Chapter 9
Policies and Priors

Karl Friston

Abstract This chapter considers addiction from a purely theoretical point of view. It tries to substantiate the idea that addictive behaviour is a natural consequence of abnormal perceptual learning. In short, addictive behaviours emerge when behaviour confounds its own acquisition. Specifically, we consider what would happen if behaviour interfered with the neurotransmitter systems responsible for optimising the conditional certainty or precision of inferences about causal structure in the world. We will pursue this within a rather abstract framework provided by free-energy formulations of action and perception. Although this treatment does not touch upon many of the neurobiological or psychosocial issues in addiction research, it provides a principled framework within which to understand exchanges with the environment and how they can be disturbed. Our focus will be on behaviour as active inference and the key role of prior expectations. These priors play the role of policies in reinforcement learning and place crucial constraints on perceptual inference and subsequent action. A dynamical treatment of these policies suggests a fundamental distinction between *fixed-point policies* that lead to a single attractive state and *itinerant policies* that support wandering behavioural orbits among sets of attractive states. Itinerant policies may provide a useful metaphor for many forms of behaviour and, in particular, addiction. Under these sorts of policies, neuromodulatory (e.g., dopaminergic) perturbations can lead to false inference and consequent learning, which produce addictive and preservative behaviour.

9.1 Introduction

This chapter provides a somewhat theoretical account of behaviour and how addiction can be seen in terms of aberrant perception. Its contribution is not to provide a detailed model of addictive behaviour (see Ahmed et al. 2009 for a nice review of current models) but rather to describe a principled framework that places existing ideas in a larger context. This exercise highlights the archi-

K. Friston (✉)
The Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London WC1N 3BG, UK
e-mail: k.friston@ucl.ac.uk

B. Gutkin, S.H. Ahmed (eds.), *Computational Neuroscience of Drug Addiction*, Springer Series in Computational Neuroscience 10,
DOI 10.1007/978-1-4614-0751-5_9,

tecture of adaptive behaviour, in relation to perception, and the ways in which things can go wrong. Its main conclusion is that addictive behaviour may be an unfortunate and rather unique consequence of a pathological coupling between behaviour (e.g., drug taking) and the perceptual learning (e.g., abnormal modulation of synaptic plasticity) that supports behaviour (cf., Alcaro et al. 2007; Zack and Poulos 2009). This coupling can be particularly disruptive because learning is fundamental for making predictions about exchanges with the world and these predictions prescribe behaviour. In what follows, we will spend some time developing a normative framework for perception and action, with a special emphasis on behavioural policies as prior expectations about how the world unfolds. Having established the basic structure of the problem faced by adaptive agents, we will consider how pathologies of learning manifest behaviourally and show that addictive behaviour is almost impossible to avoid, unless perceptual inference and learning are optimal

This chapter comprises three sections. In Sect 9.2, we review a free-energy principle for the brain. In Sect. 9.3, we focus on a key element of this formulation; namely, prior expectations that reflect innate or epigenetic constraints. In Sect. 9.4, we use the policies from Sect. 9.3 to illustrate failures in learning and behaviour using simulations.

9.2 The Free-Energy Formulation

This section considers the fundaments of normal behaviour using a free-energy account of action and perception (Friston et al. 2006). Its agenda is to establish an intimate relationship between action and perception and to sketch their neurobiological substrates. In brief, we will see that an imperative for all adaptive (biological) agents is to resist a natural tendency to disorder (Evans 2003) by minimising the surprise (unexpectedness) of sensory exchanges with the world. This imperative can be captured succinctly by requiring agents to minimise their free-energy, where free-energy is an upper bound on surprise. When one unpacks this mathematically, minimisation of surprise entails two things. First, it requires an optimisation of perceptual representations of sensory input of the sort implied by the Bayesian brain hypothesis. Second, it requires an active sampling of the sensorium to select sensory inputs that are predicted and predictable. These two facets of free-energy minimisation correspond to perception and action respectively. Basically, we will see that perceptual predictions enslave action to ensure they come true. We will start with a heuristic overview of the free-energy principle and then reprise the basic ideas more formally. By the end of this section we will have expressed perceptual inference, learning and action in terms of ordinary differential equations that describe putative neuronal dynamics underlying active inference. These dynamics can be regarded as a form of evidence accumulation, because free-energy is a bound approximation to log model-evidence. The ensuing scheme rests on internal models of the world used by agents to make predictions. In the subsequent section, we will look at the basic forms that these models can take and the prior expectations about state-transitions (i.e., policies) they entail.

