New Mathematics and Natural Computation
Vol. 5, No. 1 (2009) 1-32

ATTRACTORS IN SONG

KARL FRISTON and STEFAN KIEBEL

The Wellcome Trust Centre of Neuroimaging
University College London,
Queen Square, London WC1N 3BG
k.friston@fil.ion.ucl.ac.uk
s.kiebel@fil.ion.ucl.ac.uk

This paper summarizes our recent attempts to integrate action and perception within a single optimization framework. We start with a statistical formulation of Helmholtz's ideas about neural energy to furnish a model of perceptual inference and learning that can explain a remarkable range of neurobiological facts. Using constructs from statistical physics it can be shown that the problems of inferring the causes of our sensory inputs and learning regularities in the sensorium can be resolved using exactly the same principles. Furthermore, inference and learning can proceed in a biologically plausible fashion. The ensuing scheme rests on Empirical Bayes and hierarchical models of how sensory information is generated. The use of hierarchical models enables the brain to construct prior expectations in a dynamic and context-sensitive fashion. This scheme provides a principled way to understand many aspects of the brain's organization and responses. We will demonstrate the brain-like dynamics that this scheme entails by using models of bird songs that are based on chaotic attractors with autonomous dynamics. This provides a nice example of how nonlinear dynamics can be exploited by the brain to represent and predict dynamics in the environment.

Keywords: Generative models; Predictive coding; Hierarchical; Dynamic; Nonlinear; Variational; Birdsong.

1 Introduction

The seminal work of Walter Freeman introduced a number of key concepts into neuroscience. Among these was the notion that the causes of sensory input could be encoded in a distributed fashion by neuronal dynamics and associated attractors.[1] This paper considers prediction and perceptual categorisation as an inference problem that is solved by the brain. We assume that the brain models the world as a hierarchy or cascade of dynamic systems that encode causal structure in the sensorium. Perception is equated with the optimisation or inversion of these internal models, to explain sensory data. Given a model of how sensory data are generated, we can use a generic

approach to model inversion, based on a free-energy bound on the model's evidence. The ensuing free-energy principle furnishes equations that prescribe recognition; *i.e.*, the dynamics of neuronal representations that represent the causes of sensory input. Here, we focus on a very general model, whose hierarchical and dynamical structure enables simulated brains to recognise and predict trajectories or sequences of sensory states. We first review hierarchical dynamic models and their inversion. We then show that the brain has the necessary infrastructure to implement this inversion and present stimulations using synthetic birds that generate and recognise birdsongs.

Critically, the nature of the inversion lends itself to a relatively simple neural network implementation that shares many formal similarities with real cortical hierarchies in the brain. The basic idea that the brain uses hierarchical inference has been described in a series of papers (see Refs. 2 and 3). These papers entertain the notion that the brain uses empirical Bayes for inference about its sensory input, given the hierarchical organisation of cortical systems. Here, we generalise this idea to cover dynamical models and consider how neural networks could be configured to invert these model and deconvolve sensory causes from sensory input.

This paper comprises six sections. In the second section, we introduce hierarchical dynamic models. These cover most observation or generative models encountered in the estimation and inference literature. An important aspect of these models is their formulation in generalised coordinates of motion; this lends them a hierarchal form in both structure and dynamics. These hierarchies induce empirical priors that provide structural and dynamic constraints, which can be exploited during inversion. In the third section, we consider model inversion in statistical terms. This summarises the material in Ref. 4. In the fourth section, we show how inversion can be formulated as a simple gradient ascent using neuronal networks and, in the fifth section, present a simple empirical experiment that substantiates some of the key predictions of this formulation. In the final section, we consider how evoked brain responses might be understood in terms of inference under hierarchical dynamic models of sensory input.

To simplify notation we will use $f_x = \partial_x f = \partial f / \partial x$ to denote 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 will be dealing with variables in generalised coordinates of motion, denoted by a tilde; $\tilde{x} = [x, x', x'', \ldots]^T$, where $x^{[i]}$ denotes *i*-th order motion.

2 Hierarchical Dynamic Models

Hierarchical Dynamic models are probabilistic models $p(y, \vartheta) = p(y \mid \vartheta) p(\vartheta)$ based on state-space models. They entail the likelihood, $p(y \mid \vartheta)$ of getting some data, y, given some parameters $\vartheta = \{x, v, \theta\}$ and priors on those parameters, $p(\vartheta)$.

We will see that the parameters subsume different quantities, some of which change with time and some which do not. A dynamic model can be written as

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

The continuous nonlinear functions f and g of the states are parameterised by θ. The states $v(t)$ can be deterministic, stochastic, or both. They 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 the stochastic innovations (*i.e.*, observation noise) $z(t)$ are analytic, such that the covariance of $\tilde{z} = [z, z', z'', \ldots]^T$ is well defined; similarly for the system or state noise, $w(t)$, which represents random fluctuations on the motion of the hidden states. Under local linearity assumptions, the generalised motion of the output or response $\tilde{y} = [y, y', y'', \ldots]^T$ is given by

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

The first (observer) equation show that the generalised states $u = [\tilde{v}, \tilde{x},]^T$ are needed to generate a generalised response that encodes a path or trajectory. The second (state) equations enforce a coupling between orders of 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 $\tilde{g} = [g, g', g'', \ldots]^T$ and motion $\tilde{f}$ 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 $p(\tilde{z}) = N(\tilde{z} : 0, \tilde{\Sigma}^z)$ provide the likelihood, $p(\tilde{y} \mid \tilde{x}, \tilde{v})$. Similarly, Gaussian assumptions about state-noise $p(\tilde{w}) = N(\tilde{w} : 0, \tilde{\Sigma}^w)$ furnish empirical priors, $p(\tilde{x} \mid \tilde{v})$ in terms of predicted motion

$$\begin{aligned} p(\tilde{y}, \tilde{x}, \tilde{v}) &= p(\tilde{y} \mid \tilde{x}, \tilde{v}) p(\tilde{x}, \tilde{v}) \\ p(\tilde{x}, \tilde{v}) &= p(\tilde{x} \mid \tilde{v}) p(\tilde{v}) \end{aligned}$$

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

Here, $x^{[i]}$ means the *i*-th generalised motion. Here, we have assumed Gaussian priors $p(\tilde{v})$ on the generalised causes, with mean $\tilde{\eta}$ and covariance $\tilde{\Sigma}^v$. The density on the hidden states $p(\tilde{x} \mid \tilde{v})$ is part of the prior on quantities needed to evaluate the likelihood of the response or output. The form of this prior means that low-order motion constrains high-order motion (and *vice versa*). 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 factorise; $\tilde{\Sigma}^i = \Sigma^i \otimes R^i$ into a covariance proper and a matrix of correlations R^i among generalised motion that encodes an autocorrelation function.

2.1 Hierarchical forms

Hierarchical dynamic models have the following form, which generalises the ($m = 1$) model above

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

Again, $f^{(i)} = f(x^{(i)}, v^{(i)})$ and $g^{(i)} = g(x^{(i)}, v^{(i)})$ are continuous nonlinear functions of the states. The innovations $z^{(i)}$ and $w^{(i)}$ are conditionally independent fluctuations that enter each level of the hierarchy. These play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causes $v = [v^{(1)}, \ldots, v^{(m)}]^T$ link levels, whereas the hidden states $x = [x^{(1)}, \ldots, x^{(m)}]^T$ link dynamics over time. In 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 generalised convolutions with 'deep' (*i.e.*, hierarchical) structure.

The conditional independence of the fluctuations means that these models have a Markov property over levels, which simplifies the architecture of attending inference schemes For example, the prediction $\tilde{g}^{(i)} = \tilde{g}(\tilde{x}^{(i)}, \tilde{v}^{(i)})$ plays the role of a prior expectation on $\tilde{v}^{(i-1)}$, yet it has to be estimated in terms of $\tilde{x}^{(i)}, \tilde{v}^{(i)}$. This makes it an empirical prior.[5] See Kass and Steffey[6] for a discussion of approximate Bayesian inference in conditionally independent hierarchical models of static data and Ref. 4 for dynamic models. In short, a hierarchical form endows models with the ability to construct their own priors. This feature is central to many inference and estimation procedures, ranging from mixed-effects analyses in classical covariance component analysis to automatic relevance determination in machine learning.

