underpinning specialization should be embodied in the neuroanatomy and physiology of extrinsic connections. Extrinsic connections couple different cortical areas, whereas intrinsic connections are confined to the cortical sheet. There are certain features of cortico-cortical connections that provide strong clues about their functional role. In brief, there appears to be a hierarchical organization that rests on the distinction

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

Hierarchical organization

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

Forward connections	**Backward connections**
Sparse axonal bifurcations	Abundant axonal bifurcation
Topographically organized	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

between forward and backward connections. The designation of a connection as forward or backward depends primarily on its cortical layers of origin and termination. Some characteristics of cortico-cortical connections are summarized in Table 1. In brief, the anatomy and physiology of cortico-cortical connections suggest that forward connections are driving and commit cells to a prespecified response given the appropriate pattern of inputs. Backward connections, on the other hand, are less topographic and are in a position to modulate the responses of lower areas to driving inputs from either higher or lower areas (see Table 1). Reversible inactivation (e.g., Sandell & Schiller 1982, Girard & Bullier 1989) and functional neuroimaging (e.g., Büchel & Friston 1997) studies are consistent with this distinction. The notion that forward connections are concerned with the promulgation and segregation of sensory information is consistent with their (*a*) sparse axonal bifurcation, (*b*) patchy axonal terminations, and (*c*) 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 their (*a*) frequent bifurcation, (*b*) diffuse axonal terminations, and (*c*) nontopographic projections (Salin & Bullier 1995, Crick & Koch 1998).

In summary, backward connections are abundant and are in a position to exert powerful effects on evoked responses, in lower levels, that define the specialization of any area or neuronal population. The idea pursued in this review is that specialization depends on backward 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 specialization of brain areas that mediate perceptual synthesis and adaptive behavioral responses.

Functional Integration and Effective Connectivity

Electrophysiology and imaging neuroscience have firmly established functional specialization as a principle of brain organization in man. The functional integration of specialized areas has proven more difficult to assess. Functional integration refers to the interactions among specialized neuronal populations and how these interactions depend on 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 they can reflect stimulus-induced oscillations mediated by synaptic connections (Gerstein & 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 & Preißl 1991). This speaks to two important points: (*a*) Effective connectivity is dynamic, i.e., activity- and time-dependent and (*b*) it depends on a model of the interactions. The models employed in functional neuroimaging can be divided into those based on regression models (Friston 1995) and those based on structural equation models (McIntosh & Gonzalez-Lima 1994). A more important distinction is whether these models are linear or nonlinear. Recent characterizations of effective connectivity have focused on nonlinear models that accommodate the modulatory or nonlinear effects described above. The most general model one could envisage is provided by nonlinear system identification through the use of Volterra series (see Box 1). This model has been used to address nonlinear coupling among brain areas induced by attention (Friston & Büchel 2000). We will use this example in the functional architectures assessed with brain imaging section.

Box 1 Dynamical Systems, Volterra Kernels, and Effective Connectivity

Input-State-Output Systems and Volterra Series

Neuronal systems are inherently nonlinear and lend themselves to modeling 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 & Green 1994) and focus on the inputs and outputs. Consider the single input–single output system:

$$\dot{x}(t) = f(x(t), u(t))$$
$$y(t) = \lambda(x(t)).$$

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, which expresses the output $y(t)$ as 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 $y(t)$.

$$y(t) = F(u(t-\sigma)) = \kappa_0 + \sum_{i=1}^{\infty} \int_0^t \cdots \int_0^t \kappa_i(\sigma_1, \ldots \sigma_i)\, u(t-\sigma_1) \ldots u(t-\sigma_i)\, d\sigma_1 \ldots d\sigma_i$$

$$\kappa_i(\sigma_1, \ldots \sigma_i) = \frac{\partial^i y(t)}{\partial u(t-\sigma_1) \ldots \partial u(t-\sigma_i)},$$

where $\kappa_i(\sigma_1, \ldots \sigma_i)$ is the ith-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. When the inputs and outputs are measured neuronal activity, the Volterra kernels have a special interpretation.

Volterra Kernels and Effective Connectivity

Volterra kernels are essential in characterizing 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:

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

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 evoked response at $t - \sigma_2$, and so on for higher orders.

The important thing about this formulation of effective connectivity is that it can be defined and estimated using just the inputs and responses of a system (e.g., cortical area). In other words, effective connectivity does not refer to the (hidden) state variables (e.g., depolarization of every cell membrane, the electrochemical status of every cell compartment, or the configuration of every membrane channel) that actually mediate the input-output transformation. This is important because these state variables are often unmeasurable, particularly in functional neuroimaging.

THEORETICAL AND COMPUTATIONAL PERSPECTIVES

This section compares and contrasts two prevalent computational approaches to perceptual categorization and synthesis: information theoretic and predictive coding frameworks. This section restricts itself to sensory processing in cortical hierarchies. This precludes a discussion of other important ideas [e.g., reinforcement learning (Sutton & Barto 1990, Friston et al. 1994), neuronal selection (Edelman 1993), and dynamical systems theory (Freeman & Barrie (1994)].

The relationship between modeled and real neuronal architectures is central to basic and cognitive neuroscience. This section addresses this relationship, in terms of representations. We start with an overview of representations so that the distinctions among various approaches can be seen clearly. An important focus of this section is the interactions among "causes" of sensory input. These interactions posit the problem of contextual invariance, which has severe implications for supervised (i.e., connectionist) models of cognitive architectures. In brief the problem of contextual invariance points to the adoption of unsupervised models in which interactions among causes of a percept are modeled explicitly. Within the class of unsupervised models we compare classical information theoretic approaches and predictive coding. These schemes allow the emergence of natural representations that can accommodate contextual invariance but do so in a very different way. The question then reduces to how this difference would be expressed in terms of measurable brain responses and effective connectivity. This issue is taken up in subsequent sections.