9.2.1 Free-Energy and Self-organisation: Overview

Free-energy is a quantity from information theory that bounds the evidence for a model of data (Feynman 1972; Hinton and van Camp 1993; MacKay 1995). Here, the data are sensory inputs and the model is encoded by the brain. More precisely, free-energy is greater than the negative log-evidence or 'surprise' inherent in sensory data, given a model of how they were generated. Critically, unlike surprise itself, free-energy can be evaluated because it is a function of sensory data and brain states. In fact, under simplifying assumptions (see below), it is just the amount of prediction error.

The motivation for the free-energy principle is simple but fundamental. It rests upon the fact that self-organising biological agents resist a tendency to disorder and therefore minimise the entropy of their sensory states. Under ergodic assumptions, minimising entropy corresponds to suppressing surprise over time. In brief, for a well-defined agent to exist it must occupy a limited repertoire of states (e.g., a fish in water). This means the equilibrium density of an ensemble of agents, describing the probability of finding an agent in a particular state, must have low entropy: A distribution with low entropy just means a small number of states are occupied most of the time. Because entropy is the long-term average of surprise, agents must avoid surprising states (e.g., a fish out of water). But there is a problem; agents cannot evaluate surprise directly because this would require access to all the hidden states in the world causing sensory input. However, an agent can avoid surprising exchanges with the world if it minimises its free-energy, because free-energy is always bigger than surprise.

Mathematically, the difference between free-energy and surprise is the divergence between a probabilistic representation (recognition density) encoded by the agent and the true conditional distribution of causes of sensory input. This enables the brain to reduce free-energy by changing its representation, which makes the recognition density an approximate conditional density. This corresponds to Bayesian inference on unknown states of the world causing sensory data (Knill and Pouget 2004; Kersten et al. 2004). In short, the free-energy principle subsumes the Bayesian brain hypothesis; or the notion that the brain is an inference machine (von Helmholtz 1866; MacKay 1956; Neisser 1967; Gregory 1968, 1980; Ballard et al. 1983; Dayan et al. 1995; Lee and Mumford 2003; Friston 2005). In other words, biological agents must engage in some form of Bayesian perception to avoid surprises. However, perception is only half the story; it makes free-energy a good proxy for surprise but it does not change the sensations themselves or their surprise.

To reduce surprise, we have to change sensory input. This is where the free-energy principle comes into its own: it says that action should also minimise free-energy (Friston et al. 2009, 2010). We are open systems in exchange with the environment; the environment acts on us to produce sensory impressions and we act on the environment to change its states. This exchange rests upon sensory and effector organs (like photoreceptors and oculomotor muscles). If we change the environment

or our relationship to it, sensory input changes. Therefore, action can reduce free-energy (i.e., prediction errors) by changing sensory input, while perception reduces free-energy by changing predictions. In short, we sample the world to ensure our predictions become a self-fulfilling prophecy and that surprises are avoided. In this view, perception enables action by providing veridical predictions (more formally, by making the free-energy a tight bound on surprise) that guide active sampling of the sensorium. This is active inference.

In summary, (i) agents resist a natural tendency to disorder by minimising a free-energy bound on surprise; (ii) this entails acting on the environment to avoid surprises, which (iii) rests on making Bayesian inferences about the world. In this view, the Bayesian brain is mandated by the free-energy principle. Free-energy is not used to finesse perception, perceptual inference is necessary to minimise free-energy. This provides a principled explanation for action and perception that serve jointly to suppress surprise or prediction error; but it does not explain how the brain does this or how it encodes the representations that are optimised. In what follows, we look more formally at what minimising free-energy means for the brain.