2.2 Summary

In this section, we have introduced hierarchical dynamic models in generalised coordinates of motion. These models are about 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 the hierarchy implicit in generalised motion. We now consider how these models are inverted to disclose the unknown states generating observed sensory data.

3. Model inversion and variational Bayes

This section considers variational inversion and provides a heuristic summary of the material in Friston *et al.*[4] Variational Bayes is a generic approach to model inversion that approximates the conditional density $p(\vartheta \mid y, m)$ on some model parameters, ϑ, given a model m and data y. This is achieved by optimising the sufficient statistics of an approximate conditional density $q(\vartheta)$ with respect to a lower bound on the

evidence $p(y \mid m)$ of the model itself.[7, 8] The log-evidence can be expressed in terms of a free-energy and divergence term

$$\ln p(y \mid m) = F + K(q(\vartheta) \,||\, p(\vartheta \mid y, m)) \Rightarrow$$

$$\begin{aligned} F &= G - H \\ G &= \left\langle \ln p(y, \vartheta) \right\rangle_q \\ H &= \left\langle \ln q(\vartheta) \right\rangle_q \end{aligned} \tag{3.1}$$

The free-energy comprises an energy term, G, corresponding to an internal energy, $U(y, \vartheta) = \ln p(y, \vartheta)$ expected under the density $q(\vartheta)$ and its entropy, H, which is a measure of its uncertainty. Eq. (3.1) shows that $F(y, q)$ is a lower-bound on the log-evidence because the divergence, $K \geq 0$ is always positive. The objective is to optimise the sufficient statistics of $q(\vartheta)$ by maximising the free-energy and minimising the divergence. This renders $q(\vartheta) \approx p(\vartheta \mid y, m)$ an approximate posterior, which is exact for simple (*e.g.*, linear) systems[a].

3.1 Mean-field and Laplace approximations

Invoking the density, $q(\vartheta)$ converts a difficult integration problem (inherent in computing the evidence) into an easier optimisation problem. This rests on inducing a bound that can be optimised with respect to $q(\vartheta)$. To finesse optimisation, one usually assumes $q(\vartheta)$ factorises over a partition of the parameters

$$\begin{aligned} q(\vartheta) &= \prod_i q(\vartheta^i) \\ \vartheta &= \{u, \theta\} \end{aligned} \tag{3.2}$$

In statistical physics, this is called a mean-field approximation. Under our hierarchical dynamic model we will assume, $q(\vartheta) = q(u(t))q(\theta)$, where $u(t)$ are time-varying generalised states and θ are all the other unknown time-invariant parameters. In a dynamic setting, the conditional density on the states and the free-energy are functionals of time. By analogy with Lagrangian mechanics, this calls on the notion of *action.* Action is the anti-derivative or path-integral of energy. We will

[a] By convention, the free energy in machine learning is usually the negative of the free-energy in physics. This means the free-energy increases with log-evidence and has to be maximised.

denote the action associated with the free energy by $\bar{F}$, such that $\partial_t \bar{F} = F$. We now seek $q(\vartheta^i)$ that maximise the free-action. It is fairly easy to show[4] that the solution for the states is a functional of their instantaneous energy, $U(t) := \ln p(\tilde{y}, u \mid \theta)$

$$\begin{aligned} q(u(t)) &\propto \exp(V(t)) \\ V(t) &= \langle U(t) \rangle_{q(\theta)} \end{aligned} \tag{3.3}$$

where $V(t)$ is their variational energy. The variational energy of the states is simply their instantaneous energy expected under the conditional density of the parameters. In contrast, the conditional density of the parameters is a function of their variational action

$$\begin{aligned} q(\theta) &\propto \exp(\bar{V}^\theta) \\ V^\theta &= \langle U(t) \rangle_{q(u)} \\ \bar{V}^\theta &= \int V^\theta dt + U^\theta \end{aligned} \tag{3.4}$$

$U^\theta = \ln p(\theta)$ are the prior energies of the parameters and play the role of integration constants in the corresponding variational action; $\bar{V}^\theta$.

These equations provide closed-form expressions for the conditional or variational density in terms of the internal energy defined by our model (see previous section). They are intuitively sensible, because the conditional density of the states should reflect the instantaneous energy (3.3); whereas the conditional density of the parameters can only be determined after all the data have been observed (3.4). In other words, the variational energy involves the prior energy and an integral of time-dependent energy. In the absence of data, when the integrals are zero, the conditional density reduces to the prior density.

To further simplify things, we will assume the brain uses something called the Laplace approximation. This enables us to focus on a single quantity for each unknown, the conditional mean: Under the Laplace approximation, the conditional density on the states assumes a fixed Gaussian form $q(u(t)) = N(u : \tilde{\mu}, C)$ with sufficient statistics $\tilde{\mu}(t)$ and $C(t)$, corresponding to the conditional mean and covariance; similarly, $q(\theta) = N(u : \mu^\theta, C^\theta)$ for the parameters. The advantage of the Laplace assumption is that the conditional precisions (inverse variances) are functions of the mean (the curvature of the instantaneous energy at the mean for the states and the curvature of the corresponding action for the parameters). This means we can reduce model inversion to optimising one sufficient statistic; namely, the conditional mean.

3.2 Optimising the conditional means

For the parameters, we envisage the brain uses a simple gradient ascent (see Section 5) to maximise its variational action, under which (from Eq. 3.4)

$$\begin{aligned}\dot{\mu}^{\theta} &= \bar{V}_{\theta}^{\theta} \\ \dot{\bar{V}}_{\theta}^{\theta} &= V_{\theta}^{\theta}\end{aligned} \tag{3.5}$$

Similarly, the trajectory of the conditional mean of the states maximises variational action, which is the solution to the ansatz

$$\dot{\tilde{\mu}} - D\tilde{\mu} = V(t)_u \tag{3.6}$$

This can be regarded as an augmented form of the gradient ascent $\dot{\tilde{\mu}} = V(t)_u$. Here, $\dot{\tilde{\mu}} - D\tilde{\mu}$ is motion in a frame of reference that moves along the trajectory encoded in generalised coordinates. Critically, the stationary solution, in this moving frame of reference, maximises variational action. This can be seen easily by noting $\dot{\tilde{\mu}} - D\tilde{\mu} = 0$ means the gradient of the variational energy is zero and

$$\begin{aligned}\partial_u V(t) = 0 \Leftrightarrow \delta_u \bar{V} = 0 \\ \partial_t \bar{V} = V(t)\end{aligned} \tag{3.7}$$

This is sufficient for the mode to maximise variational action (by the Fundamental lemma of variational calculus). Intuitively, this means tiny perturbations to its path do not change the variational energy and it has the greatest variational action (*i.e.*, path-integral of variational energy) of all possible paths. This may sound a little complicated but it is simply a version of Hamilton's principle of stationary action, which allows the conditional mean in Eq. (3.6) to converge on a 'moving target'. At this point the trajectory of the mean becomes the mean of the trajectory and $\dot{\tilde{\mu}} = D\tilde{\mu}$.

3.3 Summary

In this section, we have seen how the inversion of dynamic models can be formulated as an optimization of free-action. This action is the path-integral of free-energy associated with changing states. By assuming a fixed-form (Laplace) approximation to the conditional density, one can reduce optimisation to finding the conditional means of unknown quantities. For the states, this entails finding a path or trajectory with stationary variational action. This can be formulated as a gradient ascent in a frame of

reference that moves along the path encoded in generalised coordinates. The only thing we need to implement this recognition scheme is the variational or internal energy, $U(t) := \ln p(\tilde{y}, u)$, which is specified by the generative model of the previous section (Eq. 2.5).

4 Hierarchical models in the brain

A key architectural principle of the brain is its hierarchical organisation.[9,10] 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.[11,12,13] 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 supra-granular layers and terminate on spiny stellate cells of layer four in higher cortical areas.[12,14] Conversely, backward connections arise largely from deep pyramidal cells in infra-granular layers and target cells in the infra and supra-granular 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 that renders backward connections more modulatory or nonlinear in their effects on neuronal responses.[15,16] This is consistent with the deployment of voltage-sensitive NMDA receptors in the supra-granular layers that are targeted by backward connections.[17] 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.

