PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 1211–1221
doi:10.1098/rstb.2008.0300

Predictive coding under the free-energy principle

Karl Friston* **and Stefan Kiebel**

The Wellcome Trust Centre of Neuroimaging, Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK

This paper considers prediction and perceptual categorization as an inference problem that is solved by the brain. We assume that the brain models the world as a hierarchy or cascade of dynamical systems that encode causal structure in the sensorium. Perception is equated with the optimization or inversion of these internal models, to explain sensory data. Given a model of how sensory data are generated, we can invoke a generic approach to model inversion, based on a free energy bound on the model's evidence. The ensuing free-energy formulation furnishes equations that prescribe the process of recognition, i.e. the dynamics of neuronal activity that represent the causes of sensory input. Here, we focus on a very general model, whose hierarchical and dynamical structure enables simulated brains to recognize and predict trajectories or sequences of sensory states. We first review hierarchical dynamical models and their inversion. We then show that the brain has the necessary infrastructure to implement this inversion and illustrate this point using synthetic birds that can recognize and categorize birdsongs.

Keywords: generative models; predictive coding; hierarchical; birdsong

1. INTRODUCTION

This paper reviews generic models of our sensorium and a Bayesian scheme for their inversion. We then show that the brain has the necessary anatomical and physiological equipment to invert these models, given sensory data. Critically, the scheme lends itself to a relatively simple neural network implementation that shares many features with real cortical hierarchies in the brain. The basic idea that the brain tries to infer the causes of sensations dates back to Helmholtz (e.g. Helmholtz 1860/1962; Barlow 1961; Neisser 1967; Ballard *et al*. 1983; Mumford 1992; Kawato *et al*. 1993; Dayan *et al*. 1995; Rao & Ballard 1998), with a recent emphasis on hierarchical inference and *empirical Bayes* (Friston 2003, 2005; Friston *et al*. 2006). Here, we generalize this idea to cover dynamics in the world and consider how neural networks could be configured to invert hierarchical dynamical models and deconvolve sensory causes from sensory input.

This paper comprises four sections. In §1, we introduce hierarchical dynamical models and their inversion. These models cover most of the models encountered in the statistical literature. An important aspect of these models is their formulation in generalized coordinates of motion, which lends them a hierarchal form in both structure and dynamics. These hierarchies induce empirical priors that provide structural and dynamical constraints, which can be exploited during inversion. In §2, we show how inversion can be formulated as a simple gradient ascent using neuronal networks; in §3, we consider how evoked brain responses might be understood in terms of inference under hierarchical dynamical models of sensory input.[1]

* Author for correspondence (k.friston@fil.ion.ucl.ac.uk).

One contribution of 18 to a Theme Issue 'Predictions in the brain: using our past to prepare for the future'.

2. HIERARCHICAL DYNAMICAL MODELS

In this section, we look at dynamical generative models $p(y,\vartheta) = p(y\,|\,\vartheta)p(\vartheta)$ that entail a likelihood, $p(y|\vartheta)$, of getting some data, y, given some causes, $\vartheta = \{x, v, \theta\}$, and priors on those causes, $p(\vartheta)$. The sorts of models we consider have the following form:

$$\begin{aligned} y &= g(x, v, \theta) + z, \\ \dot{x} &= f(x, v, \theta) + w, \end{aligned} \tag{2.1}$$

where the nonlinear functions f and g of the states are parametrized by θ. The states $v(t)$ can be deterministic, stochastic or both, and are variously referred to as inputs, sources or causes. The states $x(t)$ meditate the influence of the input on the output and endow the system with memory. They are often referred to as hidden states because they are seldom observed directly. We assume that the stochastic innovations (i.e. observation noise) $z(t)$ are analytic, such that the covariance of $\tilde{z} = [z, z', z'', \ldots]^{\mathrm{T}}$ is well defined; similarly, for $w(t)$, which represents random fluctuations on the motion of hidden states. Under local linearity assumptions, the generalized motion of the output or response $\tilde{y} = [y, y', y'', \ldots]^{\mathrm{T}}$ is given by

$$\begin{aligned} y &= g(x, v) + z & x' &= f(x, v) + w \\ y' &= g_x x' + g_v v' + z' & x'' &= f_x x' + f_v v' + w' \\ y'' &= g_x x'' + g_v v'' + z'' & x''' &= f_x x'' + f_v v'' + w'' \\ &\vdots & &\vdots \end{aligned} \tag{2.2}$$

The first (observer) equation shows that the generalized states $u = [\tilde{v}, \tilde{x},]^{\mathrm{T}}$ are needed to generate a generalized response or trajectory. The second (state) equations enforce a coupling between different orders of the motion of the hidden states and confer memory on the system. We can write these equations compactly as

$$\begin{aligned}\tilde{y} &= \tilde{g} + \tilde{z},\\ D\tilde{x} &= \tilde{f} + \tilde{w},\end{aligned} \tag{2.3}$$

where the predicted response $\tilde{g} = [g, g', g'', \ldots]^{\mathrm{T}}$ and motion $\tilde{f} = [f, f', f'', \ldots]^{\mathrm{T}}$ in the absence of random fluctuations are

$$\begin{aligned} g &= g(x, v) & f &= f(x, v)\\ g' &= g_x x' + g_v v' & f' &= f_x x' + f_v v'\\ g'' &= g_x x'' + g_v v'' & f'' &= f_x x'' + f_v v''\\ &\vdots & &\vdots \end{aligned} \tag{2.4}$$

and D is a block-matrix derivative operator, whose first leading diagonal contains identity matrices. Gaussian assumptions about the fluctuations provide the likelihood and furnish empirical priors $p(\tilde{x}\,|\,\tilde{v})$ on the motion of hidden states

$$\begin{aligned} p(\tilde{y}, \tilde{x}, \tilde{v}) &= p(\tilde{y}\,|\,\tilde{x}, \tilde{v})p(\tilde{x}, \tilde{v})\\ p(\tilde{y}\,|\,\tilde{x}, \tilde{v}) &= N(\tilde{y} : \tilde{g}, \tilde{\Sigma}^{z})\\ p(\tilde{x}, \tilde{v}) &= p(\tilde{x}\,|\,\tilde{v})p(\tilde{v})\\ p(\tilde{x}\,|\,\tilde{v}) &= N(D\tilde{x} : \tilde{f}, \tilde{\Sigma}^{w})\\ p(\tilde{v}) &= N(\tilde{v} : \tilde{\eta}, \tilde{\Sigma}^{v}). \end{aligned} \tag{2.5}$$

Here, we have assumed Gaussian priors $p(\tilde{v})$ on the generalized causes, with mean, $\tilde{\eta}$, and covariance, $\tilde{\Sigma}^{v}$. The factorization in equation (2.5) is important because one can appeal to empirical Bayes to interpret the conditional dependences that are induced. In empirical Bayes (Efron & Morris 1973), factorizations of the prior density create empirical priors that share properties of both the likelihood and priors. For example, the density on the hidden states $p(\tilde{x}\,|\,\tilde{v})$ is part of the prior on quantities needed to evaluate the likelihood. However, it is also a likelihood of the hidden states, given the causes or inputs. This renders it an empirical prior. It is these constraints that can be exploited by the brain and are accessed through plausible assumptions about noise. These assumptions are encoded by their covariances $\tilde{\Sigma}^{z}$ and $\tilde{\Sigma}^{w}$ or inverses $\tilde{\Pi}^{z}$ and $\tilde{\Pi}^{w}$ (known as precisions). Generally, these covariances factorize; $\tilde{\Sigma}^{i} = \Sigma^{i} \otimes R^{i}$ into a covariance matrix and a matrix of correlations R^{i} among generalized states that encode their autocorrelations or smoothness.