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. It can be defined operationally as the neuronal transient evoked by the cause being represented. In very general terms, let us frame the problem of representing real world causes $u(t)$ in terms of the system of equations

$$\dot{y}(t) = f(y(t), u(t)), \qquad 1.$$

where u is a vector describing the expression of causes in the environment (e.g., the presence of a particular object, direction of radiant light, etc.), and y represents sensory inputs. $\dot{y}(t)$ denotes the rate of change of y at time t. The function f can be highly nonlinear and allows for both the current state of the sensory inputs and their causes to interact when inducing changes in the activity of sensory units. The sensory input can be shown to be a function of, and only of, the causes and their recent history.

$$y(t) = F(u(t-\sigma))$$
$$= \sum_{i=1}^{\infty} \int_0^t \cdots \int_0^t \frac{\partial^i y(t)}{\partial u(t-\sigma_1) \ldots \partial u(t-\sigma_i)} u(t-\sigma_1) \ldots u(t-\sigma_i)\, d\sigma_1 \ldots d\sigma_i. \quad 2.$$

Equation 2 is simply a functional Taylor expansion to cover dynamical systems of the sort implied by Equation 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 smoothing over time. The importance of this formulation is that it highlights (*a*) the dynamical aspects of sensory input and (*b*) the role of interactions among the causes of the sensory input. For example the second-order terms with $i = 2$ in Equation 2 represent pairwise interactions among u, possibly at different points in time. Interactions can be viewed as contextual effects, for which the expression of a particular cause is highly dependent on the context induced by another. For example, the extraction of motion from the visual field depends on there being sufficient luminance or wavelength contrast. 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 objects renders 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 how to find a function of the input $y(t)$ that represents the underlying causes. To do this, the brain must effectively undo the interactions to reveal contextually invariant causes. In other words the brain must perform some form of nonlinear unmixing of causes and context without ever knowing either. Furthermore, because of the convolution implied by Equation 2, it must deconvolve the inputs to obtain these causes. In estimation theory this general problem is sometimes called blind deconvolution because the estimation is blind to the underlying causes that are convolved to give the observed variables.

Most models of perceptual categorization can be understood as trying to effect a blind deconvolution of sensory inputs to reveal the causes. Consider a formally

similar system of equations to Equation 1 that represent the dynamics of the brain:

$$\dot{x}(t) = f_\theta(x(t), y(t))$$
$$v(t) = l_\theta(x(t)) \quad 3.$$

and by analogy with Equation 2 $v(t) = F_\theta(y(t - \sigma))$. 4.

Here x represents the activity of neuronal units (i.e., neurons or populations of neurons) in the brain. A subset of units can be selected and passed through a nonlinear function to give some explicit or implicit representation v. The parameters θ of the functions in Equation 3 embody the series of dynamical transformations that the sensory input is subject to and can be thought of specifying the connection strengths and biases of a neuronal network model or effective connectivity. The problem of extracting causes from the input reduces to finding the right parameters such that the activity of the representational units v have some clearly defined relationship to the causes u. More formally one wants to find the parameters that maximize the mutual information or 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 parameters. The greater the biological validity of the constraints under which these solutions are obtained, the more plausible the relationship becomes.

In what follows we will consider models based on information theory and those based on predictive coding. Each subsection below provides the background for the approach and then describes it using the formalism above. Figure 1 provides a graphical overview of the two schemes.

Information Theoretic Approaches

There have been many compelling developments in theoretical neurobiology that have used information theory (e.g., Barlow 1961, Optican & 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 & Redlich 1990, Bell & 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 & Redlich 1990, Olshausen & 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 be configured to maximize the mutual information between the representations that they engender and the causes of sensory input. This maximization is usually considered in the light of some sensible constraints, e.g., the presence of noise in the sensory input (Atick & Redlich 1990) or dimension reduction (Oja 1988).

For any given causes we want to maximize the mutual information between $u(t)$ and the neuronal responses $v(t)$. Intuitively mutual information is like the

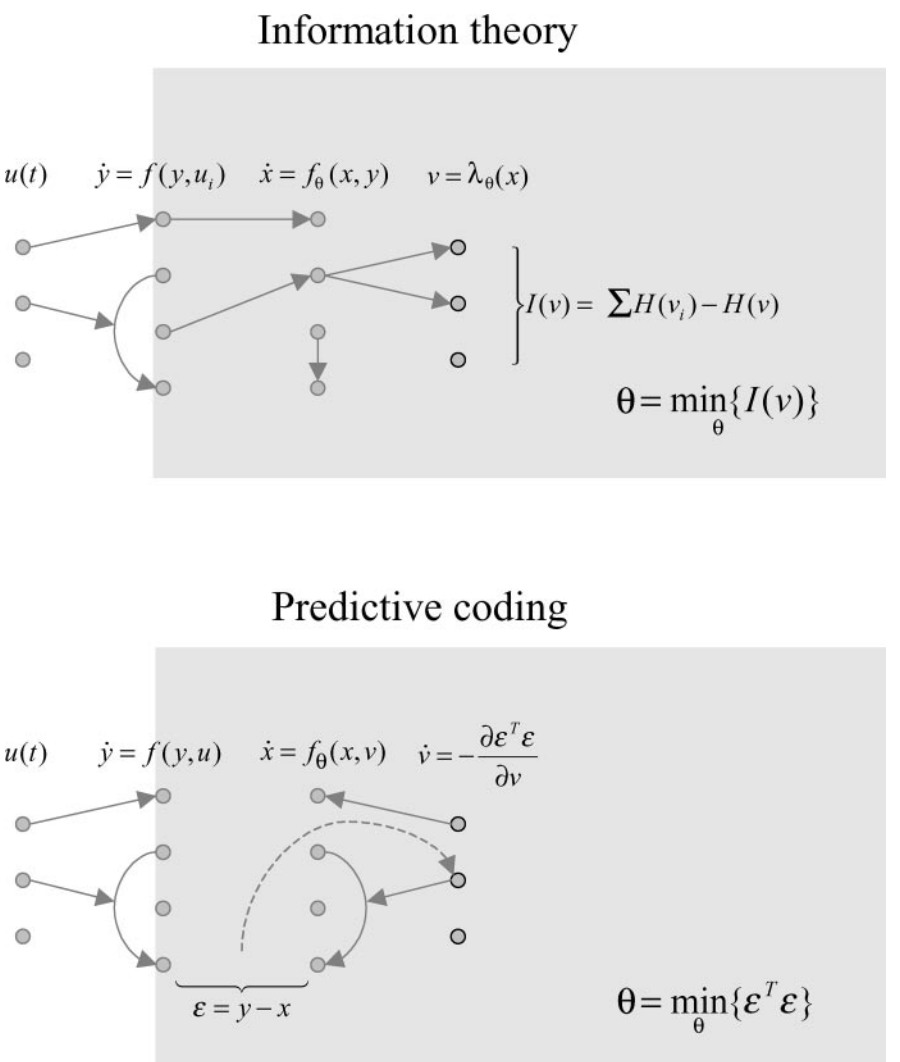


Figure 1 Schematic illustrating the architectures implied by information theory–based approaches and predictive coding. The circles represent nodes in a network and the arrows represent a few of the connections. See the main text for an explanation of the equations and designation of the variables each set of nodes represents. The light grey boxes encompass connections and nodes within the model. The strengths of connections within this area are determined by the free parameters of the model θ. 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 or interpretation). The broken arrow in the lower panel denotes connections that convey an error signal to the higher level from the input level.