The hierarchical structure of the brain speaks to hierarchical models of sensory input. We now consider how this functional architecture can be understood under the inversion of hierarchical models by the brain.

4.1 Perceptual inference

If we assume that the activity of neurons encode the conditional mean of states, then Eq. (3.6) specifies the neuronal dynamics entailed by perception or recognizing states of the world from sensory data. In Friston *et al*[4] we show how these dynamics can be expressed simply in terms of auxiliary variables

$$\varepsilon^v = \begin{bmatrix} y \\ v^{(1)} \\ \vdots \\ v^{(m)} \end{bmatrix} - \begin{bmatrix} g^{(1)} \\ g^{(2)} \\ \vdots \\ \eta \end{bmatrix} \qquad \varepsilon^x = \begin{bmatrix} Dx^{(1)} \\ \vdots \\ Dx^{(m)} \end{bmatrix} - \begin{bmatrix} f^{(1)} \\ \vdots \\ f^{(m)} \end{bmatrix} \qquad \tilde{\varepsilon} = \begin{bmatrix} \tilde{\varepsilon}^v \\ \tilde{\varepsilon}^x \end{bmatrix} \tag{4.1}$$

These correspond to prediction errors on the causes and motion of the hidden states. Using these errors we can write Eq. (3.6) as

$$\begin{aligned} \dot{\tilde{\mu}} &= V(t)_u + D\tilde{\mu} \\ &= D\tilde{\mu} - \tilde{\varepsilon}_u^T \xi \qquad\qquad \tilde{\Pi} = \begin{bmatrix} \tilde{\Pi}^z & \\ & \tilde{\Pi}^w \end{bmatrix} \\ \xi &= \tilde{\Pi}\tilde{\varepsilon} = \tilde{\varepsilon} - \Lambda\xi \end{aligned} \tag{4.2}$$

This equation describes how neuronal states self-organise, when exposed to sensory input. Its form is quite revealing and suggests two distinct populations of neurons; causal or hidden *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 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 $-\Lambda\xi$ decorrelate the error-units. The matrix $\Lambda = \tilde{\Sigma} - 1$ can be thought of as lateral connections among error-units that mediate winner-take-all like interactions and increases with higher levels of noise or uncertainty.

4.1.1 Hierarchical message passing

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

$$\begin{aligned} \dot{\tilde{\mu}}^{(i)v} &= D\tilde{\mu}^{(i)v} - \tilde{\varepsilon}_v^{(i)T}\xi^{(i)} - \xi^{(i+1)v} \\ \dot{\tilde{\mu}}^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{\varepsilon}_x^{(i)T}\xi^{(i)} \\ \\ \xi^{(i)v} &= \tilde{\mu}^{(i-1)v} - \tilde{g}(\tilde{\mu}^{(i)}) - \Lambda^{(i)z}\xi^{(i)v} \\ \xi^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{f}(\tilde{\mu}^{(i)}) - \Lambda^{(i)w}\xi^{(i)x} \end{aligned} \tag{4.3}$$

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 (Fig. 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 away 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 optimise 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.

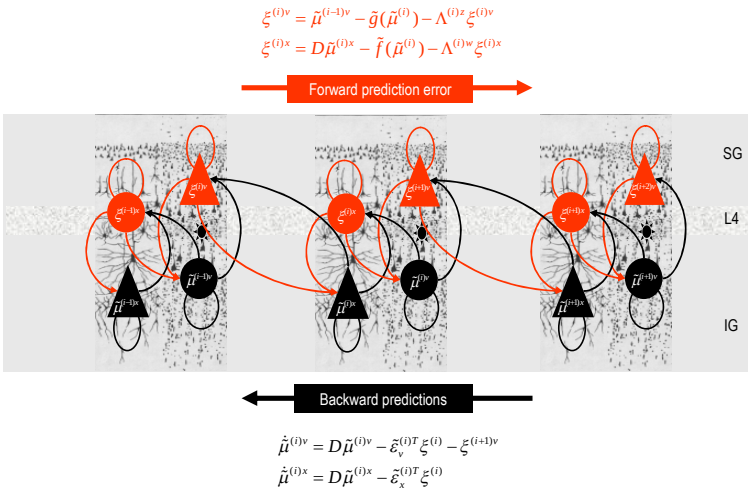

Fig. 1: Schematic detailing the neuronal architectures that encode an ensemble density on the states of a hierarchical model. This schematic shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that are used to construct predictions. These predictions try to explain away 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 optimisation scheme detailed in the main text. The state-units and their efferents are in black and the error-units in red, with causes on the right and hidden states on the left. For simplicity, we have assumed the output of each level is a function of, and only of, the hidden states. This induces a hierarchy over levels and, within each level, a hierarchical relationship between states, where causes predict hidden states. This schematic shows how the neuronal populations may be deployed hierarchically within three cortical areas (or macro-columns). Within each area the cells are shown in relation to the laminar structure of the cortex that includes supra-granular (**SG**) granular (**L4**) and infra-granular (**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 (EEG) signals that can be measured non-invasively. Similarly, the only messages that are passed down the hierarchy are the predictions from state-units that are necessary to form prediction errors in lower levels. The sources of extrinsic backward connections are the deep pyramidal cells and one might deduce that these encode the expected causes of sensory states (Fig. 1). Critically, the motion of each state-unit is a linear mixture of bottom-up prediction error (4.3). This is exactly what is observed physiologically; in that 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 nonlinear or modulatory characteristics of backward connections.

4.1.2 Encoding generalised motion

Equation (4.3) is cast in terms of generalised states. This suggests that the brain has an explicit representation of generalised motion. In other words, there are separable neuronal codes for different orders of motion. This is perfectly consistent with empirical evidence for distinct populations of neurons encoding elemental visual features and their motion (*e.g.*, motion-sensitive area V5; Ref. 10). The analysis above suggests that acceleration and higher-order motion are also encoded; each providing constraints on a lower order, through $D\tilde{\mu}$. Here, D represents a fixed connectivity matrix that mediates these temporal constraints. Notice that $\dot{\tilde{\mu}} = D\tilde{\mu}$ only when $\tilde{\varepsilon}_u^T \xi = 0$. This means it is perfectly possible to represent the motion of a state that is inconsistent with the state of motion. The motion after-effect is a nice example of this, where a motion percept co-exists with no change in the perceived location of visual stimuli. The encoding of generalised motion may mean that we represent paths or trajectories of sensory dynamics over short periods of time and that there is no perceptual instant. One could speculate that the encoding of different orders of motion may involve rate codes in distinct neuronal populations or multiplexed temporal codes in the same populations (*e.g.*, in different frequency band.

When sampling sensory data, one can imagine easily how receptors generate $\tilde{\mu}^{(0)} := \tilde{y}$. Indeed, it would be surprising to find any sensory system that did not respond to a high-order derivative of changing sensory fields (*e.g.*, acoustic edge detection; offset units in the visual system, *etc*; Chait *et al* 2007). Note that sampling high-order derivatives is formally equivalent to high-pass filtering sensory data. A simple consequence of encoding generalised motion is, in electrophysiological terms, the emergence of spatiotemporal receptive fields that belie selectivity to particular sensory trajectories.

4.2 Perceptual learning and plasticity

The conditional expectations of the parameters, μ^{θ} control the construction of prediction error through backward and lateral connections. This suggests that they are encoded in the strength of extrinsic and intrinsic connections. If we define effective connectivity as the rate of change of a unit's response with respect to its inputs, Eq. 4.2 suggests an interesting anti-symmetry in the effective connectivity between the state and error-units. The effective connectivity from the states to the error-units is $\partial_{\tilde{\mu}}\tilde{\xi} = \tilde{\varepsilon}_u$. This is simply the negative transpose of the effective connectivity that mediates recognition dynamics; $\partial_{\tilde{\xi}}\dot{\tilde{\mu}} = -\tilde{\varepsilon}_u^T$. In other words, the effective connection from any state to any error-unit has the same strength (but opposite sign) of the reciprocal connection from the error to the state-unit. This means we would expect to see connections reciprocated in the brain, which is generally the case.[10,12] Furthermore, we would not expect to see positive feedback loops (*c.f.*, Ref. 18). We now consider the synaptic efficacies underlying effective connectivity.