(a) *Hierarchical forms*

Hierarchical dynamical models generalize the $(m = 1)$ model given in §1,

$$\begin{aligned} y &= g(x^{(1)}, v^{(1)}) + z^{(1)}\\ \dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + w^{(1)}\\ &\vdots\\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + z^{(i)}\\ \dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + w^{(i)}\\ &\vdots\\ v^{(m)} &= \eta + z^{(m+1)}. \end{aligned} \tag{2.6}$$

Again, $f^{(i)} = f(x^{(i)}, v^{(i)})$ and $g^{(i)} = g(x^{(i)}, v^{(i)})$ are functions of the states. The conditionally independent fluctuations $z^{(i)}$ and $w^{(i)}$ play the role of observation noise at the first level and induce random fluctuations in the states at higher levels. The causes $v = [v^{(1)}, \ldots, v^{(m)}]^{\mathrm{T}}$ link levels, whereas the hidden states $x = [x^{(1)}, \ldots, x^{(m)}]^{\mathrm{T}}$ link dynamics over time. In a hierarchical form, the output of one level acts as an input to the next. Inputs from higher levels can enter nonlinearly into the state equations and can be regarded as changing its control parameters to produce quite complicated generalized convolutions with deep (i.e. hierarchical) structure.

(b) *Model inversion and variational Bayes*

We now consider how these models are inverted (for details, see Friston 2008). A very general approach is based on variational Bayes, which approximates the conditional density $p(\vartheta\,|\,y, m)$ on the causes, ϑ, given a model m and data y. This is achieved by optimizing the sufficient statistics of a recognition density $q(\vartheta)$ with respect to a lower bound on the evidence $p(y\,|\,m)$ of the model itself (Feynman 1972; Hinton & von Camp 1993; MacKay 1995; Neal & Hinton 1998; Friston *et al.* 2007). In this paper, we assume that the parameters are known and focus on the states; $q(\vartheta) = q(u(t))$, where $u = [\tilde{v}, \tilde{x},]^{\mathrm{T}}$. To further simplify, we assume that the brain uses something called the Laplace approximation. This enables us to focus on a single quantity for each unknown state, the conditional mean; under the Laplace approximation, the conditional density assumes a fixed Gaussian form $q(u(t)) = N(u : \tilde{\mu}, C)$ with sufficient statistics $\tilde{\mu}$ and C, corresponding to the conditional mean and covariance of the unknown states. A key advantage of the Laplace assumption is that the conditional precision is a function of the mean, which means we can focus on optimizing the mean (precision is the inverse covariance).

In static systems, the mean maximizes the energy, $U(t) = \ln p(\tilde{y}, u)$; this is the solution to a gradient ascent scheme, $\dot{\tilde{\mu}} = U(t)_u$. In dynamical systems, the trajectory of the conditional mean maximizes the path integral of energy (called action), which is the solution to the ansatz

$$\dot{\tilde{\mu}} - D\tilde{\mu} = U(t)_u. \tag{2.7}$$

Here, $\dot{\tilde{\mu}} - D\tilde{\mu}$ can be regarded as motion in a frame of reference that moves along the trajectory encoded in generalized coordinates. Critically, the stationary solution, in this moving frame of reference, maximizes the action. This may sound a little complicated but is simply a version of Hamilton's principle of stationary action, which allows the conditional mean in equation (2.7) to converge on a 'moving target'. At this point, the path of the mean becomes the mean of the path and $\dot{\tilde{\mu}} - D\tilde{\mu} = 0$.

(c) *Summary*

In this section, we have introduced hierarchical dynamical models in generalized coordinates of motion. These models are as complicated as one could imagine; they comprise causes and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. A key aspect of these models is their hierarchical form, which induces empirical priors on

the causes. These recapitulate the constraints on hidden states, furnished by constraints on their motion.

By assuming a fixed-form (Laplace) approximation to the conditional density, one can reduce model inversion to finding the conditional means of unknown causes. This can be formulated as a gradient ascent in a frame of reference that moves along the path encoded in generalized coordinates. The only thing we need to implement this recognition scheme is the internal energy, $U(t) = \ln p(\tilde{y}, u)$, which is specified by the generative model (equation (2.5)).

3. HIERARCHICAL MODELS IN THE BRAIN

A key architectural principle of the brain is its hierarchical organization (Maunsell & Van Essen 1983; Zeki & Shipp 1988; Felleman & Van Essen 1991). This has been established most thoroughly in the visual system, where lower (primary) areas receive sensory input and higher areas adopt a multimodal or associational role. The neurobiological notion of a hierarchy rests upon the distinction between forward and backward connections (Rockland & Pandya 1979; Murphy & Sillito 1987; Felleman & Van Essen 1991; Sherman & Guillery 1998; Angelucci *et al.* 2002). This distinction is based upon the specificity of cortical layers that are the predominant sources and origins of extrinsic connections. Forward connections arise largely in superficial pyramidal cells, in supragranular layers, and terminate on spiny stellate cells of layer 4 in higher cortical areas (Felleman & Van Essen 1991; DeFelipe *et al.* 2002). Conversely, backward connections arise largely from deep pyramidal cells in infragranular layers and target cells in the infra- and supragranular layers of lower cortical areas. Intrinsic connections mediate lateral interactions between neurons that are a few millimetres away. There is a key functional asymmetry between forward and backward connections, which renders backward connections more modulatory or nonlinear in their effects on neuronal responses (e.g. Sherman & Guillery 1998; see also Hupe *et al.* 1998). This is consistent with the deployment of voltage-sensitive N-methyl-D-aspartic acid receptors in the supragranular layers that are targeted by backward connections (Rosier *et al.* 1993). Typically, the synaptic dynamics of backward connections have slower time constants. This has led to the notion that forward connections are driving and illicit an obligatory response in higher levels, whereas backward connections have both driving and modulatory effects and operate over larger spatial and temporal scales. We now consider how this hierarchical functional architecture can be understood under the inversion of hierarchical models by the brain.

(a) *Perceptual inference*

If we assume that the activity of neurons encodes the conditional mean of states, then equation (2.7) specifies the neuronal dynamics entailed by recognizing states of the world from sensory data. In Friston (2008), we show how these dynamics can be expressed simply in terms of prediction errors,

$$\varepsilon^v = \begin{bmatrix} y \\ v^{(1)} \\ \vdots \\ v^{(m)} \end{bmatrix} - \begin{bmatrix} g^{(1)} \\ g^{(2)} \\ \vdots \\ \eta \end{bmatrix}$$
$$\varepsilon^x = \begin{bmatrix} Dx^{(1)} \\ \vdots \\ Dx^{(m)} \end{bmatrix} - \begin{bmatrix} f^{(1)} \\ \vdots \\ f^{(m)} \end{bmatrix} \qquad \tilde{\varepsilon} = \begin{bmatrix} \tilde{\varepsilon}^v \\ \tilde{\varepsilon}^x \end{bmatrix}. \tag{3.1}$$

Using these errors, we can rewrite equation (2.7) as

$$\dot{\tilde{\mu}} = U(t)_u + D\tilde{\mu} = D\tilde{\mu} - \tilde{\varepsilon}_u^{\mathrm{T}}\xi$$
$$\xi = \tilde{\Pi}\tilde{\varepsilon} = \tilde{\varepsilon} - \Lambda\xi \qquad \tilde{\Pi} = \begin{bmatrix} \tilde{\Pi}^z & \\ & \tilde{\Pi}^w \end{bmatrix}, \tag{3.2}$$

equation (3.2) describes how neuronal states self-organize when exposed to sensory input. Its form is quite revealing and suggests two distinct populations of neurons; *state units* whose activity encodes $\tilde{\mu}(t)$ and *error units* encoding precision-weighted prediction error $\xi = \tilde{\Pi}\tilde{\varepsilon}$, with one error unit for each state. Furthermore, the activities of error units are a function of the states and the dynamics of state units are a function of prediction error. This means that the two populations pass messages to each other and to themselves. The messages passed within the states, $D\tilde{\mu}$, mediate empirical priors on their motion, while $-\Lambda\xi$ optimize the weighting or gain of error units.