9.2.2 Free-Energy and Self-Organisation: Active Inference from Basic Principles

Our objective is to minimise the average uncertainty (entropy) about generalised sensory states $\tilde{s} = s \oplus s' \oplus s'' \ldots \in S$, sampled by a brain or model or the world m ($\oplus$ means concatenation). Generalised states comprise the state itself, its velocity, acceleration, jerk, etc. The average uncertainty is

$$H(S|m) = -\int p(\tilde{s}|m) \ln p(\tilde{s}|m)\, d\tilde{s} \tag{9.1}$$

Under ergodic assumptions, this is proportional to the long-term average of surprise, also known as negative log-evidence $-\ln p(\tilde{s}(t)|m)$

$$H(S|m) \propto -\int_0^T dt \ln p(\tilde{s}(t)|m) \tag{9.2}$$

It can be seen that sensory entropy accumulates negative log-evidence over time. Minimising sensory entropy therefore corresponds to maximising the accumulated log-evidence for an agent's model of the world. Although sensory entropy cannot be minimised directly, we can induce an upper bound $\mathcal{S}(\tilde{s}, q) \geq H(S)$ that can be evaluated using a recognition density $q(t) := q(\vartheta)$ on the generalised causes (i.e., environmental states and parameters) of sensory signals. We will see later that these causes comprise time-varying states $u(t) \subset \vartheta$ and slowly varying parameters $\varphi(t) \subset \vartheta$. This bound is the path-integral of free-energy $\mathcal{F}(t)$, which is created by simply adding a non-negative function of the recognition density to surprise:

$$\mathcal{S} = \int dt \mathcal{F}(t)$$

$$\begin{aligned}\mathcal{F}(t) &= D_{KL}\big(q(\vartheta)\|p(\vartheta|\tilde{s},m)\big) - \ln p\big(\tilde{s}(a)|m\big)\\ &= D_{KL}\big(q(\vartheta)\|p(\vartheta|m)\big) - \big\langle \ln p\big(\tilde{s}(a)|\vartheta,m\big)\big\rangle_q\\ &= \big\langle \ln q(\vartheta)\big\rangle_q - \big\langle \ln p\big(\tilde{s}(a),\vartheta|m\big)\big\rangle_q\end{aligned} \tag{9.3}$$

This non-negative function is a Kullback-Leibler divergence $D_{KL}(q(\vartheta)\|p(\vartheta|\tilde{s},m))$, which is only zero when $q(\vartheta) = p(\vartheta|\tilde{s},m)$ is the true conditional density. This means that minimising free-energy, by optimising $q(\vartheta)$, makes the recognition density an approximate conditional density on sensory causes. The free-energy can be evaluated easily because it is a function of $q(\vartheta)$ and a generative model $p(\tilde{s},u|m)$ entailed by m. One can see this by rewriting the last equality in Eq. (9.3) in terms of $\mathcal{H}(t)$, the neg-entropy of $q(t)$ and an energy $\mathcal{L}(t)$ expected under $q(t)$.

$$\begin{aligned}\mathcal{F}(t) &= \big\langle \mathcal{L}(t)\big\rangle_q - \mathcal{H}(t)\\ \mathcal{L}(t) &= -\ln p\big(\tilde{s}(a),\vartheta|m\big)\\ \mathcal{H}(t) &= -\big\langle \ln q(\vartheta)\big\rangle_q\end{aligned} \tag{9.4}$$

In physics, $\mathcal{L}(t)$ is called Gibb's energy and reports the joint surprise about sensations and their causes. If we assume that the recognition density $q(\vartheta) = \mathcal{N}(\mu,\mathcal{C})$ is Gaussian (the Laplace assumption), then we can express free-energy in terms of the mean and covariance of the recognition density

$$\mathcal{F} = \mathcal{L}(\mu) + \frac{1}{2}\operatorname{tr}(\mathcal{C}\mathcal{L}_{\mu\mu}) - \frac{1}{2}\ln|\mathcal{C}| - \frac{n}{2}\ln 2\pi e \tag{9.5}$$

Where $n = \dim(\mu)$. Here and throughout, subscripts denote derivatives. We can now minimise free-energy with respect to the conditional precision $\mathcal{P} = \mathcal{C}^{-1}$ (inverse covariance) by solving $\partial_\Sigma \mathcal{F} = 0 \Rightarrow \delta_\Sigma \mathcal{S} = 0$ to give

$$\mathcal{F}_\Sigma = \frac{1}{2}\mathcal{L}_{\mu\mu} - \frac{1}{2}\mathcal{P} = 0 \Rightarrow \mathcal{P} = \mathcal{L}_{\mu\mu} \tag{9.6}$$