If synaptic efficacy encodes the parameter estimates, we can cast parameter optimisation as changing synaptic connections. These changes have a relatively simple form that is recognisable as associative plasticity. To show this, we will make the simplifying but plausible assumption that the brain's generative model is based on nonlinear functions a of linear mixtures of states

$$\begin{aligned} f^{(i)} &= a(\theta^{(i)_1} x^{(i)} + \theta^{(i)_2} v^{(i)}) \\ g^{(i)} &= a(\theta^{(i)_3} x^{(i)} + \theta^{(i)_4} v^{(i)}) \end{aligned} \tag{4.4}$$

Under this assumption $\theta^{(i)_j}$ correspond to matrices of synaptic strengths or weights and a can be understood as a neuronal activation function that models nonlinear summation of presynaptic inputs over the dendritic tree.[19] This means that the synaptic connection to the *i*-th error from the *j*-th state depends on only one parameter, μ_{ij}^{θ}, which changes according to Eq. (3.5)[b]

$$\begin{aligned} \dot{\mu}_{ij}^{\theta} &= \alpha_{ij}^{\theta} - \Pi_{ij}^{\theta}\varepsilon_{ij}^{\theta} \\ \\ \dot{\alpha}_{ij}^{\theta} &= V_{\theta_{ij}}^{\theta} \approx -\tilde{\varepsilon}^T \tilde{\Pi} \tilde{\varepsilon}_{\theta_{ij}} = a_i' \xi_i^T \tilde{\mu}_j \\ \varepsilon_{ij}^{\theta} &= \mu_{ij}^{\theta} - \eta_{ij}^{\theta} \end{aligned} \tag{4.5}$$

[b] The approximate equality follows from the fact we are ignoring uncertainty in the states, when taking the expected instantaneous energy to get the variational energy of the parameters.

This suggests that plasticity due to parameter optimisation comprises an associative term α_{ij}^{θ} and a decay term mediating priors on the parameters[c]. The associative term could be regarded as a synaptic tag,[20] which is simply the covariance between presynaptic input and postsynaptic prediction error, summed over orders of motion. In short, it mediates associative or Hebbian plasticity. The contribution of the product of pre and postsynaptic signals $\xi_i^T \tilde{\mu}_j$ is modulated by an activity-dependent term, a_i', which is the gradient of the activation function at its current level of input (and is constant for linear models). Critically, updating the conditional estimates of the parameters, through synaptic efficacies, μ_{ij}^{θ}, uses local information that is available at each error-unit. Furthermore, the same information is available at the synaptic terminal of the reciprocal connection, where the *i*-th error-unit delivers presynaptic inputs to the *j*-th state. In principle, this enables reciprocal connections to change in tandem. Finally, because plasticity is governed by two coupled ordinary differential equations (4.5), connection strengths should change more slowly than the neuronal states they mediate. These theoretical predictions are entirely consistent with empirical and computational characterisations of plasticity.[20]

4.3 Summary

We have seen that the brain has, in principle, the infrastructure needed to invert hierarchical dynamic models of the sort considered in previous sections. It is perhaps remarkable that such a comprehensive treatment of generative models can be reduced to recognition dynamics that are as simple as Eq. 4.2. Having said this, the notion that the brain inverts hierarchical models speaks to a range of empirical facts about the brain:

- The hierarchical organisation of cortical areas.
- Each area comprises distinct neuronal subpopulations, encoding expected states of the world and prediction error.
- Extrinsic forward connections convey prediction error (from superficial pyramidal cells) and backward connections mediate predictions, based on hidden and causal states (from deep pyramidal cells).[21]
- Recurrent dynamics are intrinsically stable because they are trying to suppress prediction error.[22]
- Functional asymmetries in forwards (linear) and backwards (nonlinear) connections may reflect their distinct roles in recognition.

These observations pertain to the anatomy and physiology of neuronal architectures; see Friston et al[2] for a discussion of operational and cognitive issues. In the next

[c] We have assumed Gaussian priors here; *i.e.*, $p(\theta) = N(\theta : \eta^{\theta}, \Sigma^{\theta})$.

section we consider the sort of empirical evidence that can be garnished in support of this perspective.

5 A simple experiment

In this section, we describe a simple experiment using functional magnetic resonance imaging (fMRI) to measure visually evoked responses at different levels in the visual cortical hierarchy. This experiment was first reported in Harrison *et al.*[23] Here, we focus on its implications for the functional architecture of neuronal computations underlying inference. Specifically, the implementation described in the previous section makes two key predictions. First, differences in responses evoked by predictable and unpredictable stimuli must be mediated by top-down predictions and second, the responses evoked at low levels of the visual hierarchy must, in part, be mediated by the activity of neurons encoding prediction error. In other words, low-level responses should be greater for unpredictable, relative to predictable stimuli.

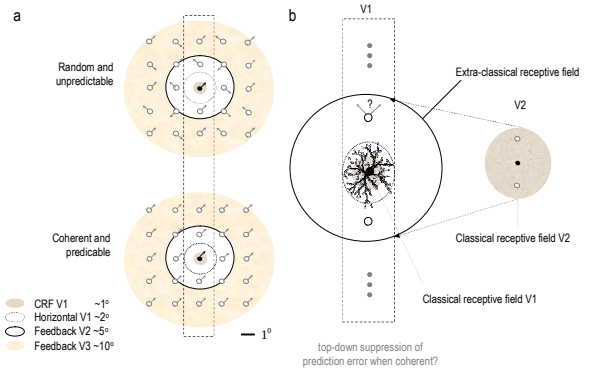

Fig. 2: **(a)** Schematic of the stimuli used in the brain imaging experiment to establish the role of top-down influences in visual recognition. The stimuli comprised random dot arrays, whose motion was either incoherent (upper panel) or coherent (lower panel). Critically the dot stimuli were always separated by more than 3°. This ensured that no two stimuli fell within the classical receptive field of any V1 unit or within the range of its horizontal connections with neighbouring V1 units. This means that any differences in V1 responses to coherent versus incoherent stimuli must be mediated by backward connections from higher areas. **(b)** This schematic quantifies the classical receptive fields of V2 units and shows that their projection to V1 subsumes several dot stimuli. This means, in principle, backward influences from V2 can mediate a sensitivity of V1 responses to coherence.

In our experiment we exploited the known anatomy of intrinsic and extrinsic connections in the visual system to preclude neuronal responses that could be mediated by lateral interactions within the lowest level; namely striate cortex or V1. We did this by presenting moving dot stimuli, where the dots were sufficiently far apart to fall beyond the range of V1 horizontal connections, which extend to about 2 degrees of visual angle. We used predictable and unpredictable stimuli by changing the coherence of the dots' motion. We then simply measured the evoked responses to coherent and

incoherent stimuli in V1 and all other parts of the brain. Fig. 2 provides a schematic detailing the spacing of the dots in relation to the lateral extent of horizontal connections in V1 and the extent of classical and extra-classical receptive fields in V1 and higher areas.