(b) *Hierarchical message passing*

If we unpack these equations, we can see the hierarchical nature of this message passing,

$$\begin{aligned} \dot{\tilde{\mu}}^{(i)v} &= D\tilde{\mu}^{(i)v} - \tilde{\varepsilon}_v^{(i)\mathrm{T}}\xi^{(i)} - \xi^{(i+1)v} \\ \dot{\tilde{\mu}}^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{\varepsilon}_x^{(i)\mathrm{T}}\xi^{(i)} \\ \xi^{(i)v} &= \tilde{\mu}^{(i-1)v} - \tilde{g}(\tilde{\mu}^{(i)}) - \Lambda^{(i)z}\xi^{(i)v} \\ \xi^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{f}(\tilde{\mu}^{(i)}) - \Lambda^{(i)w}\xi^{(i)x}. \end{aligned} \tag{3.3}$$

This shows that error units receive messages from the states in the same level and the level above, whereas states are driven by error units in the same level and the level below (figure 1). Critically, inference requires only the prediction error from the lower level $\xi^{(i)}$ and the level in question, $\xi^{(i+1)}$. These provide bottom-up and lateral messages that drive conditional expectations $\tilde{\mu}^{(i)}$ towards a better prediction, to explain the prediction error in the level below. These top-down and lateral predictions correspond to $\tilde{g}^{(i)}$ and $\tilde{f}(i)$. This is the essence of recurrent message passing between hierarchical levels to optimize free energy or suppress prediction error, i.e. recognition dynamics. In summary, all connections between error and state units are reciprocal, where the only connections that link levels are forward connections conveying prediction error to state units and reciprocal backward connections that mediate predictions.

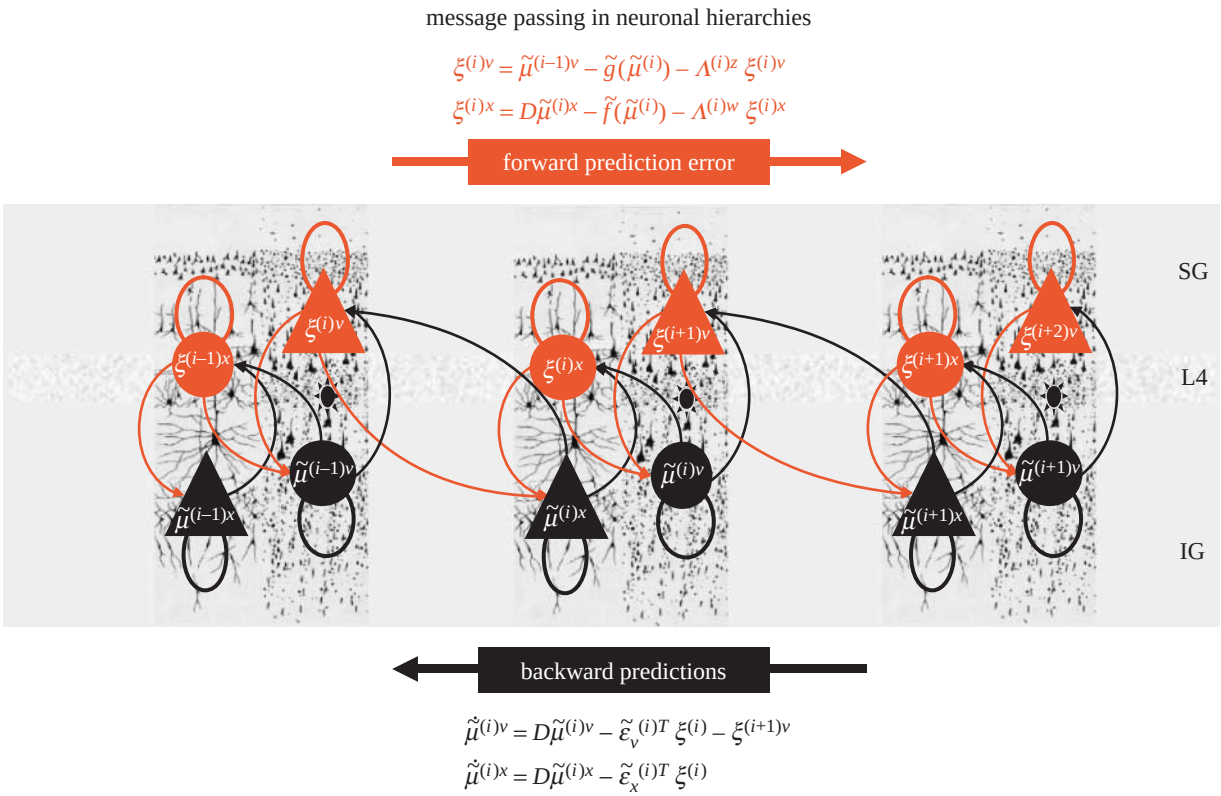

Figure 1. Schematic detailing the neuronal architectures that encode an ensemble density on the states of a hierarchical model. This schematic shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that are used to construct predictions. These predictions try to explain the input from lower areas by suppressing prediction error. In this scheme, the sources of forward connections are the superficial pyramidal cell population and the sources of backward connections are the deep pyramidal cell population. The differential equations relate to the optimization scheme detailed in the main text. Within each area, the cells are shown in relation to the laminar structure of the cortex that includes supragranular (SG), granular (L4) and infragranular (IG) layers.

We can identify error units with superficial pyramidal cells because the only messages that pass up the hierarchy are prediction errors, and superficial pyramidal cells originate forward connections in the brain. This is useful because it is these cells that are primarily responsible for electroencephalographic signals that can be measured non-invasively. Similarly, the only messages that are passed down the hierarchy are the predictions from state units. The sources of extrinsic backward connections are the deep pyramidal cells, and one might deduce that these encode the expected causes of sensory states (see Mumford 1992; Raizada & Grossberg 2003; figure 1). Critically, the motion of each state unit is a linear mixture of bottom-up prediction error (equation (3.3)). This is exactly what is observed physiologically; bottom-up driving inputs elicit obligatory responses that do not depend on other bottom-up inputs. The prediction error itself is formed by predictions conveyed by backward and lateral connections. These influences embody the nonlinearities implicit in $\tilde{g}^{(i)}$ and $\tilde{f}(i)$. Again, this is entirely consistent with the modulatory characteristics of backward connections.

(c) *Summary*

We have seen how the inversion of a generic hierarchical and dynamical model of sensory inputs can be transcribed onto neuronal quantities that optimize a bound on the evidence for that model. Under some simplifying assumptions, this optimization corresponds to the suppression of prediction error at all levels in a cortical hierarchy. This suppression rests upon a balance between bottom-up (prediction error) and top-down (empirical prior) influences. In §3, we use this scheme to simulate neuronal responses. Specifically, we consider the electrophysiological correlates of prediction error and ask whether we can understand some common phenomena in event-related potential (ERP) research in terms of message passing in the brain.

4. BIRDSONG AND ATTRACTORS

In this section, we examine the emergent properties of a system that uses hierarchical dynamics or attractors as generative models of sensory input. The example we use is birdsong and the empirical measures we focus on are local field potentials (LFP) or evoked (ERP) responses that can be recorded non-invasively. Our aim is to show that canonical features of empirical electrophysiological responses can be reproduced easily under attractor models of sensory input. We first describe the model of birdsong and demonstrate its dynamics through simulated lesion experiments. We then use simplified versions to show how attractors can be used to categorize sequences of stimuli quickly and efficiently. Throughout this section, we exploit the fact that superficial pyramidal cells are major contributors to observed LFP and ERP signals, which means that we can ascribe these signals to prediction error (figure 1).

(a) *Attractors in the brain*

Here, the basic idea is that the environment unfolds as an ordered sequence of spatio-temporal dynamics, whose equations of motion induce attractor manifolds that contain sensory trajectories. Critically, the shape of the manifold generating sensory data is itself changed by other dynamical systems that could have their own attractors. If we consider the brain has a generative model of these coupled attractors, then we would expect to see attractors in neuronal dynamics that are trying to predict sensory input. In a hierarchical setting, the states of a high-level attractor enter the equations of motion of a low-level attractor in a nonlinear way, to change the shape of its manifold. This form of this generative model has a number of sensible and compelling characteristics.