covariance or correlation between two variables but extended to cover multivariate observations. 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 is given by

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

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

$$\frac{\partial}{\partial \theta} I(u, v) = \frac{\partial}{\partial \theta} H(v). \quad 6.$$

It follows that maximizing the mutual information between the outputs and the causes is the same as maximizing the entropy of the responses. The infomax principle (maximum information transfer) is closely related to the idea of efficient coding. Generally speaking, redundancy minimization and efficient coding are all variations on the same theme and can be considered as the infomax principle operating under some appropriate constraints. The key thing that distinguishes among the various information theoretic schema is the nature of the constraints under which entropy is maximized. These constraints render the infomax a viable approach to recovering the original causes of data, especially if one can enforce the outputs to comply with the same constraints as the causes. One useful way of looking at constraints is in terms of efficiency.

EFFICIENCY, REDUNDANCY, AND INFORMATION Efficiency 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 & Richmond 1993).

$$I(v) = \sum H(v_i) - H(v). \quad 7.$$

Here $H(v_i)$ is the entropy of the *i*th unit. Equation 7 implies that redundancy is the difference between the joint entropy and the sum of the entropies of the individual units (componential entropies). Intuitively this expression makes sense if one considers that the variability in activity of any one unit corresponds to its entropy. Therefore, an efficient system represents its inputs with the minimum changes in firing. Another way of thinking about Equation 7 is to note that maximizing efficiency is equivalent to minimizing the mutual information among the outputs. This is the basis of approaches that seek to decorrelate or orthogonalize the outputs.

To minimize redundancy one can either minimize the entropy of the output units or maximize their joint entropy. Olshausen & Field (1996) present a very nice analysis based on sparse coding. Sparse coding minimizes the redundancy by minimizing componential entropies. This minimization is implicit in sparse coding

because a neuron that fires very sparsely will generally not be firing. We can, therefore, be relatively certain about its (quiescent) state, conferring low entropy upon it.

Approaches that seek to maximize the joint entropy of the outputs include principle component analysis (PCA), learning algorithms (that sample the subspace of the inputs that have the highest entropy) (e.g., Foldiak 1990, Friston et al. 1993), and independent component analysis (ICA). ICA finds nonlinear functions of the inputs that maximize the joint entropy (Common 1994, Bell & Sejnowski 1995). In PCA the componential entropies are constrained by setting the sum of squared connection strengths to be one. In ICA they are maintained at low levels by the application of a sigmoid squashing function to the outputs.

IMPLEMENTATION In terms of the above formulation, information theoretic approaches can be construed as finding the parameters that maximize the efficiency or minimize the redundancy

$$\theta = \min_{\theta} I(v). \qquad 8.$$

Compared to supervised schemes this has the fundamental advantage that the algorithm is unsupervised (the causes do not enter into Equation 8). Furthermore, the inputs can, in principle be generated dynamically and can interact. However, for simple variants of this information theoretic approach (e.g., ICA) only linear mixtures of independent causes can be recovered (up to some permutation and scaling). For example a typical model adopted by PCA for Gaussian and ICA for non-Gaussian causes is

$$y(t) = F(u(t - \sigma)) = Wu(t), \qquad 9.$$

where W is linear mixing matrix. This example highlights the operational shortcomings of information theoretic approaches that are based on feedforward architectures: Not only does the model of real world mixing of causes in Equation 9 preclude any dynamics, it also ignores interactions among causes. Nonlinear variants of ICA and PCA do exist (e.g., Karhunen & Joutsensalo 1994, Dong & McAvoy 1996) and typically employ a "bottleneck" architecture that forces the inputs through a small number of nodes. The output from these nodes then diverges to predict the original inputs (see predictive coding below). However, these architectures are better regarded as generative models in the sense that the nonlinear transformations, from the bottleneck nodes to the output layer, recapitulate the nonlinear mixing of the original causes and constitute a generative model. Generative models are presented in the next subsection.

Finally ICA, like parallel and distributed processing models, assumes the existence of an operator that can deconvolve (a nonlinear function of) the causes out of the inputs. For very simple mixtures of causes this may be tenable. However, generally nonlinear mixing applied by the real world renders the existence of this deconvolution very questionable. In the inverse solution literature these problems are known as ill-posed or underdetermined. The solution is to render the problem tractable using constraints on the solution. This means that information

theoretic approaches that try to solve the unmixing or inverse problem rely heavily on constraints (e.g., efficiency, sparse coding, etc.). In the alternative approach, considered here, we discuss predictive coding models that moderate this constraint-dependency and suggest a more natural form for representations.

Generative Models and Predictive Coding

Over the past years generative models have supervened over other modeling approaches to brain function and represent one of the most promising avenues, offered by computational neuroscience, to understanding neuronal dynamics in relation to perceptual categorization. In generative models the dynamics of units in a network are trying to predict the inputs. The representational aspects of any unit emerge spontaneously as the capacity to predict improves with learning. There is no a priori "labeling" of the units or any supervision in terms of what a correct response should be (cf. connectionism). 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. There are many forms of generative models that range from conventional statistical models (e.g., factor and cluster analysis) and those motivated by Bayesian learning (e.g., Dayan et al. 1995, Hinton et al. 1995) to biologically plausible models of visual processing (e.g., Rao & Ballard 1998). 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). These models emphasize the role of backward connections in mediating the prediction, at lower or input levels, based on the activity of units in higher levels.