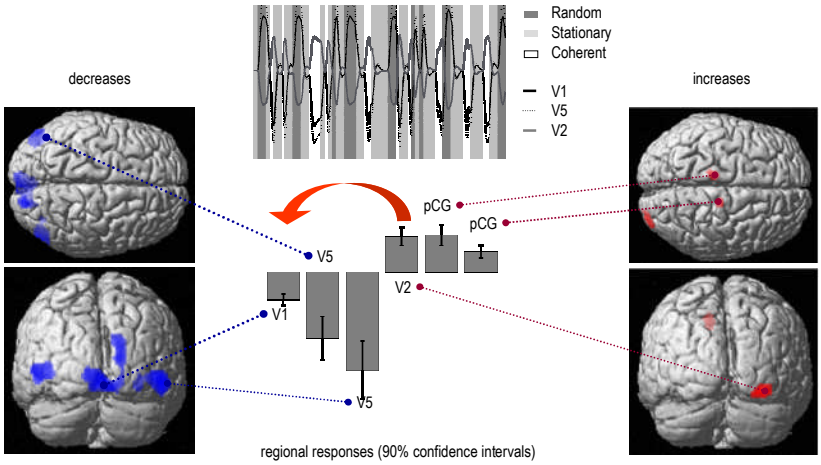

Fig. 3: This is a summary of the results of the fMRI study described in the previous Figure. The upper middle panel shows the time course of activity in three regions (striatal cortex V1; motion sensitive area V5 and second order visual area V2). The shaded bars indicate whether motion was coherent (clear) or random (dark grey). The moving stimuli were interspersed with stationary displays (light grey). A reciprocal activity profile is clearly evident on comparing the dynamics of V1 and V2, with a marked suppression of V1 activity during coherent motion. Left panels: these are statistical parametric maps (SPMs) rendered on the cortical surface showing parts of the brain that exhibited a reduction in activity during predictable or coherent stimulation. The corresponding parameter estimates modulating a regressor encoding the presence of coherent visual motion are provided in the middle panel along with their 90% confidence intervals. Right panels: the corresponding regional activations due to coherence in V2 and posterior cingulate gyrus pCG. The parameter estimates in the middle panel were derived from the peak voxels of each regional effect detailed in the SPMs. See Ref. 23 for a fuller description of these results and the experimental paradigm.

The results of this experiment are shown in Fig. 3 which shows, as predicted, responses in V1 were smaller for predictable coherent stimuli than unpredictable incoherent stimuli. Furthermore, the reverse pattern was seen in the higher cortical area, V2. Interestingly, V5 (a motion sensitive area) behaved like a low-level area with reduced responses to predictable stimuli. This may reflect fast extra-geniculate pathways that deliver subcortical afferents directly to V5. These results have some profound implications for computation in the cortex. First, they show that backward connections mediate evoked responses, even in early visual areas. This is because the sensory input seen by any V1 neuron is exactly the same for the coherent and incoherent stimuli. Because we precluded lateral interactions, the only explanation for differential responses rests upon top-down message passing. This is an important

result because it discounts theories of perceptual processing (although not necessarily elemental sensory processing) that rely only on forward connections. These accounts usually try to maximize the mutual information between the inputs and outputs of any cortical level by optimizing the forward connections. In this view, the visual system represents a series of optimized filters, without recurrent dynamics or self-organisation. In the context of our experimental paradigm, these explanations are clearly inadequate. The second prediction, namely that predictable stimuli enable prediction error to be explained away more efficiently and evoke smaller responses was also confirmed. This is important because it shows that a substantial and measurable proportion of neuronal activity in V1 might be attributable to prediction error. Clearly, V1 is encoding the visual attributes it represents (e.g., in the activity of state units); however, the existence of error units can, in some form, be deduced from these results. This finding challenges any theory of cortical computations that does not include an explicit representation of prediction error. On the other hand, it is exactly consistent with the message passing scheme described above. The only messages required by higher levels for optimal inference are the prediction errors that have not yet been explained away. Although there is no mathematical reason why prediction errors (4.1) should be encoded explicitly by the brain as an auxiliary variable; the physical constraints on message passing in biological systems and the empirical evidence of the sort reported here, suggests that they may be.

5.1 Summary

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

6 Attractors in the brain

In this section, we examine the emergent properties of a system that uses hierarchical dynamics or attractors as generative models of sensory input. We take Walter Freeman's idea and examine the emergent properties of a system that uses attractors as forward models of their 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 the nature and form of this model through simulated lesion experiments. We will then use simplified versions of this model to show how attractors can be used to categorize sequences of stimuli quickly and efficiently. Finally, we will consider perceptual learning of single chirps by simulating a roving mismatch negativity paradigm and looking at the ensuing electrophysiological responses. These examples cover optimisation of states (perceptual inference) and parameters (perceptual learning). Throughout this section, we will exploit the fact that superficial pyramidal cells are the major contributors to observed LFP and ERP signals, which means we can ascribe these signals to prediction error; because the superficial pyramidal cells are the source of bottom-up messages in the brain (see Fig. 1)

6.1 Perceptual Inference

The basic idea here is that the environment unfolds as an ordered sequence of spatiotemporal dynamics, whose equations of motion entail attractor manifolds that contain sensory trajectories. Generally these attractors will support autonomous and probably chaotic dynamics. 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 dynamical systems, 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 generative model has a number of sensible and appealing 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[24] or can exhibit complicated sequences of the sort associated with chaotic and itinerant dynamics.[25,26,27,28,29,30] The notion of attractors as the basis of generative models extends the notion of generalised coordinates, encoding trajectories, to families of trajectories that lie on the attractor manifold; i.e., paths that are contained in the flow-field specified by the control parameters provided by the states of the level above.

Second, hierarchically deployed attractors enable the brain to generate and therefore predict or represent 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 one 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 being perceived and the neuronal activity encodes *where* the current percept is located on the manifold or within the sequence.

Thirdly, 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 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 at any cortical level providing control parameters for attractor dynamics at the level below. This necessarily entails a nonlinear interaction between the top-down effects of the higher attractor and the states of the recipient attractor. Again, this is entirely consistent with the known functional asymmetries between forward and backward connections and speaks to the nonlinear effects of top-down connections in the real brain.

Finally, this particular model has implications for the temporal structure of perception. Put simply, 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 of the higher level having a fixed point attractor, its fixed states will encode lower level dynamics, which could change quite rapidly. We will see an example of this below when considering the perceptual categorisation 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.[31,32,33] One can turn this argument on its head and use the fact that we are able to recognise sequences of sequences[34] as an existence proof for this sort of generative model. In the examples below, we will try to show how autonomous dynamics furnish generative models of sensory input, which behave much like real brains, when measured electrophysiologically.

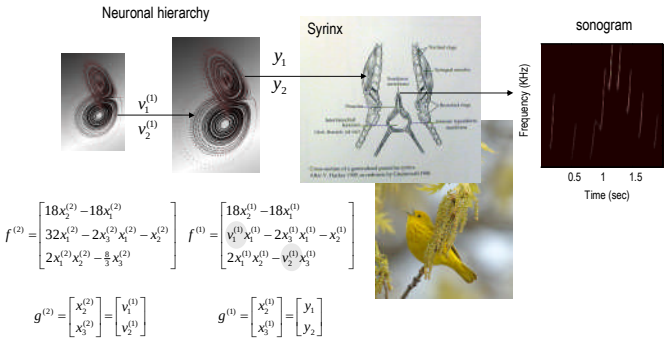

Fig. 4: Schematic showing the construction of the generative model for birdsongs. This comprises two Lorenz attractors where the higher attractor delivers two control parameters (grey circles) to a lower level attractor, which, in turn, delivers two control parameters to a synthetic syrinx to produce amplitude and frequency modulated stimuli. This stimulus is represented as a sonogram in the right panel. The equations represent the hierarchical dynamic model in the form of (2.6).

6.1.1 A synthetic avian brain

The toy example used here deals with the generation and recognition of birdsongs.[35] 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 (see Fig. 4). There has been an extensive modelling effort using attractor models at the biomechanical level to understand the generation of birdsong.[36] Here we use the attractors at a higher level to provide time-varying control over the resulting sonograms. We drive the syrinx with two states of a Lorenz attractor, one controlling the frequency (between two to five KHz) and the other (after rectification) controlling the amplitude or volume. The parameters of the Lorenz attractor were chosen to generate a short sequence of chirps every second or so. To endow the generative model 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 Raleigh and Prandtl number) to control the dynamics exhibited by 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 (about twenty four) 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 (see Fig. 5).