First, at any level, the model can generate and therefore encode structured sequences of events, as the states flow over different parts of the manifold. These sequences can be simple, such as the quasi-periodic attractors of central pattern generators (McCrea & Rybak 2008), or can exhibit complicated sequences of the sort associated with chaotic and itinerant dynamics (e.g. Haken *et al.* 1990; Friston 1997; Jirsa *et al.* 1998; Kopell *et al.* 2000; Breakspear & Stam 2005; Canolty *et al.* 2006; Rabinovich *et al.* 2008). The notion of attractors as the basis of generative models extends the notion of generalized coordinates, encoding trajectories, to families of trajectories that lie on attractor manifolds, i.e. paths that are contained in the flow field specified by the states of a supraordinate attractor.

Second, hierarchically deployed attractors enable the brain to generate and predict different categories of sequences. This is because any low-level attractor embodies a family of trajectories that correspond to a structured sequence. The neuronal activity encoding the particular state at any time determines where the current dynamics are within the sequence, while the shape of the attractor manifold determines which sequence is currently being expressed. In other words, the attractor manifold encodes *what* is perceived and the neuronal activity encodes *where* the percept is on the manifold or within the sequence.

Third, if the state of a higher attractor changes the manifold of a subordinate attractor, then the states of the higher attractor come to encode the category of the sequence or dynamics represented by the lower attractor. This means that it is possible to generate and represent sequences of sequences and, by induction, sequences of sequences of sequences, etc. This rests upon the states of neuronal attractors providing control parameters for attractor dynamics in the level below. This necessarily entails a nonlinear interaction between the top-down predictions and the states of the recipient attractor. Again, this is entirely consistent with the known nonlinear effects of top-down connections in the real brain.

Finally, this particular model has implications for the temporal structure of perception. In other words, the dynamics of high-level representations unfold more slowly than the dynamics of lower level representations. This is because the state of a higher attractor prescribes a manifold that guides the flow of lower states. In the limiting case, a fixed-point attractor will encode lower level dynamics, which could change quite rapidly. In the following, we see an example of this when considering the perceptual categorization of different sequences of chirps subtending birdsongs. This attribute of hierarchically coupled attractors enables the representation of arbitrarily long sequences of sequences and suggests that neuronal representations in the brain will change more slowly at higher levels (Kiebel *et al.* 2008; see also Botvinick 2007; Hasson *et al.* 2008). One can turn this argument on its head and use the fact that we are able to recognize sequences of sequences (e.g. Chait *et al.* 2007) as an existence proof for this sort of generative model. In the following examples, we try to show how autonomous dynamics furnish generative models of sensory input, which behave like real brains, when measured electrophysiologically.

(b) *A synthetic avian brain*

The toy example used here deals with the generation and recognition of birdsongs (Laje & Mindlin 2002). We imagine that birdsongs are produced by two time-varying control parameters that control the frequency and amplitude of vibrations emanating from the syrinx of a songbird (figure 2). There has been an extensive modelling effort using attractor models at the biomechanical level to understand the generation of birdsong (e.g. Laje *et al.* 2002). Here, we use the attractors at a higher level to provide time-varying control over the resulting sonograms (Fletcher 2000). We drive the syrinx with two states of a Lorenz attractor, one controlling the frequency (between 2 and 5 KHz) and the other controlling the amplitude or volume (after rectification). The parameters of the Lorenz attractor were chosen to generate a short sequence of chirps every second or so. To endow the songs with a hierarchical structure, we placed a second Lorenz attractor, whose dynamics were an order of magnitude slower, over the first. The states of the slower attractor entered as control parameters (the Rayleigh and Prandtl numbers) to control the dynamics of the first. These dynamics could range from a fixed-point attractor, where the states of the first are all zero, through to quasi-periodic and chaotic behaviour, when the value of the Prandtl number exceeds an appropriate threshold (approx. 24) and induces a bifurcation. Because higher states evolve more slowly, they switch the lower attractor on and off, generating distinct songs, where each song comprises a series of distinct chirps (figure 3).

(c) *Song recognition*

This model generates spontaneous sequences of songs using autonomous dynamics. We generated a single song, corresponding roughly to a cycle of the higher attractor and then inverted the ensuing sonogram (summarized as peak amplitude and volume), using the message-passing scheme described in §2. The results are shown in figure 3 and demonstrate that, after several hundreds of milliseconds, the veridical hidden states and supraordinate causes can be recovered. Interestingly, the third chirp is not perceived, suggesting that the first-level prediction error was not sufficient to overcome the dynamical and

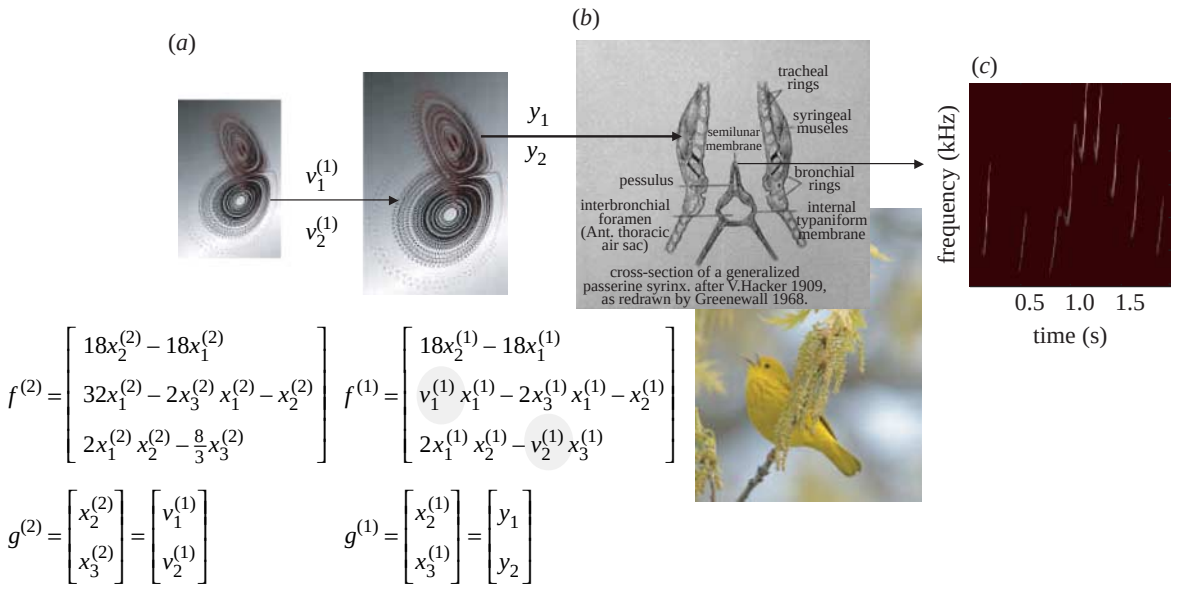

Figure 2. Schematic showing the construction of the generative model for birdsongs. This comprises two Lorenz attractors where the higher attractor delivers two control parameters (grey circles) to a lower level attractor, which, in turn, delivers two control parameters to a synthetic syrinx to produce amplitude- and frequency-modulated stimuli ((*a*) neuronal hierarchy and (*b*) syrinx). This stimulus is represented as a sonogram in (*c*). The equations represent the hierarchical dynamical model in the form of equation (2.6).