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 their causes, they find functions of estimated causes that predict the inputs. As in approaches based on information theory, the causes do not enter into the learning rules and they are therefore unsupervised. Furthermore, they do not require the convolution of causes, engendering the inputs to be invertible. This is because generative models instantiate the forward solution, not the inverse solution. Here the forward solution is the nonlinear mixing of causes that by definition must exist. The estimation of the causes still rests on constraints, but these are now framed in terms of the generative model and have a much more direct relationship to casual processes in the real world. The ensuing mirror symmetry of the architecture is illustrated in Figure 1. Notice that the connections within the model are now going backward. In the predictive coding scheme the outputs now become the inputs such that

$$\dot{x}(t) = f_\theta(x(t), v(t)) \Rightarrow$$
$$x(t) = F_\theta(v(t - \sigma)), \qquad 10.$$

cf. Equation 3. The parameters now change so as to minimize some function G of the prediction error at the input level

$$\theta = \min_{\theta} G(\varepsilon)$$
$$\varepsilon = y - F_\theta(v). \quad 11.$$

The top-down inputs $v(t)$ now drive the predictions $x(t)$ of the input and the parameters of the backward connections forming these predictions change so as to minimize the prediction error. But what drives the top-down inputs? The casual estimates or representations change in the same way as the other free parameters of the model. They change to minimize prediction error, usually through gradient descent

$$\dot{v} = -\frac{\partial G(\varepsilon)}{\partial v}. \quad 12.$$

The error is conveyed from the input layer to the higher layer by forward connections that are rendered as a broken line in the lower panel of Figure 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 happen after training or learning and prediction error is largely eliminated. This implies that the prediction of the input becomes very precise $x(t) \to y(t)$, and consequently from Equation 2 and Equation 10

$$F_\theta(v(t-\sigma)) \to F(u(t-\sigma)). \quad 13.$$

In other words the brain's nonlinear convolution of the estimated causes reproduces exactly 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 connections or parameters of F_θ.

The dynamics of representational units or populations implied by Equation 12 represents the essential difference between this class of approaches and others. Only in predictive coding is the activity of the units driven explicitly to improve the representational capacity of the system. Predictive coding is a strategy that has some compelling (Bayesian) underpinnings (see below) and is not simply using a connectionist architecture in auto-associative mode or using error minimization to maximize mutual information transfer. It is a real time, dynamical scheme that embeds two concurrent processes. (*a*) The parameters of the model are changing so that the generative model emulates the real world mixing of causes, using their current estimates, and (*b*) the representations are converging to the best estimate of the causes extant at any time, using the generative model. Both the parameters and the state variables change in a mathematically identical way to minimize prediction error. The predictive coding scheme can easily accommodate dynamical and nonlinear mixing of causes in the real world. It does not require this mixing to be invertible, and it only requires the sensory inputs to be known. Before considering how the brain might perform predictive coding, we look at its motivation from another point of view.

PREDICTIVE CODING AND BAYESIAN INFERENCE One of the most important aspects of generative models is that they emphasize the role of the brain as an inferential machine (Dayan et al. 1995). From this perspective functional architectures exist, not to unmix the input to obtain the causes, but to make inferences about the causes and test the predictions against observed input. A compelling aspect of predictive coding schemes is that they lend themselves to a hierarchical extension that can be viewed in terms of Bayesian inference. In the simplest extension, let us suppose we had some expected values $\bar{u}$ of the causes, which were used to generate a prior prediction error $G(v - \bar{u})$ not at the level of the inputs but at the higher level of the causal representations v. The changes in v are now required to minimize the error at both levels so that

$$\dot{v} = -\frac{\partial}{\partial v}\left[G(y - F_\theta(v)) + G(v - \bar{u})\right]. \qquad 14.$$

The addition of this extra term renders the ensuing estimation of causes a Bayesian one in the following way. Bayesian inference allows one to posit the probability of the causes of some input or data given that data. This is in contradistinction to maximum likelihood estimates, which simply identify the causes that maximize the likelihood of the input. The difference rests on 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

$$\begin{aligned} p(u, y) &= p(u|y)p(y) = p(y|u)p(u) \\ &\Rightarrow p(u|y) \propto p(y|u)p(u). \end{aligned} \qquad 15.$$

The Bayesian, posterior, or conditional estimator of the causes is that which is most likely given the data.

$$\max_u p(u|y) = \max_u \{\ln p(y|u) + \ln p(u)\}. \qquad 16.$$

This is referred to as the maximum posterior or MAP estimator. The first term on the right is known as the log likelihood or likelihood potential and the second is the prior potential. If we take the Gibb's form for $p(y|u) = \exp(-\frac{1}{2}G\{y - F(u)\})$, then Equation 16 becomes

$$\max_u p(u|y) = \min_u \{G\{y - F(u)\} + G(u - \bar{u})\}. \qquad 17.$$

A gradient descent to find the MAP estimator would be

$$\dot{u} = -\frac{\partial}{\partial u}\left[G(y - F(u) + G(u - \bar{u})\right]. \qquad 18.$$

This is formally identical to Equation 14, the dynamics of the representations. This suggests that if the connectivity has properly captured the dynamical structure of the real world, i.e., $F_\theta(u) \to F(u)$ then the activities of representational units or

populations strive to encode the most probable causes given the input. In this Bayesian formulation the state of the brain changes, not to minimize error per se, but to attain an estimate of the causes that maximizes both the likelihood of the input given that estimate and the prior probability of the estimate being true.

This notion can be extended in a hierarchical fashion to any number of levels as depicted in Figure 2. In the forgoing we simply assumed some expected values for the causes. These expected values can, of course, be predictions from higher level causes. This extension models the world as a hierarchy of dynamical systems in which supraordinate causes induce, and moderate, changes in subordinate causes. For example, the presence of a particular object in the visual field induces changes in the incident light falling on a particular part of the retina. A more abstract example, which illustrates the brain's inferential capacities, is presented in Figure 3. On reading the first sentence "Jack and Jill went up the hill" we perceive the word