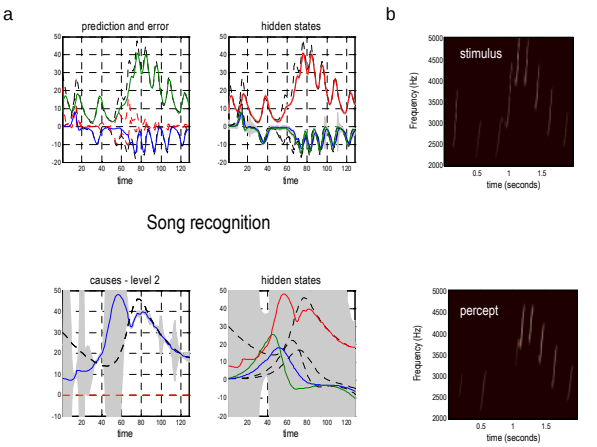

Fig. 5: Results of a Bayesian inversion or deconvolution of the sonogram shown in the previous Figure. **(a)** Upper panels show the time courses of hidden and causal states. Upper left: These are the true and predicted states driving the syrinx and are simple mappings from two of the three hidden states of the first-level attractor. The coloured lines respond to the conditional mode and the dotted lines to the true values. The discrepancy is the prediction error and is shown as a broken red line. Upper right: 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. The panels below show the corresponding causes and hidden states at the second level. Again the conditional expectations are shown as coloured 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)** The stimulus and percept in sonogram format, detailing the expression of different frequencies generated over peristimulus time.

6.1.2 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 (summarised as peak amplitude and volume) using the message-passing scheme described in the previous section. The results are shown in Fig. 5 and demonstrate that, after several hundred milliseconds, the veridical hidden states and supraordinate causes can be recovered. Interestingly, the third chirp is not perceived, in that the first-level prediction error was not sufficient to overcome the dynamical and structural priors entailed by the model. However, once the subsequent chirp had been predicted correctly the following sequence of chirps was recognised with a high degree of conditional confidence. Note that when the second and third chirps in the sequence are not recognised, 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 synthetic bird listening to the song did not know which song was being emitted and was unable to predict subsequent chirps.

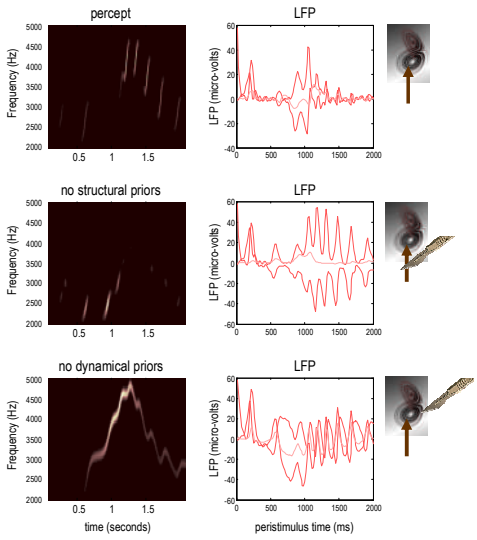

Fig. 6: Results of simulated lesion studies using the birdsong model of the previous figures. The left panels show the percept in terms of the predicted sonograms and the right panels show the corresponding prediction error (at the both levels); these are the differences between the incoming sensory 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. **Top row**: the recognition dynamics in the intact bird. **Middle row**: 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. **Lower panels**: the effects of retaining the structural priors but removing the dynamical priors by cutting the

connections that mediate inversion in generalised coordinates. These results suggest that both structural and dynamical priors are necessary for veridical perception.

6.1.3 Structural and dynamic priors

This example provides a nice opportunity to illustrate the relative roles of structural and dynamic priors. Structural priors are provided by the top-down inputs that dynamically reshape the manifold of the low-level attractor. However, this attractor itself contains an abundance of dynamical priors that unfold in generalised coordinates. Both provide important constraints on the evolution of sensory states, 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 dynamic priors. The latter involves cutting the self-connections in Fig. 1, among the causal and state units. The results of these two simulated lesion experiments are shown in Fig. 6. The top panel shows the percept as in the previous panel, in terms of the predicted sonogram and prediction error at the first and second level. The subsequent two panels show exactly the same information but without structural (middle) and dynamic (lower) priors. In both cases, the synthetic bird fails to recognise the sequence with a corresponding inflation of prediction error, particularly at the last level. Interestingly, the removal of structural priors has a less marked effect on recognition than removing the dynamical priors. 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 completely 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. These simulations address the predictive capacity of the brain. In this example the predictions rests upon the internal construction of an attractor manifold that defines a family of trajectories, each corresponding to the realisation of a particular song. In the next set of simulations we look more closely at the perceptual categorisation of these songs.

6.2 Perceptual categorisation

In the previous simulations, we saw that a song corresponds to a sequence of chirps that are preordained by the shape of an attractor manifold that is 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, recognising or categorising a particular song corresponds to finding a fixed location in the higher state-space. This provides a nice metaphor for perceptual categorisation; 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 categorisation. This can be demonstrated in our synthetic songbird by ignoring the dynamics of the second-level attractor and exposing the bird to a song and letting the states at the second level optimise their location in perceptual space, to best predict the sensory input. To illustrate this, we generated three songs by fixing the Raleigh and Prandtl variables to three distinct values. We then placed uninformative priors on the second-level causes (that were previously driven by the hidden states of the second-level attractor) and inverted the model in the usual way. Fig. 7a shows the results of this simulation for a single song. This song comprises a series of relatively low-frequency chirps emitted every 250 milliseconds or so. The causes of this song (song C in panel b) are recovered after the second chirp, with relatively tight confidence intervals (the blue and green lines in the lower left panel). We then repeated this exercise for three songs. The results are shown in Fig. 7b.

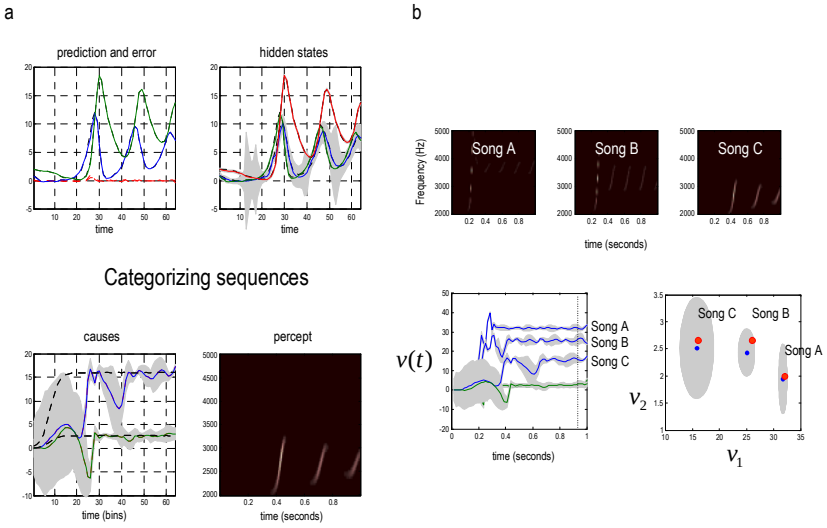

Fig. 7: (a) Schematic demonstration of perceptual categorisation. This figure follows the same format as Figures 3. However, here there are no hidden states at the second level and the causes were subject to stationary and uninformative priors. This song was generated by a first level attractor with 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 about 50 iterations (about 600 milliseconds) before they stabilise. 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 on the right. **(b)** The results of inversion for three songs each produced with three distinct pairs of values for the second level causes (the Raleigh and Prandtl variables of the first level attractor). **Upper panel**: the three songs shown in sonogram format corresponding to a series of relatively high frequency chirps that fall progressively in both frequency and

number as the Raleigh number is decreased. **Lower left**: 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 about 600 milliseconds with high conditional precision. **Lower right**: this shows the conditional density on the causes shortly before the end of peristimulus time (dotted line on the left). 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 categorisation, in the sense that there is no ambiguity (from the point of view of the synthetic bird) about which song was heard.

The songs are portrayed in sonogram format in the top panels and the inferred perceptual causes in the bottom panels. The left panel shows the evolution of these causes for all three songs as a function of peristimulus time and the right shows the corresponding conditional density in the causal or perceptual space of these two states 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 categorisation is almost 100%. This is a simple but nice example of perceptual categorisation, 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.