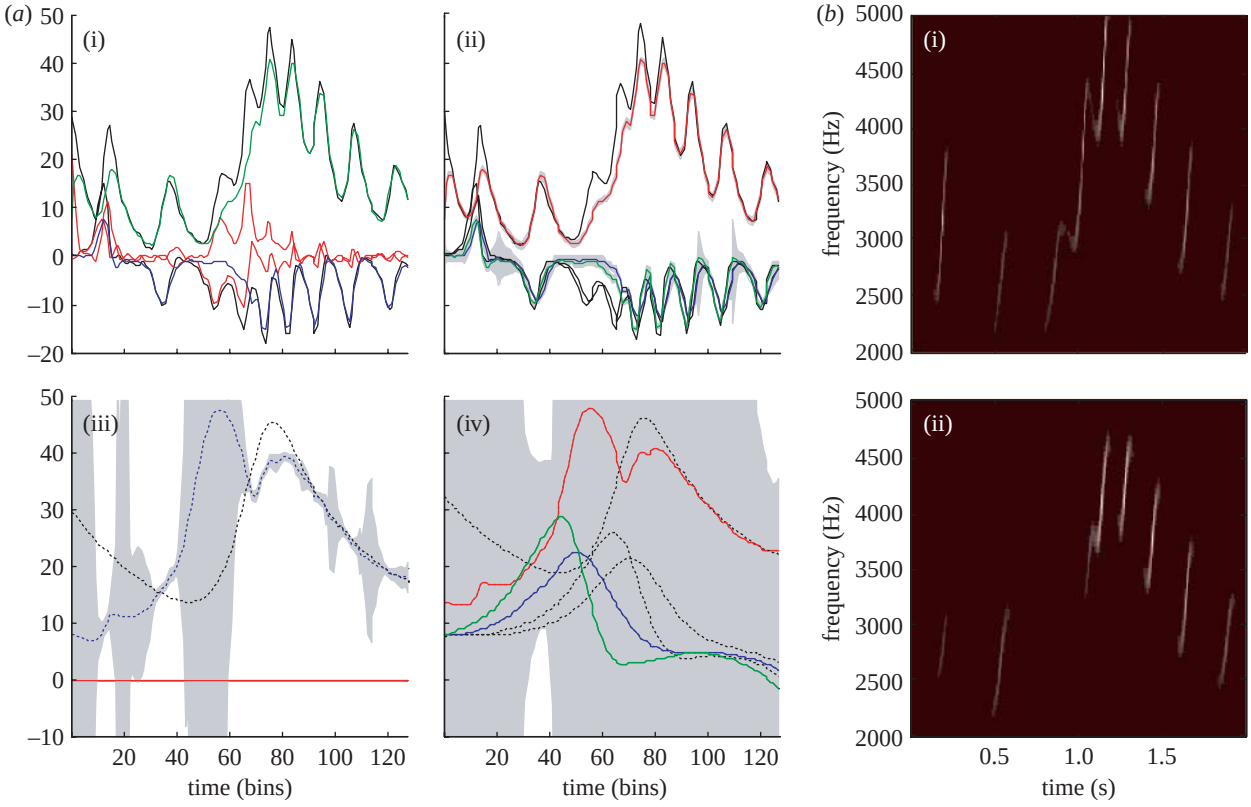

Figure 3. Results of a Bayesian inversion or deconvolution of the sonogram shown in figure 2. (*a*) Song recognition: the time courses of (i) causal and (ii) hidden states. (i) These are the true and predicted states driving the syrinx and are simple mappings from two of the three hidden states of the first-level attractor. The solid lines correspond to the conditional mode and the dotted lines correspond to the true values. The discrepancy is the prediction error and is shown as a red line. (ii) The true and estimated hidden states of the first-level attractor. Note that the third hidden state has to be inferred from the sensory data. Confidence intervals on the conditional expectations are shown in grey and demonstrate a high degree of confidence, because a low level of sensory noise was used in these simulations. Also shown are the corresponding (iii) causes and (iv) hidden states at the second level. Again, the conditional expectations are shown as solid lines and the true values as broken lines. Note the inflated conditional confidence interval halfway through the song when the third and fourth chirps are misperceived. (*b*)(i) The stimulus and (ii) percept in sonogram format, detailing the expression of different frequencies generated over peristimulus time.

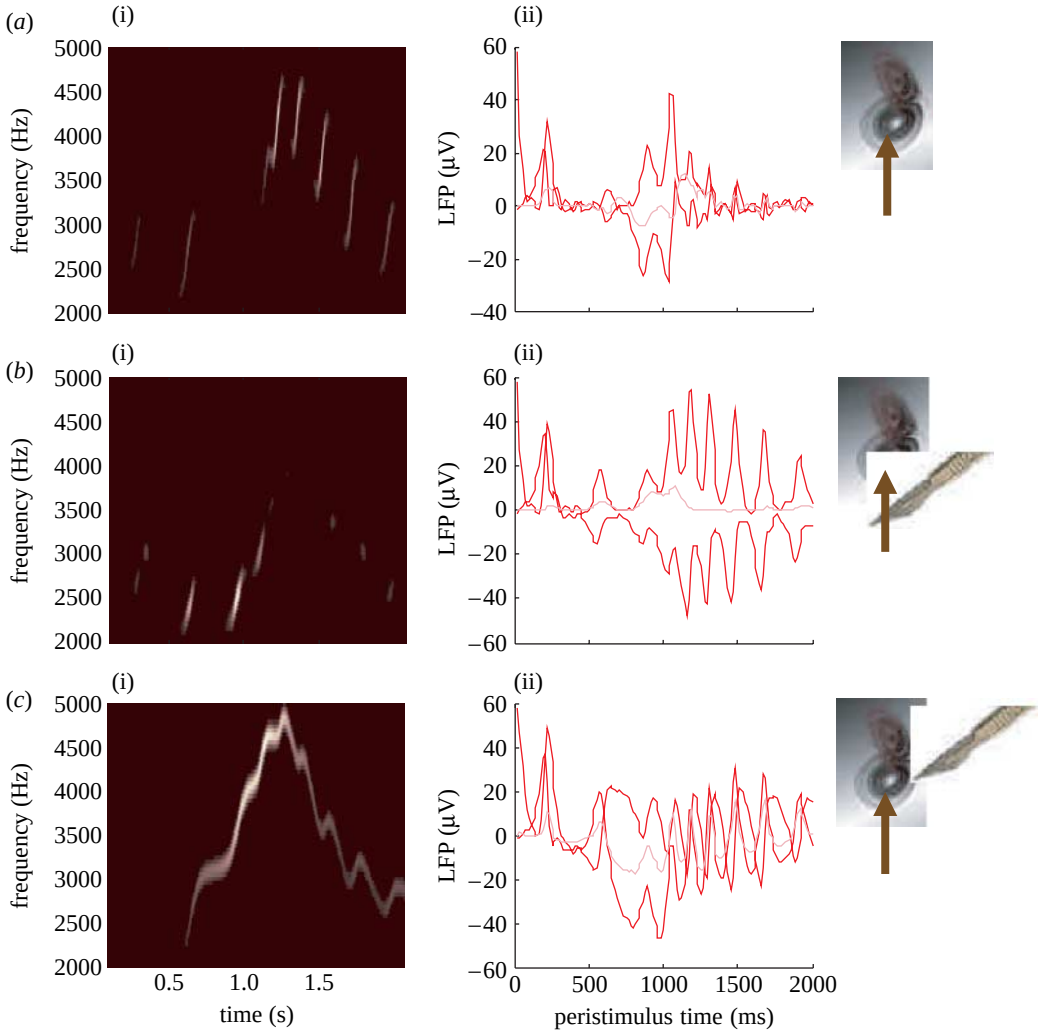

Figure 4. Results of simulated lesion studies using the birdsong model of figure 3. (i) The percept in terms of the predicted sonograms and (ii) the corresponding prediction error (at both the levels); these are the differences between the incoming sensory information and the prediction and the discrepancy between the conditional expectation of the second-level cause and that predicted by the second-level hidden states. (*a*) The recognition dynamics in the intact bird. (*b*) The percept and corresponding prediction errors when the connections between the hidden states at the second level and their corresponding causes are removed. This effectively removes structural priors on the evolution of the attractor manifold prescribing the sensory dynamics at the first level. (*c*) The effects of retaining the structural priors but removing the dynamical priors by cutting the connections that mediate inversion in generalized coordinates. These results suggest that both structural and dynamical priors are necessary for veridical perception.

structural priors entailed by the model. However, once the subsequent chirp had been predicted correctly, the following sequence of chirps was recognized with a high degree of conditional confidence. Note that when the second and third chirps in the sequence are not recognized, the first-level prediction error is high and the conditional confidence about the causes at the second level is low (reflected in the wide 90% confidence intervals). Heuristically, this means that the bird did not know which song was being emitted and was unable to predict subsequent chirps.