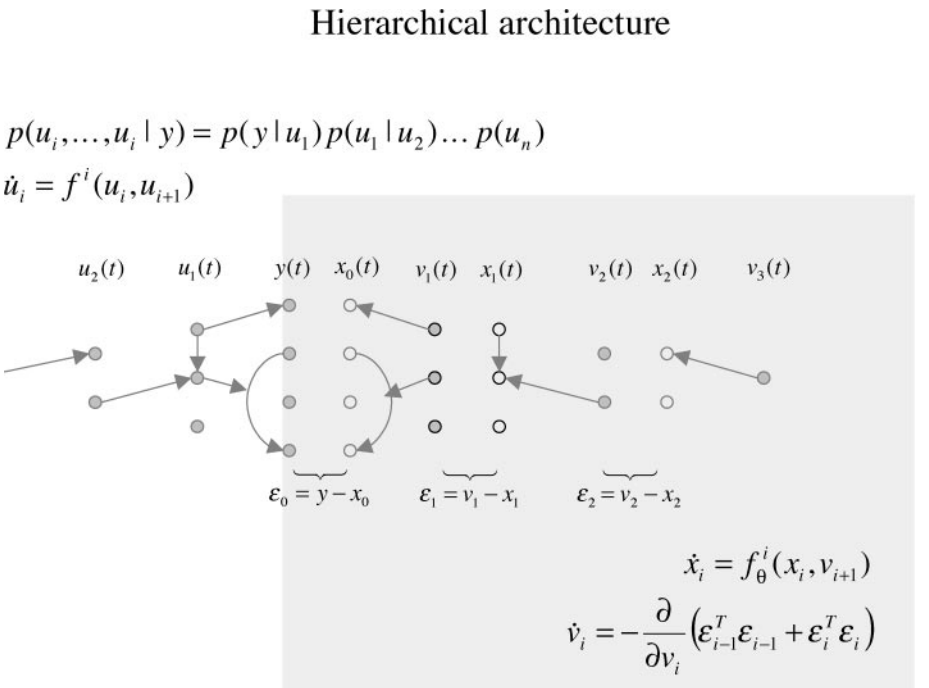


Figure 2 Schematic depicting a hierarchical extension to the predictive coding architecture, with the same format as Figure 1. Here hierarchical arrangements within the model serve to provide predictions or priors to representations in the level below. The open circles are the predictions and the filled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

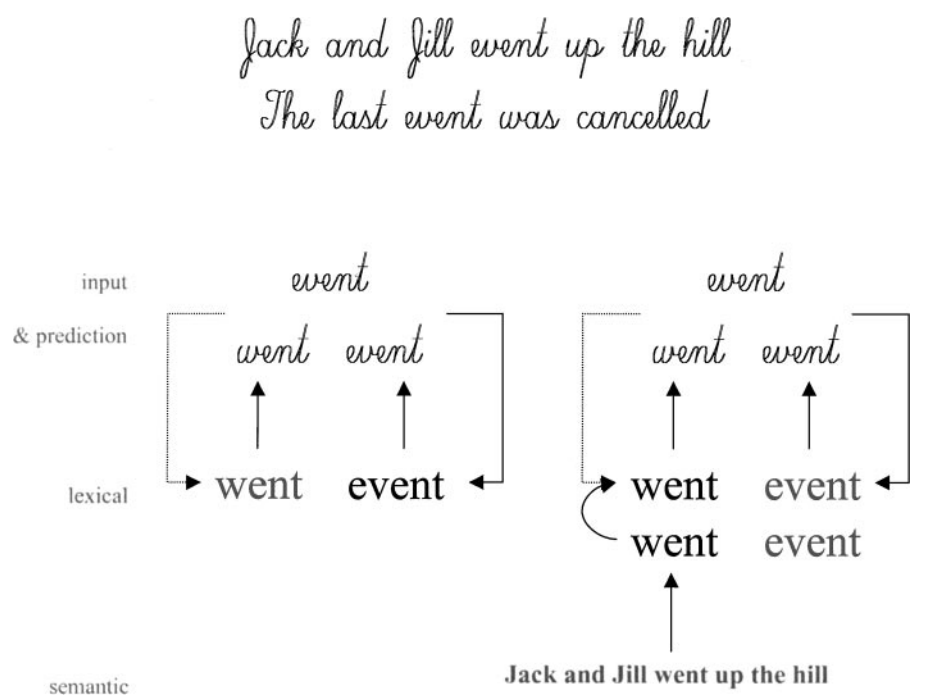


Figure 3 Schematic illustrating the role of priors in biasing toward one representation of an input or another. (*Left*) The word event is selected as the most like cause of the visual input. (*Right*) The word went is selected as the most likely word that is (*a*) a reasonable explanation for the sensory input and (*b*) conforms to prior expectations based on semantic context.

"event" as "went." However, 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 the most likely word. However, within a hierarchical scheme, semantic context can provide top-down predictions, to which the posterior estimate is accountable. When this prior strongly biases in favor of "went," we tolerate a small error at a lower level of visual analysis to minimize the overall prediction error at both 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 toward the most likely cause of the input. Note that predictions at higher levels do not arise by magic. They are themselves 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.

The hierarchical structure of the real world literally comes to be "reflected" by the hierarchical architectures trying to minimize prediction error, not just at the

level of sensory input but at all levels of the hierarchy (again notice the deliberate mirror symmetry in Figure 2). The nice thing about this architecture is that the dynamics of casual representations at the *i*th 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, for which one only needs to know the immediately supraordinate causes to determine the evolution of causes at any level in question. 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 minimization have only to run forward from one level to the next (see Box 2).

In summary the predictive coding approach lends itself naturally to a hierarchical Bayesian treatment, which considers the brain as an inferential device. This perspective arises because the dynamics of the units or populations are driven to minimize error at all levels and implicitly render themselves posterior estimates of the causes given the data. They can do this even with data generated by hierarchies of highly nonlinear dynamical systems. Unlike information theoretic approaches they do not require strong constraints to be built into the architecture; these constraints emerge spontaneously as priors from higher levels. The implicit Bayesian estimation can be formalized from a number of different perspectives. Rao & Ballard (1998) give an extremely nice example using the Kalman filter.

Box 2 Hierarchical Bayes in the Brain

The biological plausibility of the scheme depicted in Figure 2 can be established very simply. Consider any level *i* in a cortical hierarchy containing units (neurons or neuronal populations), whose activity v_i is being predicted by equivalent units in the level above v_{i+1}. The prediction error being reflected in the activities of units is denoted by ε_i. Assuming the simplest generative model possible

$$x_i = F_\theta^i(v_{i+1}) = \theta_i v_{i+1}$$
$$\varepsilon_i = v_i - x_i,$$

where θ_i are backwards connection strengths; we require units in the higher level v_{i+1} to maximize the probability of v_{i+1} given v_i. Assuming the errors have a Gaussian distribution with variance λ_i (i.e., $G_i(\varepsilon) = \varepsilon_i^T \varepsilon_i / \lambda_i$) we have

$$p(v_{i+1}|v_i) \propto p(v_i|v_{i+1})p(v_{i+1}) \propto \exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i^T \varepsilon_i}{\lambda_i} + \frac{\varepsilon_{i+1}^T \varepsilon_{i+1}}{\lambda_{i+1}}\right)\right).$$