6.3 Perceptual learning

In the foregoing examples we have looked exclusively at recognition and perceptual inference. In the final simulations, we turn to perceptual learning and the optimisation of parameters or synaptic efficacy. In these examples, we consider that the quantities controlling the attractor manifolds generating sensory trajectories are represented not by supraordinate states but by slowly changing parameters. In the brain, these would correspond to the strength of synaptic connections so that learning corresponds to optimising these strengths to predict input accurately. To illustrate perceptual learning we will use a classical paradigm (the mismatch negativity paradigm) in which stimuli are repeated successively to induce sensory learning. The products of this learning are disclosed by presenting a deviant stimulus that elicits a greater prediction error than the standard stimuli. In this model of the mismatch negativity the difference between the responses evoked by the deviant and standard stimuli correspond to an empirically determined mismatch response that is usually expressed in ERP research as a negative going deflection shortly after the N1 component. Here, we attribute this difference to short-term plasticity that underlies sensory learning on repeated exposure to the same stimulus. The generative model for the stimuli used a simple linear convolution model with two hidden states. This does not show autonomous dynamics but is sufficient for our purposes. The two hidden states are perturbed by a Gaussian bump function (representing a single cause) and express a damped transient. As above, we use one of

the hidden states to modulate the frequency of a chirp and the other to modulate its amplitude (after rectification). This provides a simple model for a single chirp, with a well defined pitch-glide. Fig. 8 shows an example of a standard chirp used in our simulated mismatch negativity paradigm. Here, the cause is equivalent to prediction error at the second level because these had no empirical priors. As above, we assume that the depolarisation of superficial pyramidal cells subtending the EEG reflects prediction error at the appropriate level of the cortical hierarchy and focus on changes in these evoked responses during perceptual learning.

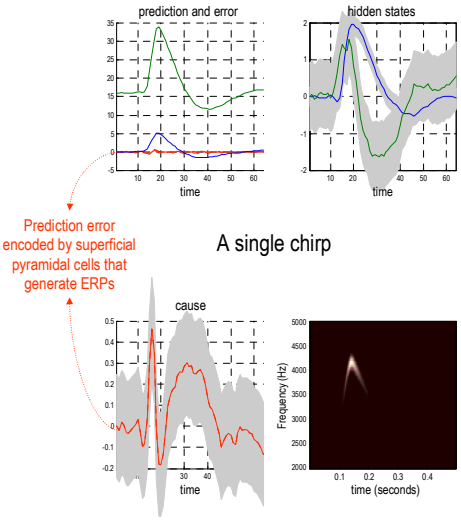

Fig. 8: Results of a deconvolution using a simple linear convolution model of a single chirp. The format of this figure corresponds to Fig. 7. Here, there are no second level hidden states and the second level causes were subject to uninformative and static priors. In this instance, the chirp was generated with a Gaussian bump function to elicit a damped oscillation in two (linearly coupled) hidden states at the first level. These controlled the frequency and amplitude of the chirp shown in sonogram format on the lower right. The prediction error at the first and second levels are shown as solid red lines and are taken to model observed depolarisation in superficial pyramidal cells that dominate observed EEG measurements.

6.3.1 A simulated oddball paradigm

The specific experimental paradigm we simulated corresponds to a roving paradigm where standard stimuli are repeated a small number of times and then one or more of their attributes is changed. In our example, we changed the stimulus by changing the parameters coupling the two hidden states to produce a deviant chirp that was subsequently repeated. After each stimulus we updated the parameters by solving (4.5). The results of this are shown in Fig. 9. The true and predicted hidden states generating the chirps are shown on the left. The sonograms of the corresponding percept are shown in the middle row, clearly indicating the stimulus change after the second chirp. The right column shows the prediction errors (*i.e.*, observed EEG responses) at both the first and second levels. This example highlights a number of important features. First, the distinction between perceptual inference and learning under the free energy formulation can be seen by the progressive suppression of prediction error; both within

the peristimulus time and between trials, over peristimulus time. The elimination of prediction error through self-organising and recurrent dynamics shapes the observed ERP and can be seen on every trial. If we now compare the third and sixth stimuli, we see that there has been a reduction in prediction error between homologous points in peristimulus time. This corresponds to an optimisation of the synaptic connections that models perceptual learning. This occurs despite the fact that the percept is nearly identical between the first and fourth presentations of the new or deviant stimulus. This progressive learning expresses itself in terms of changes in the dynamics of the hidden states that slowly converge to the true values (left column, Fig. 9). From an empirical point of view, the most interesting comparison here is between the first and last presentation of the deviant stimulus. By its fourth presentation this has become a standard stimulus. Therefore, any differences between the can only be explained by perceptual learning. We show these differences in greater detail in Fig. 10.

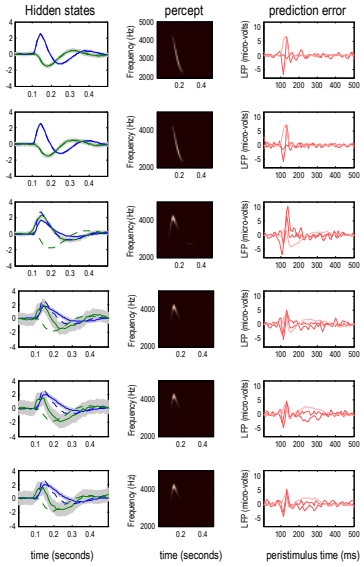

Fig. 9: This shows the results of a simulated roving oddball paradigm, in which trains or sequences of simple stimuli are presented and the stimulus is changed sporadically to elicit an oddball or deviant response. The left hand panels show the evolution of the hidden states at the first level as a function of peristimulus time. These states control the frequency and, after rectification, the amplitude of the chirp; the corresponding percepts as shown in the middle panel. The right hand panels show the evolution of prediction error at the first and second levels, again as a function of peristimulus time. The results here are shown for two occurrences of a previously learned chirp and the first four responses to a new chirp. The new chirp was generated by changing the parameters of the equations of motion of the first level hidden states. The true states are shown as dotted lines and the conditional expectations as solid lines. It can be seen, following the first presentation of the new or oddball stimulus, the hidden states change and converge to the true states. This is due to progressive learning and optimisation of the first-level (control) parameters. This is taken as a model of perceptual learning that is mediated by short-term changes in synaptic efficacy. The concomitant reduction in prediction error at the first and second levels is evident on the right hand panels that demonstrate a suppression of prediction error within trial, over repetition time, and between trials, over repetitions. Of particular interest here is the difference in responses to the first and last presentations of the new stimulus that correspond to the deviant and standard responses respectively. These are considered in more detail in the next figure.

The top panel shows the summed (precision-weighted) prediction error over peristimulus time for all six stimuli. This shows that the first presentation of the deviant

stimuli elicits an enormous prediction error, relative to the others. The evoked responses to the final presentation of the novel stimulus (middle right) correspond to the response to a standard, whereas the response to its first presentation (middle left) corresponds to an oddball or deviant response. The differences between these two responses are shown in the lower panel. These differences correspond to the difference waveforms studied empirically and show and an interesting dissociation. There is evidence of an enhanced negativity at around 100 ms that could correspond to the enhancement of a classical component named the N1 component. This is largely limited to the lower level. Conversely, at the higher level there is a pronounced negativity from about 150 to 250 ms that may be the synthetic homologue of the mismatch negativity. These simulations suggest the interesting possibility that difference waveforms (in mismatch negativity paradigms) may have components that are dissociable into early components that arise in lower cortical sources and later components that are generated by high-level sources. It is interesting to consider how one would model the P300, another classical surprised-related component that is normally associated with changes in high-level attributes of sensory streams. One might imagine that this could be generated by models with a deeper hierarchical structure.