(**d**) *Structural and dynamical priors*

This example provides a nice opportunity to illustrate the relative roles of structural and dynamical priors. Structural priors are provided by the top-down inputs that dynamically reshape the manifold of the low-level attractor. However, the low-level attractor itself contains an abundance of dynamical priors that unfold in generalized coordinates. Both structural (extrinsic) and dynamical (intrinsic) priors provide important constraints, which facilitate recognition. We can selectively destroy these priors by lesioning the top-down connections to remove structural priors or by cutting the intrinsic connections that mediate dynamical priors. The latter involves cutting the self-connections in figure 1, among the state units. The results of these two simulated lesion experiments are shown in figure 4. Figure 4*a* shows the percept as in figure 3, in terms of the predicted sonogram and prediction error at the first and second levels. Figure 4*b*,*c* shows exactly the same thing but without structural and dynamical priors, respectively. In both cases, the synthetic bird fails to recognize the sequence with a corresponding inflation of prediction error, particularly at the sensory level. Interestingly, the removal of structural priors has a less marked effect on recognition than removing the dynamical priors.

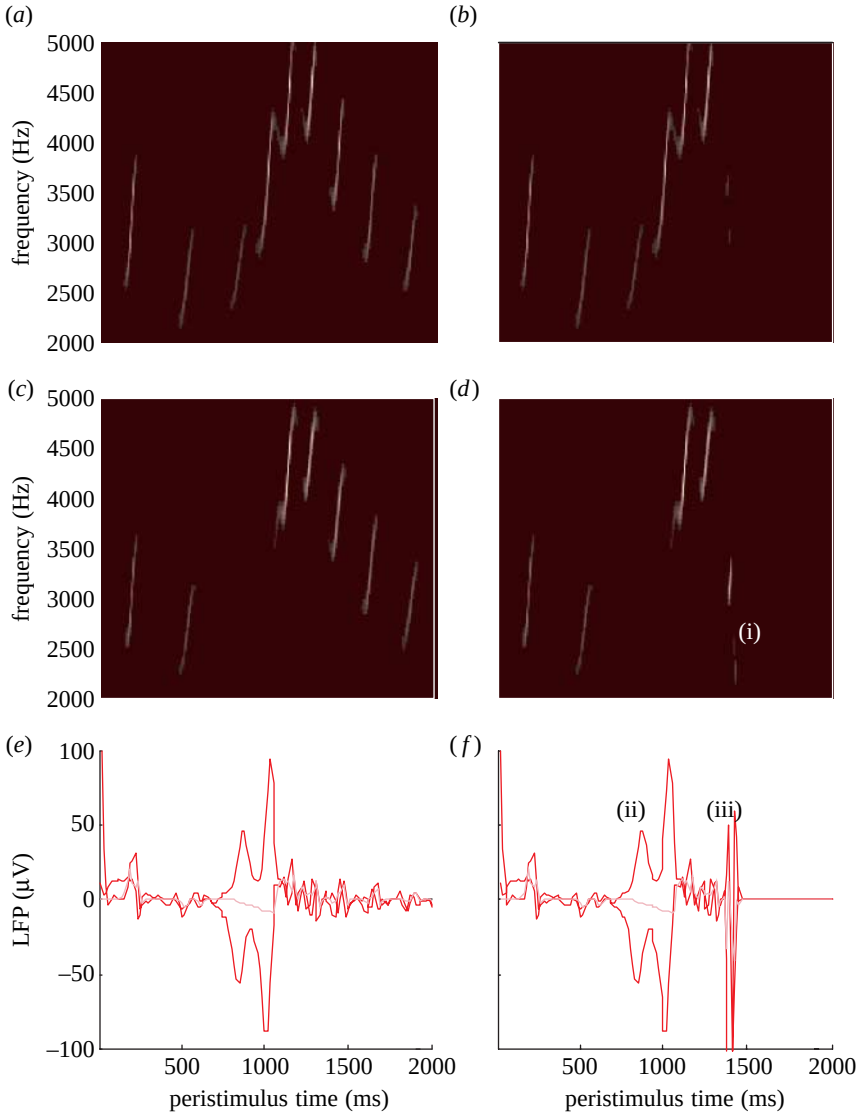

Figure 5. Omission-related responses. Here, we have omitted the last few chirps from the stimulus. (*a*,*c*,*e*) The original sequence and responses evoked. (*b*,*d*,*f*) The equivalent dynamics on omission of the last chirps. (*a*,*b*) The stimulus and (*c*,*d*) the corresponding percept in sonogram format. The interesting thing to note here is the occurrence of an anomalous percept after termination of the song in the lower right (i). This corresponds roughly to the chirp that would have been perceived in the absence of omission. (*e*,*f*) The corresponding (precision-weighted) prediction error under the two stimuli at both the levels. A comparison of the two reveals a burst of prediction error when a stimulus is missed (ii) and at the point that the stimulus terminates (iii), despite the fact that there is no stimulus present at this time. The red lines correspond to prediction error at the first level and the pink lines correspond to prediction error at the second level.

Without dynamical priors, there is a failure to segment the sensory stream and, although there is a preservation of frequency tracking, the dynamics *per se* have lost their sequential structure. Although it is interesting to compare and contrast the relative roles of structural and dynamics priors, the important message here is that both are necessary for veridical perception and that destruction of either leads to suboptimal inference. Both of these empirical priors prescribe dynamics, which enable the synthetic bird to predict what will be heard next. This leads to the question, 'what would happen if the song terminated prematurely?'

(e) *Omission and violation of predictions*

We repeated the above simulation but terminated the song after the fifth chirp. The corresponding sonograms and percepts are shown with their prediction errors in figure 5. Figure 5*a*,*c*,*e* shows the stimulus and percept as in figure 4, while figure 5*b*,*d*,*f* shows the stimulus and responses to omission of the last syllables. These results illustrate two important phenomena. First, there is a vigorous expression of prediction error after the song terminates abruptly. This reflects the dynamical nature of the recognition process because, at this point, there is no sensory input to predict. In other words, the prediction error is generated entirely by the predictions afforded by the dynamical model of sensory input. It can be seen that this prediction error (with a percept but no stimulus) is almost as large as the prediction error associated with the third and fourth stimuli that are not perceived (stimulus but no percept). Second, it can be seen that there is a transient percept, when the omitted chirp should have occurred. Its frequency is slightly too low but its timing is preserved in relation to the expected stimulus train. This is an interesting stimulation from the point of view