Both the dynamics of v_{i+1} and the connection strengths perform a gradient ascent on the log of this probability.

$$\dot{v}_{i+1} = \frac{\partial \ln p(v_{i+1}|v_i)}{\partial v_{i+1}} = \lambda_i^{-1}\theta_i^T\varepsilon_i - \lambda_{i+1}^{-1}\varepsilon_{i+1}$$

$$\dot{\theta}_i = \frac{\partial \ln p(v_{i+1}|v_i)}{\partial \theta_i} = \lambda_i^{-1}\varepsilon_i v_{i+1}^T$$

Despite the complicated nature of the hierarchical generative model and the abstract theorizing, three simple and biologically plausible things emerge: (*a*) The forwards and backwards connections are exactly the same, consistent with the reciprocity of anatomical connections; (*b*) changes in connection strengths reduce to simple Hebbian or associative plasticity; and (*c*) 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 a prior term conveyed by error units in the same level.

GENERATIVE MODELS AND THE BRAIN

The arguments in the preceding section clearly favor predictive coding over information theoretic frameworks as a more plausible account of functional brain architectures. However, it should be noted that the differences between them have been deliberately emphasized. For example, predictive coding and the implicit error minimization results in the maximization of information transfer. In other words, predictive coding conforms to the principle of maximum information transfer, but it does so in a very distinct way (see Olshausen & Field 1996 for a nice integration of predictive and sparse coding). 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 & 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.

Predictive coding, especially in its hierarchical formulation, also conforms to the same parallel and distributed processing principles that underpin connectionist schema (Rumelhart & McClelland 1986). The representation of any cause depends on 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 parallel and distributed processing phenomena (e.g., lateral interactions leading to competition among representations) may be observed. However, in predictive coding, these dynamics are driven explicitly by error minimization, whereas in connectionist simulations the activity is determined solely by the connection strengths that are established during training.

In addition to the theoretical bias toward generative models and predictive coding, the clear emphasis on backward and reentrant dynamics makes it a more natural framework for understanding neuronal infrastructures. Figure 1 shows the fundamental difference between infomax and generative schemes. In infomax schemes the connections are universally forward. In the predictive coding scheme the forward connections (broken line) drive the casual representations to minimize error, whereas backward connections (solid lines) use these representations to emulate mixing enacted by the real world. The nonlinear aspects of this mixing imply that backward connections are modulatory in predictive coding, whereas the nonlinear *un*mixing in infomax schemes is mediated by forward connections. The section on functional specialization and integration assembled some of the anatomical and physiological evidence that backward connections are prevalent in the real brain and can 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 neurons.

Context, Causes, and Representations

The Bayesian perspective suggests something quite profound for the classical view of receptive fields. If neuronal responses encompass both 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. In other words, when a neuron or population is predicted by top-down inputs, it will be much easier to drive than when it is not. Consider the example in Figure 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.

Neuronal Responses and Representations

Classical models (e.g., classical receptive fields) assume that evoked responses will be invariably expressed in the same units or neuronal populations irrespective of the context. However, real neuronal responses are not invariant but depend on 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 nonclassical receptive field effects modeled in Rao & Ballard 1999). Another example

is attentional modulation of evoked responses that can change the sensitivity of neurons to different perceptual attributes (e.g., Treue & Maunsell 1996). 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 presynaptic inputs (e.g., Abbot et al. 1997). In brief, the underlying connection strengths, which define what that unit represents, are a strong function of the immediately preceding neuronal transient (i.e., preceding representation).

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 & 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 suggest that lateral and backward interactions may convey contextual information that shapes the responses of any neuron to its inputs (e.g., Kay & Phillips 1996, Phillips & 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 characterizations of effective connectivity using neuroimaging.

AN EXAMPLE FROM ELECTROPHYSIOLOGY In the next section we will illustrate the context-sensitive nature of cortical activations, and implicit specialization, in the infero-temporal (IT) lobe using neuroimaging. Here we consider the evidence for contextual representations in terms of single-cell responses, to visual stimuli, in the inferior 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 dynamically 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 categorized 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, categorizing 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.

The preceding arguments have been based largely on electrophysiological responses. They can be extended to the population responses elicited in functional neuroimaging where functional specialization (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 specialization is conferred in a context-sensitive fashion by backward connections from higher brain areas.

FUNCTIONAL ARCHITECTURES ASSESSED WITH BRAIN IMAGING

Information theory and predictive coding schema posit alternative architectures that the brain might adopt for perceptual synthesis. The former relies on forward connections, whereas the latter suggests that most of the brain's infrastructure would be used to predict the 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 in one level was partially predicted by activity in a higher level would be sufficient to confirm the existence of backward influences. The problem is that this statistical dependency does not permit any causal inference. Statistical dependencies could easily arise in purely feedforward 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 lesions. 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 modeling techniques that can address this problem, we take a simpler approach and note that interactions between bottom-up and top-down influences cannot be explained by a 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 on the context established by top-down inputs.

In summary, a critical feature of the functional architectures implied by predictive coding is the expression of interactions between bottom-up and top-down influences from other brain regions at a unit, population, or cortical level. The remainder of this article focuses on the evidence for these interactions. From the point of view of functionally specialized responses these interactions manifest as context-sensitive or contextual specialization, where modality-, category-, or exemplar-specific responses, driven by bottom-up inputs are modulated by top-down influences induced by perceptual set. The first half of this section adopts this perspective. 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 involve eliciting changes in perceptual or cognitive (attentional) set.

Context-Sensitive Specialization

If the contextual nature of specialization is mediated by backward 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 following example shows that this is indeed possible.