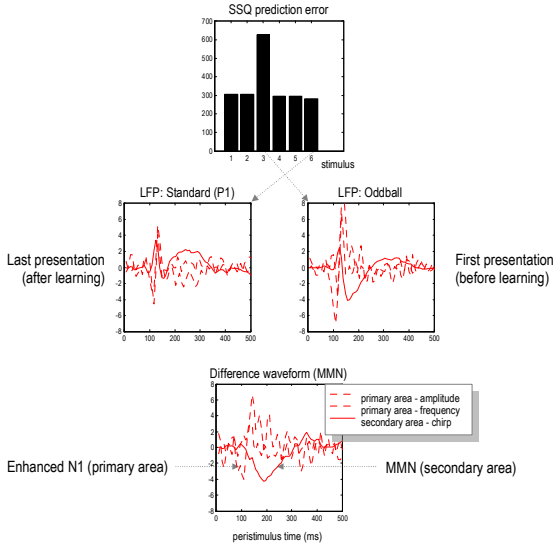

Fig. 10: These show the results of the simulation depicted in the previous figure, with a focus on the responses to the oddball and standard stimuli (first and last presentations of the new stimulus in the previous figure). Upper panel: the sum of squared (precision-weighted) prediction error, accumulated over peristimulus time, for each of the six successive stimuli. It is evident that the presentation of the oddball stimulus elicits a large amount of prediction error, which is rapidly attenuated by perceptual learning, so that

on its second occurrence the prediction error is even lower that the preceding stimulus. Middle panels: these show the prediction error at the first and second level (broken and solid lines respectively) as a function of peristimulus time. Lower panel: this is the difference between the simulated evoked responses to the standard and oddball, showing an enhanced negativity at the first level early in peristimulus time and a later negativity at the higher or second level. These differences could correspond to an enhanced N1 effect and the mismatch negativity found in empirical difference waveforms.

7 Conclusion

We have suggested that the architecture of cortical systems speak to hierarchical generative models in the brain. The estimation or inversion of these models corresponds to a generalised deconvolution of sensory inputs to disclose their causes. This deconvolution could be implemented in a neuronally plausible fashion, where neuronal dynamics self-organised when exposed to inputs to suppress free energy. The focus of this paper has been 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 recognised. 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 spatiotemporal sensory trajectories to points in abstract perceptual spaces.

The ideas presented in this paper have a long history, starting with the notion of neuronal energy;[37] covering ideas like efficient coding[38] and analysis by synthesis[39] to more recent formulations in terms of Bayesian inversion and predictive coding.[22,40,41,42] This work has tried to provide support for the notion that the brain uses attractors to represent and predict causes in the sensorium[43,44,54] first proposed by Walter Freeman two decades ago.

Acknowledgements

The Wellcome Trust funded this work. We would like to thank our colleagues for invaluable discussion about these ideas and Marcia Bennett for helping prepare this manuscript.

Software note

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.

References

1. Freeman WJ. (1987) Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biol Cybern*. **56**(2-3):139-50.
2. Friston KJ. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci*. **360**:815-36.
3. Friston K, Kilner J, Harrison L. (2006). A free energy principle for the brain. *J Physiol Paris.* Jul-Sep;**100**(1-3):70-87.
4. Friston KJ, Trujillo-Barreto N, Daunizeau J (2008). DEM: A variational treatment of dynamic systems. *NeuroImage*. Jul 1;**41**(3):849-85.
5. Efron B and Morris C (1973) Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Am. Stats. Assoc*. **68**:117-130.
6. Kass RE and Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc*. **407**:717-726.
7. Feynman RP (1972). *Statistical mechanics*. Benjamin, Reading MA, USA.
8. Hinton GE and von Cramp D. (1993). Keeping neural networks simple by minimising the description length of weights. In: *Proceedings of COLT*-**93** pp5-13.
9. Maunsell JH and van Essen DC (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.* **3**:2563-86.
10. Zeki S and Shipp S (1988). The functional logic of cortical connections. *Nature* **335**:311-31.
11. Rockland K.S. and Pandya D.N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res*. **179**:3-20.
12. Felleman DJ and Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**:1-47.
13. Angelucci A, Levitt JB, Walton EJ, Hupe JM, Bullier J, Lund JS. (2002) Circuits for local and global signal integration in primary visual cortex. *J Neurosci*. **22**:8633-46.
14. DeFelipe J, Alonso-Nanclares L, Arellano JI. (2002). Microstructure of the neocortex: comparative aspects. *J Neurocytol*. **31**:299-316.
15. Sherman SM, and Guillery RW. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proc Natl Acad Sci* USA **95**:7121-6.
16. Hupe JM, James AC, Payne BR Lomber SG, Girard P and Bullier J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*. **394**:784-7.
17. Rosier AM, Arckens L, Orban GA, Vandesande F (1993) Laminar distribution of NMDA receptors in cat and monkey visual cortex visualized by [3H]-MK-801 binding. *Journal of Comparative Neurology* **335**:369-380.
18. Crick F, Koch C. (1998). Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature*. Jan 15;**391**(6664):245-50.
19. London M. and Häusser M. (2005) Dendritic Computation. *Annual Review of Neuroscience*. **28**:503-532.

20. Martin SJ, Grimwood PD, Morris RG. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu Rev Neurosci* **23**:649-711.
21. Mumford D (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern*. **66**:241-51.
22. Rao RP & Ballard DH (1998) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* **2**:79-87.
23. Harrison LM, Stephan KE, Rees G, Friston KJ. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage*. Feb 1;**34**(3):1199-208.
24. McCrea DA, Rybak IA. (2008). Organization of mammalian locomotor rhythm and pattern generation. Brain Res Rev. Jan;57(1):134-46.
25. Haken H, Kelso JAS, Fuchs A, Pandya AS (1990). Dynamic Pattern-Recognition of Coordinated Biological Motion. *Neural Networks* **3**:395-401.
26. Jirsa VK, Fuchs A, Kelso JA (1998). Connecting cortical and behavioral dynamics: bimanual coordination. *Neural Comput* **10**:2019-2045.
27. Kopell N, Ermentrout GB, Whittington MA, Traub RD (2000). Gamma rhythms and beta rhythms have different synchronization properties. *Proc Natl Acad Sci USA* **97**:1867-1872.
28. Breakspear M, Stam CJ (2005). Dynamics of a neural system with a multiscale architecture. *Philos Trans R Soc Lond B Biol Sci* **360**: 1051-107.
29. Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight RT (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**: 1626-1628.
30. Rabinovich M, Huerta R, Laurent G. (2008). Neuroscience. Transient dynamics for neural processing. *Science*. Jul 4;**321**(5885):48-50.
31. Kiebel S, Daunizeau J., and Friston KJ. (2009). A hierarchy of time-scales and the brain: under review.
32. Botvinick MM. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philos Trans R Soc Lond B Biol Sci*. Sep 29;**362**(1485):1615-26.
33. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. *J Neurosci* **28**:2539-2550.
34. Chait M, Poeppel D, de Cheveigné A, Simon JZ. (2007). Processing asymmetry of transitions between order and disorder in human auditory cortex .*J Neurosci*. May 9;**27**(19):5207-14.
35. Laje R, Mindlin GB (2002). Diversity within a birdsong. *Phys Rev Lett* **89**: 288102.
36. Laje R, Gardner TJ, Mindlin GB (2002). Neuromuscular control of vocalizations in birdsong: a model. *Phys Rev E Stat Nonlin Soft Matter Phys* **65**:051921.
37. Helmholtz H (1860/1962). *Handbuch der physiologischen optik*. (English trans., Southall JPC, Ed.) Vol. 3, New York: Dover.
38. Barlow HB (1961). Possible principles underlying the transformation of sensory messages. *In sensory communication*. Rosenblith. WA ed. MIT press Cambridge MA.
39. Neisser U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.
40. Ballard DH, Hinton GE, Sejnowski TJ (1983) Parallel visual computation. *Nature.* **306**:21-6.
41. Kawato M Hayakawa H and Inui T (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network.* **4**:415-422.

42. Dayan P, Hinton GE and Neal RM (1995) The Helmholtz machine. *Neural Computation* **7**:889-904.
43. Tsodyks M. (1999). Attractor neural network models of spatial maps in hippocampus. *Hippocampus*. **9**(4):481-9.
44. Deco G, Rolls ET. (2003). Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *Eur J Neurosci*. Oct;**18**(8):2374-90.
45. Byrne P, Becker S, Burgess N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychol Rev*. Apr;**114**(2):340-75.