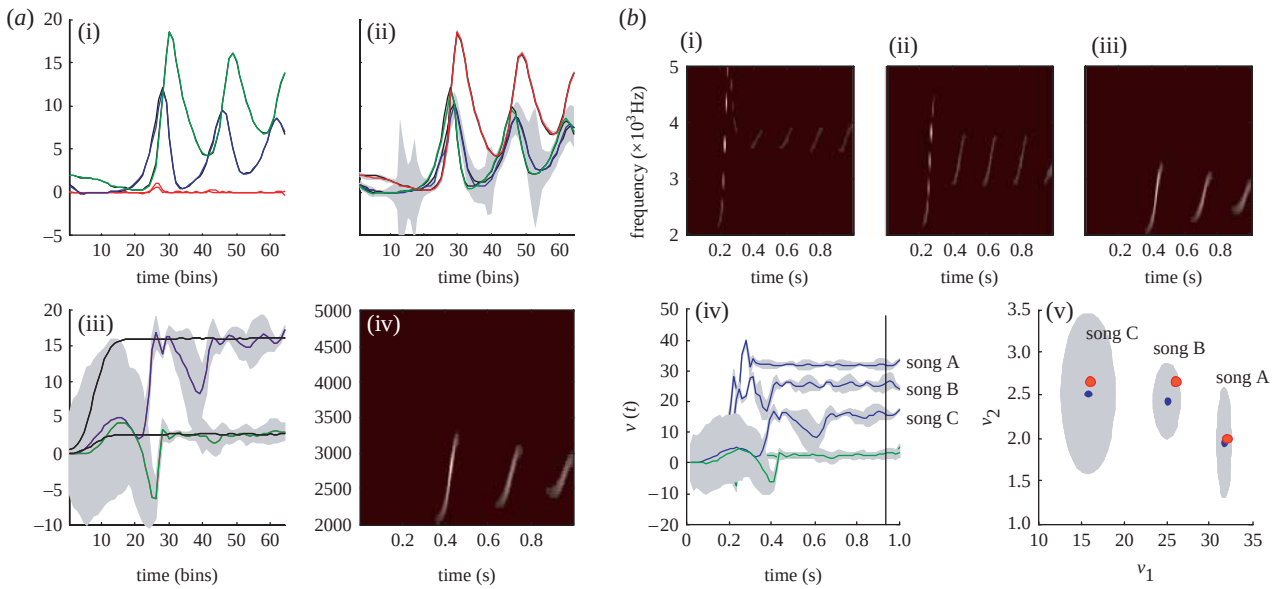

Figure 6. (*a*) Schematic of perceptual categorization ((i) prediction and error, (ii) hidden states, (iii) causes and (iv) percept). This follows the same format as in figure 3. However, here, there are no hidden states at the second level and the causes were subject to stationary and uninformative priors. This song was generated by a first-level attractor with fixed control parameters of $v_1^{(1)} = 16$ and $v_2^{(1)} = 8/3$, respectively. It can be seen that, on inversion of this model, these two control variables, corresponding to causes or states at the second level, are recovered with relatively high conditional precision. However, it takes approximately 50 iterations (approx. 600 ms) before they stabilize. In other words, the sensory sequence has been mapped correctly to a point in perceptual space after the occurrence of the second chirp. This song corresponds to song C (*b*(iii)). (*b*) The results of inversion for three songs ((i) song A, (ii) song B and (iii) song C) each produced with three distinct pairs of values for the second-level causes (the Rayleigh and Prandtl variables of the first-level attractor). (i–iii) The three songs shown in sonogram format correspond to a series of relatively high-frequency chirps that fall progressively in both frequency and number as the Rayleigh number is decreased. (iv) These are the second-level causes shown as a function of peristimulus time for the three songs. It can be seen that the causes are identified after approximately 600 ms with high conditional precision. (v) The conditional density on the causes shortly before the end of peristimulus time (vertical line in (iv)). The blue dots correspond to conditional means or expectations and the grey areas correspond to the conditional confidence regions. Note that these encompass the true values (red dots) used to generate the songs. These results indicate that there has been a successful categorization, in the sense that there is no ambiguity (from the point of view of the synthetic bird) about which song was heard.

of ERP studies of omission-related responses. These simulations and related empirical studies (e.g. Nordby *et al.* 1994; Yabe *et al.* 1997) provide clear evidence for the predictive capacity of the brain. In this example, prediction rests upon the internal construction of an attractor manifold that defines a family of trajectories, each corresponding to the realization of a particular song. In the last simulation, we look more closely at perceptual categorization of these songs.

(f) *Perceptual categorization*

In the previous simulations, we saw that a song corresponds to a sequence of chirps that are pre-ordained by the shape of an attractor manifold controlled by top-down inputs. This means that, for every point in the state space of the higher attractor, there is a corresponding manifold or category of song. In other words, recognizing or categorizing a particular song corresponds to finding a location in the higher state space. This provides a nice metaphor for perceptual categorization, because the neuronal states of the higher attractor represent, implicitly, a category of song. Inverting the generative model means that, probabilistically, we can map from a sequence of sensory events to a point in some perceptual space, where this mapping corresponds to perceptual recognition or categorization. This can be demonstrated in our synthetic songbird by suppressing the dynamics of the second-level attractor and letting its states optimize their location in perceptual space, to best predict the sensory input. To illustrate this, we generated three songs by fixing the Rayleigh and Prandtl variables to three distinct values. We then placed uninformative priors on the top-down causes (that were previously driven by the hidden states of the second-level attractor) and inverted the model in the usual way. Figure 6*a* shows the results of this simulation for a single song. This song comprises a series of relatively low-frequency chirps emitted every 250 ms or so. The causes of this song (song C in (*b*)) are recovered after the second chirp, with relatively tight confidence intervals (the blue and green lines in (iv)). We then repeated this for three songs. The results are shown in figure 6*b*. The songs are portrayed in sonogram format in figure 6*b*(i–iii) and the inferred perceptual causes in figure 6*b*(iv)(v). Figure 6*b*(iv) shows the evolution of these causes for all three songs as a function of peristimulus time and figure 6*b*(v) shows the corresponding conditional density in the causal or perceptual space after convergence. It can be seen that, for all three songs, the 90% confidence interval encompasses the true values (red dots). Furthermore, there is very little overlap between the conditional densities (grey regions), which means that the precision of the

perceptual categorization is almost 100 per cent. This is a simple but nice example of perceptual categorization, where sequences of sensory events, with extended temporal support, can be mapped to locations in perceptual space, through Bayesian deconvolution of the sort entailed by the free-energy principle.

5. CONCLUSION

This paper suggests that the architecture of cortical systems speak to hierarchical generative models in the brain. The estimation or inversion of these models corresponds to a generalized deconvolution of sensory inputs to disclose their causes. This deconvolution can be implemented in a neuronally plausible fashion, where neuronal dynamics self-organize when exposed to inputs to suppress free energy. The focus of this paper is on the nature of the hierarchical models and, in particular, models that show autonomous dynamics. These models may be relevant for the brain because they enable sequences of sequences to be inferred or recognized. We have tried to demonstrate their plausibility, in relation to empirical observations, by interpreting the prediction error, associated with model inversion, with observed electrophysiological responses. These models provide a graceful way to map from complicated spatio-temporal sensory trajectories to points in abstract perceptual spaces. Furthermore, in a hierarchical setting, this mapping may involve trajectories in perceptual spaces of increasingly higher order.

The ideas presented in this paper have a long history, starting with the notion of neuronal energy (Helmholtz 1860/1962), covering ideas such as efficient coding and analysis by synthesis (Barlow 1961; Neisser 1967) to more recent formulations in terms of Bayesian inversion and predictive coding (e.g. Ballard *et al.* 1983; Mumford 1992; Kawato *et al.* 1993; Dayan *et al.* 1995; Rao & Ballard 1998). This work tries to provide support for the notion that the brain uses attractors to represent and predict causes in the sensorium (Freeman 1987; Tsodyks 1999; Deco & Rolls 2003; Byrne *et al.* 2007).

This work was funded by the Wellcome Trust. We would like to thank our colleagues for their invaluable discussion about these ideas and Marcia Bennett for her help in preparing this manuscript.

All the schemes described in this paper are available in MATLAB code as academic freeware (http://www.fil.ion.ucl.ac.uk/spm). The simulation figures in this paper can be reproduced from a graphical user interface called from the DEM toolbox.

ENDNOTE

[1]To simplify notation, we use $f_x = \partial_x f = \partial f/\partial x$ for the partial derivative of the function, f, with respect to the variable x. We also use $\dot{x} = \partial_t x$ for temporal derivatives. Furthermore, we deal with variables in generalized coordinates of motion, denoted by a tilde; $\tilde{x} = [x, x', x'', \ldots]^{\mathrm{T}}$.

REFERENCES

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J. & Lund, J. S. 2002 Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.* **22**, 8633–8646.

Ballard, D. H., Hinton, G. E. & Sejnowski, T. J. 1983 Parallel visual computation. *Nature* **306**, 21–26. (doi:10.1038/306021a0)

Barlow, H. B. 1961 Possible principles underlying the transformation of sensory messages. In *Sensory communication* (ed. W. A. Rosenblith), pp. 217–234. Cambridge, MA: MIT press.