PSYCHOPHYSIOLOGICAL INTERACTIONS Psychophysiological interactions speak directly to the interactions between bottom-up and top-down influences, where one is modeled as an experimental factor and the other constitutes a measured brain response. An analysis of psychophysiological interactions tries 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 or lateral 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 attribute 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 selective responses in the target area. 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). This class of model speaks directly to functional specialization of a nonlinear and contextual sort. Figure 4 illustrates a specific example (see Dolan et al. 1997 for details). Subjects were asked to view (degraded) faces and nonface (object) controls. The interaction between activity in the parietal region and the presence of faces was expressed most significantly in the right IT region. Changes in parietal activity were induced experimentally by pre-exposure of the (undegraded) stimuli before some scans but not others to

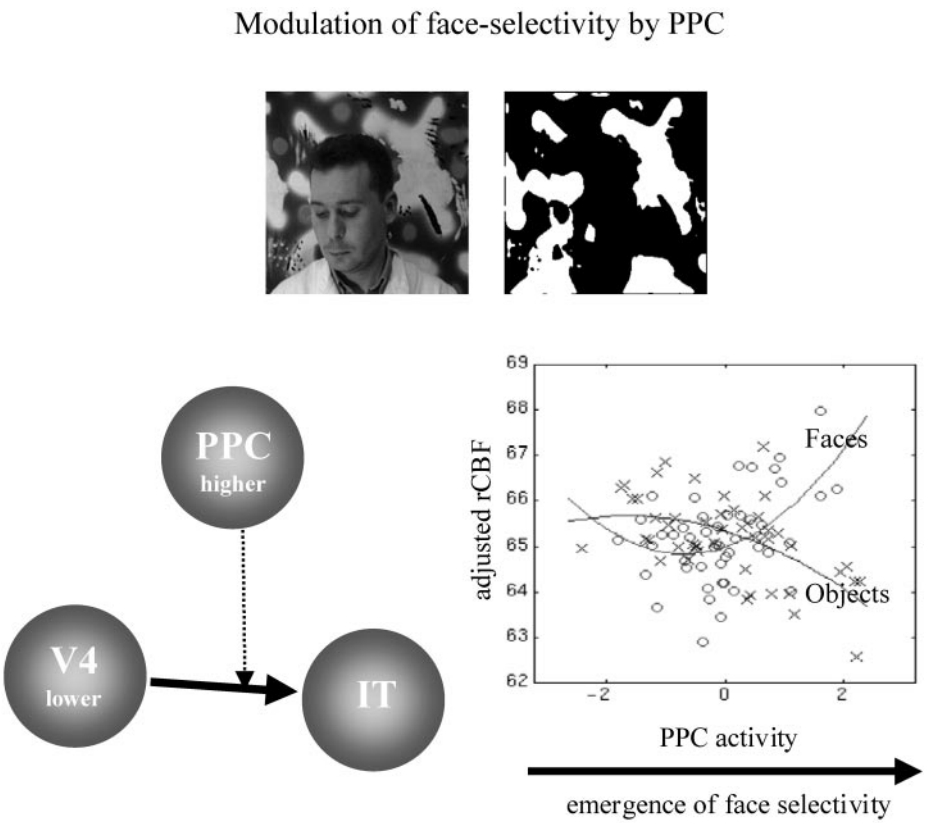


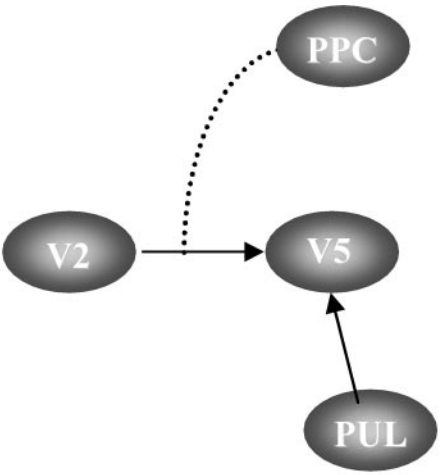
Figure 4 (*Top*) Examples of the stimuli presented to subjects. During the measurement of brain responses only degraded stimuli where shown (e.g., *the right-hand picture*). In half the scans the subject was given the underlying cause of these stimuli through presentation of the original picture (e.g., *left*) before scanning. This priming induced a profound difference in perceptual set for the primed, relative to nonprimed, stimuli. (*Right*) Activity observed in a right infero-temporal (IT) 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 while viewing nonface stimuli and the circles to faces. The essence of this effect can be seen by noting that this region differentiates between faces and nonfaces 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 responses in IT regions subject to permissive modulation by PPC projections.

prime them. The data in the right panel of Figure 4 suggest that the IT region shows face-specific responses, relative to nonface objects, when, and only when, parietal activity is high. These results can be interpreted as a priming-dependent face-specific response, in IT regions that are mediated by interactions with medial parietal cortex. This is a clear example of contextual specialization that depends on top-down nonlinear effects.