Botvinick, M. M. 2007 Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Phil. Trans. R. Soc. B* **362**, 1615–1626. (doi:10.1098/rstb.2007.2056)

Breakspear, M. & Stam, C. J. 2005 Dynamics of a neural system with a multiscale architecture. *Phil. Trans. R. Soc. B* **360**, 1051–1107. (doi:10.1098/rstb.2005.1643)

Byrne, P., Becker, S. & Burgess, N. 2007 Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychol. Rev.* **114**, 340–375. (doi:10.1037/0033-295X.114.2.340)

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., Berger, M. S., Barbaro, N. M. & Knight, R. T. 2006 High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**, 1626–1628. (doi:10.1126/science.1128115)

Chait, M., Poeppel, D., de Cheveigné, A. & Simon, J. Z. 2007 Processing asymmetry of transitions between order and disorder in human auditory cortex. *J. Neurosci.* **27**, 5207–5214. (doi:10.1523/JNEUROSCI.0318-07.2007)

Dayan, P., Hinton, G. E. & Neal, R. M. 1995 The Helmholtz machine. *Neural Comput.* **7**, 889–904. (doi:10.1162/neco.1995.7.5.889)

Deco, G. & Rolls, E. T. 2003 Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *Eur. J. Neurosci.* **18**, 2374–2390. (doi:10.1046/j.1460-9568.2003.02956.x)

DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. 2002 Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* **31**, 299–316. (doi:10.1023/A:1024130211265)

Efron, B. & Morris, C. 1973 Stein's estimation rule and its competitors: an empirical Bayes approach. *J. Am. Stat. Assoc.* **68**, 117–130. (doi:10.2307/2284155)

Felleman, D. J. & Van Essen, D. C. 1991 Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47. (doi:10.1093/cercor/1.1.1-a)

Feynman, R. P. 1972 *Statistical mechanics*. Reading, MA: Benjamin.

Fletcher, N. H. 2000 A class of chaotic bird calls? *J. Acoust. Soc. Am.* **108**, 821–826. (doi:10.1121/1.429615)

Freeman, W. J. 1987 Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biol. Cybern.* **56**, 139–150. (doi:10.1007/BF00317988)

Friston, K. J. 1997 Transients, metastability, and neuronal dynamics. *NeuroImage* **5**, 164–171. (doi:10.1006/nimg.1997.0259)

Friston, K. J. 2003 Learning and inference in the brain. *Neural Netw.* **16**, 1325–1352. (doi:10.1016/j.neunet.2003.06.005)

Friston, K. J. 2005 A theory of cortical responses. *Phil. Trans. R. Soc. B* **360**, 815–836. (doi:10.1098/rstb.2005.1622)

Friston, K. 2008 Hierarchical models in the brain. *PLoS Comput. Biol.* **4**, e1000211. (doi:10.1371/journal.pcbi.1000211)

Friston, K., Kilner, J. & Harrison, L. 2006 A free-energy principle for the brain. *J. Physiol. Paris* **100**, 70–87. (doi:10.1016/j.jphysparis.2006.10.001)

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J. & Penny, W. 2007 Variational Bayes and the Laplace approximation. *NeuroImage* **34**, 220–234. (doi:10.1016/j.neuroimage.2006.08.035)

Haken, H., Kelso, J. A. S., Fuchs, A. & Pandya, A. S. 1990 Dynamic pattern-recognition of coordinated biological motion. *Neural Netw.* **3**, 395–401. (doi:10.1016/0893-6080(90)90022-D)

Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. 2008 A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550. (doi:10.1523/JNEUROSCI.5487-07.2008)

Helmholtz, H. 1860/1962 *Handbuch der physiologischen optik*. (English transl. J. P. C. Southall), vol. 3. New York, NY: Dover.

Hinton, G. E. & von Cramp, D. 1993 Keeping neural networks simple by minimising the description length of weights. In *Proc. COLT*, vol. 93. pp. 5–13.

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P. & Bullier, J. 1998 Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* **394**, 784–787. (doi:10.1038/29537)

Jirsa, V. K., Fuchs, A. & Kelso, J. A. 1998 Connecting cortical and behavioral dynamics: bimanual coordination. *Neural Comput.* **10**, 2019–2045. (doi:10.1162/089976698300016954)

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

Kiebel, S. J., Daunizeau, J. & Friston, K. J. 2008 A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* **4**, e1000209. (doi:10.1371/journal.pcbi.1000209)

Kopell, N., Ermentrout, G. B., Whittington, M. A. & Traub, R. D. 2000 Gamma rhythms and beta rhythms have different synchronization properties. *Proc. Natl Acad. Sci. USA* **97**, 1867–1872. (doi:10.1073/pnas.97.4.1867)

Laje, R. & Mindlin, G. B. 2002 Diversity within a birdsong. *Phys. Rev. Lett.* **89**, 288102. (doi:10.1103/PhysRevLett.89.288102)

Laje, R., Gardner, T. J. & Mindlin, G. B. 2002 Neuromuscular control of vocalizations in birdsong: a model. *Phys. Rev. E Stat. Nonlin. Softw. Matter Phys.* **65**, 051921. (doi:10.1103/PhysRevE.65.051921)

MacKay, D. J. C. 1995 Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electron. Lett.* **31**, 445–447. (doi:10.1049/el:19950331)

Maunsell, J. H. & Van Essen, D. C. 1983 The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.* **3**, 2563–2586.

McCrea, D. A. & Rybak, I. A. 2008 Organization of mammalian locomotor rhythm and pattern generation. *Brain Res. Rev.* **57**, 134–146. (doi:10.1016/j.brainresrev.2007.08.006)

Mumford, D. 1992 On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* **66**, 241–251. (doi:10.1007/BF00198477)

Murphy, P. C. & Sillito, A. M. 1987 Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature* **329**, 727–729. (doi:10.1038/329727a0)

Neal, R. M. & Hinton, G. E. 1998 A view of the EM algorithm that justifies incremental sparse and other variants. In *Learning in graphical models* (ed. M. I. Jordan), pp. 355–368. Dordrecht, The Netherlands: Kulwer Academic Press.

Neisser, U. 1967 *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.

Nordby, H., Hammerborg, D., Roth, W. T. & Hugdahl, K. 1994 ERPs for infrequent omissions and inclusions of stimulus elements. *Psychophysiology* **31**, 544–552. (doi:10.1111/j.1469-8986.1994.tb02347.x)

Rabinovich, M., Huerta, R. & Laurent, G. 2008 Neuroscience. Transient dynamics for neural processing. *Science* **321**, 48–50. (doi:10.1126/science.1155564)

Raizada, R. D. & Grossberg, S. 2003 Towards a theory of the laminar architecture of cerebral cortex: computational clues from the visual system. *Cereb. Cortex* **13**, 100–113. (doi:10.1093/cercor/13.1.100)

Rao, R. P. & Ballard, D. H. 1998 Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci.* **2**, 79–87. (doi:10.1038/4580)

Rockland, K. S. & Pandya, D. N. 1979 Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res.* **179**, 3–20. (doi:10.1016/0006-8993(79)90485-2)

Rosier, A. M., Arckens, L., Orban, G. A. & Vandesande, F. 1993 Laminar distribution of NMDA receptors in cat and monkey visual cortex visualized by [3H]-MK-801 binding. *J. Comp. Neurol.* **335**, 369–380. (doi:10.1002/cne.903350307)

Sherman, S. M. & Guillery, R. W. 1998 On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proc. Natl Acad. Sci. USA* **95**, 7121–7126. (doi:10.1073/pnas.95.12.7121)

Tsodyks, M. 1999 Attractor neural network models of spatial maps in hippocampus. *Hippocampus* **9**, 481–489. (doi:10.1002/(SICI)1098-1063(1999)9:4<481::AID-HIPO14>3.0.CO;2-S)

Yabe, H., Tervaniemi, M., Reinikainen, K. & Näätänen, R. 1997 Temporal window of integration revealed by MMN to sound omission. *NeuroReport* **8**, 1971–1974. (doi:10.1097/00001756-199705260-00035)

Zeki, S. & Shipp, S. 1988 The functional logic of cortical connections. *Nature* **335**, 311–331. (doi:10.1038/335311a0)