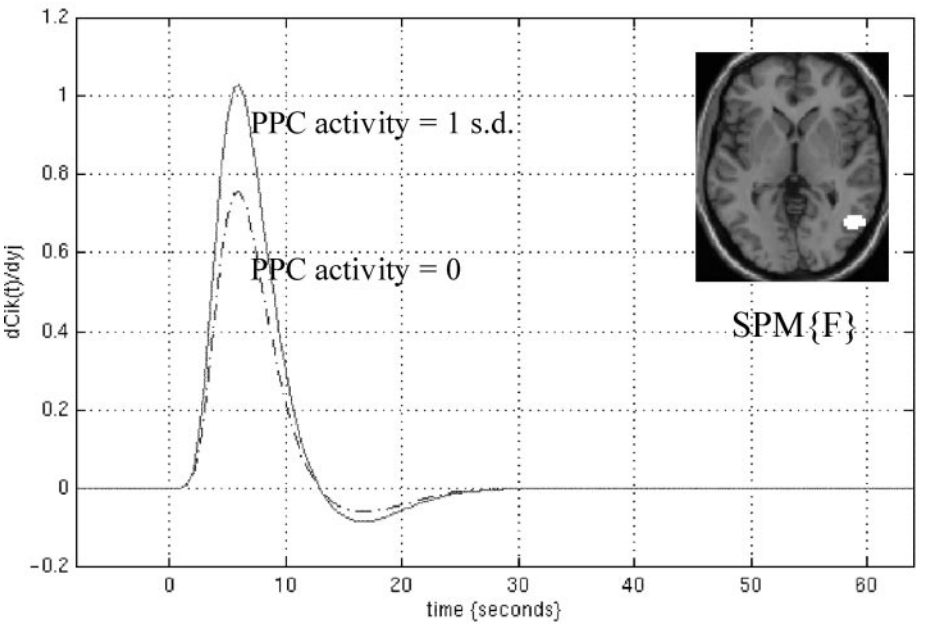
Nonlinear Coupling Among Brain Areas

The previous example, demonstrating contextual specialization, is consistent with functional architectures implied by predictive coding. However, it does not provide definitive evidence for an interaction between top-down and bottom-up influences. In this subsection we look for direct evidence using functional imaging. This rests on being able to measure effective connectivity in a way that is sensitive to interactions among inputs. 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 (see Box 1 and the second section). These interactions can be construed as a context- or activity-dependent modulation of the influence that one region exerts over another. These nonlinearities can also be introduced into structural equation modeling using so-called moderator variables that represent the interaction between two regions when causing activity in a third (Büchel & Friston 1997). From a dynamical point of view, modulatory effects are modeled by second-order kernels. Within these models the influence of one region on another has two components: (*a*) the direct or driving influence of input from the first (e.g., hierarchically lower) region, irrespective of the activities elsewhere and (*b*) an activity-dependent, modulatory component that represents an interaction with inputs from the remaining (e.g., hierarchically higher) regions. These are mediated by the first- and second-order kernels respectively. The example provided in Figure 5 addresses the modulation of visual cortical responses by attentional mechanisms (e.g., Treue & Maunsell 1996) and the mediating role of activity-dependent changes in effective connectivity.

The bottom panel in Figure 5 shows a characterization of this modulatory effect in terms of the increase in V5 responses, to a simulated V2 input, when posterior parietal activity is zero (broken line) and when it is high (solid lines). In this study subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) while manipulating the attentional component of the task (detection of velocity changes). The brain regions and connections comprising the model are shown in the upper panel. The lower panel shows a characterization 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



Changes in V5 responses to inputs from V2 with PPC activity



evidenced a modulatory effect ($p < 0.05$ uncorrected). These voxels were identified by thresholding statistical parametric maps of the F statistic (Friston et al. 1995) 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 & 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. V5 cortical responses evidence an interaction between bottom-up input from early visual cortex and top-down influences from parietal cortex.

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 specialization. The arguments in this review were developed under generative models of brain function, where higher-level systems provide a prediction of the inputs to lower-level regions. Conflict between the two is resolved by changes in the higher-level representations, which are driven by the ensuing error in lower regions, until the mismatch is "cancelled." From this perspective the specialization of any region is determined both by bottom-up driving inputs and by top-down predictions. Specialization is therefore not an intrinsic property of any region but depends on both forward and backward connections with other areas. Because the latter have access to the context in which the inputs are generated, they are in a position to modulate the selectivity or specialization of lower areas.

←

Figure 5 (*Top*) Brain regions and connections comprising the model. (*Bottom*) Characterization 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. 1995) of the F statistic testing for the contribution of second-order kernels involving V2 and PPC (treating all other terms as nuisance variables). The data were obtained with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) while manipulating the attentional component of the task (detection of velocity changes).

The implications for classical models (e.g., classical receptive fields in electrophysiology, classical specialization 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 generative models, not only accounts for many extra-classical phenomena seen empirically but enforces a view of the brain as an inference machine, through its Bayesian motivation.

ACKNOWLEDGMENT

The Wellcome Trust funded this work.

The *Annual Review of Neuroscience* is online at http://neuro.annualreviews.org

LITERATURE CITED

Abbot LF, Varela JA, Sen K, Nelson SB. 1997. Synaptic depression and cortical gain control. *Science* 275:220–23

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

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

Atick JJ, Redlich AN. 1990. Towards a theory of early visual processing. *Neural Comput.* 2:308–20

Barlow HB. 1961. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, ed. WA Rosenblith. Cambridge: MIT

Bell AJ, Sejnowski TJ. 1995. An information maximisation approach to blind separation and blind de-convolution. *Neural Comput.* 7:1129–59

Bendat JS. 1990. *Nonlinear System Analysis and Identification from Random Data*. New York: Wiley

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

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

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

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

Dolan RJ, Fink GR, Rolls E, Booth M, Holmes A, et al. 1997. How the brain learns to see objects and faces in an impoverished context. *Nature* 389:596–98

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

Edelman GM. 1993. Neural Darwinism: selection and reentrant signalling in higher brain function. *Neuron* 10:115–25

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

Fliess M, Lamnabhi M, Lamnabhi-Lagarrigue F. 1983. An algebraic approach to nonlinear functional expansions. *IEEE Trans. Circuits Syst.* 30:554–70

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

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

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

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

Friston KJ, Büchel C, Fink GR, Morris J, Rolls E, Dolan RJ. 1997. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* 6:218–29

Friston KJ, Frith CD, Frackowiak RSJ. 1993. Principal component analysis learning algorithms: a neurobiological analysis. *Proc. R. Soc. B* 254:47–54

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ. 1995. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2:189–210

Friston KJ, Tononi G, Reeke GH, Sporns O, Edelman GE. 1994. Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* 39:229–43

Gawne TJ, Richmond BJ. 1993. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* 13:2758–71

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

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

Hinton GE, Dayan P, Frey BJ, Neal RM. 1995. The "Wake-Sleep" algorithm for unsupervised neural networks. *Science* 268:1158–61

Hirsch JA, Gilbert CD. 1991. Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.* 11:1800–9

Karhunen J, Joutsensalo J. 1994. Representation and separation of signals using nonlinear PCA type learning. *Neural Netw.* 7:113–27

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

Linsker R. 1988. Self-organisation in a perceptual network. *Computer* March:105–17

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

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

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

Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–9

Optican L, Richmond BJ. 1987. Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II. Information theoretic analysis. *J. Neurophysiol.* 57:132–46

Pearl J. 2000. *Causality, Models, Reasoning and Inference*. Cambridge, UK: Cambridge Univ. Press

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

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

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

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

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

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

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

Sugase Y, Yamane S, Ueno S, Kawano K. 1999. Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400:869–73

Sutton RS, Barto AG. 1990. Time derivative models of Pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, ed. M Gabriel, J Moore, pp. 497–538. Cambridge: MIT

Tononi G, Sporns O, Edelman GM. 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci.* 91: 5033–37

Tovee MJ, Rolls ET, Treves A, Bellis RP. 1993. Information encoding and the response of single neurons in the primate temporal visual cortex. *J. Neurophysiol.* 70:640–54

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

Wray J, Green GGR. 1994. Calculation of the Volterra kernels of non-linear dynamic systems using an artificial neuronal network. *Biol. Cybern.* 71:187–95

Zeki S. 1990. The motion pathways of the visual cortex. In *Vision: Coding and Efficiency*, ed. C Blakemore, pp. 321–45. Cambridge, UK: Cambridge Univ. Press

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

Annual Review of Neuroscience
Volume 25, 2002

CONTENTS