This allows one to simplify the expression for free-energy by eliminating $\mathcal{C}$ to give

$$\mathcal{F} = \mathcal{L}(\mu) + \frac{1}{2}\ln|\mathcal{L}_{\mu\mu}| - \frac{n}{2}\ln 2\pi \tag{9.7}$$

Crucially, Eq. (9.7) shows that free-energy is a function of the conditional mean, which means all we have worry about is optimising the means or (approximate) conditional expectations. Their optimal values are the solution to the following differential equations. For the generalised states $\tilde{u}(t) \subset \vartheta$

$$\begin{aligned}\dot{\mu}^{(u)} &= \mu'^{(u)} - \mathcal{F}_u\\ \dot{\mu}'^{(u)} &= \mu''^{(u)} - \mathcal{F}_{u'}\\ &\vdots\\ &\triangleq\\ \dot{\tilde{\mu}}^{(u)} &= \mathcal{D}\tilde{\mu}^{(u)} - \mathcal{F}_{\tilde{u}}\end{aligned} \tag{9.8}$$

Where $\mathcal{D}$ is a derivative matrix operator with identity matrices above the leading diagonal, such that $\mathcal{D}\tilde{u} = [u', u'', \ldots]^{\mathrm{T}}$. Here and throughout, we assume all gradients are evaluated at the mean; here $\tilde{u} = \tilde{\mu}^{(u)}$. The stationary solution of Eq. (9.8), in a frame of reference that moves with the generalised motion of the mean, minimises free-energy and its path integral. This can be seen by noting $\dot{\tilde{\mu}}^{(u)} - \mathcal{D}\tilde{\mu}^{(u)} = 0 \Rightarrow \mathcal{F}_{\tilde{u}} = 0 \Rightarrow \delta_{\tilde{u}}\mathcal{S} = 0$. This ensures that when free-energy is minimised the mean of the motion is the motion of the mean: i.e., $\dot{\tilde{\mu}}^{(u)} = \mathcal{D}\tilde{\mu}^{(u)}$. For slowly varying parameters $\varphi(t) \subset \vartheta$, we can use the a formally related scheme, which ensures their motion disappears

$$\begin{aligned} \dot{\mu}^{(\varphi)} &= \mu'^{(\varphi)} \\ \dot{\mu}'^{(\varphi)} &= -\mathcal{F}_{\varphi} - \kappa\mu'^{(\varphi)} \end{aligned} \tag{9.9}$$

Here, the solution $\dot{\tilde{\mu}}^{(\varphi)} = 0$ minimises free-energy, under constraint that the motion of the expected parameters is small: i.e., $\mu'^{(\varphi)} \to 0$. One can see this by noting that when $\dot{\mu}^{(\varphi)} = \dot{\mu}'^{(\varphi)} = 0 \Rightarrow \mathcal{F}_{\varphi} = 0 \Rightarrow \delta_{\varphi}\mathcal{S} = 0$. Equations (9.8) and (9.9) prescribe recognition dynamics for the expected states and parameters respectively. The dynamics for states can be thought of as a gradient descent in a frame of reference that moves with the expected motion of the world (cf., a moving target). Conversely, the dynamics for the parameters can be thought of as a gradient descent that resists transient fluctuations with the damping term $\mathcal{F}_{\varphi'} = \kappa\mu'^{(\varphi)}$ (see Appendix A for a perspective from conventional decent schemes). It is this damping that instantiates prior knowledge that fluctuations in the parameters are small. These recognition dynamics minimise free-energy with respect to the conditional expectations underlying perception but what about action?

9.2.2.1 Action and Perception

The second equality in Eq. (9.3) equality shows that free-energy can also be suppressed by action, through its effects on hidden states and ensuing sensory signals. The key term here is the accuracy term, $\langle \ln p(\tilde{s}(a)|\vartheta, m)\rangle_q$ which, under Gaussian assumptions, this is just the amount of prediction error. This means action should change the motion of sensory states so that they conform to conditional expectations. This minimises surprise, provided perception makes free-energy a tight bound on surprise. In short, the free-energy principle prescribes optimal perception and action

$$\begin{aligned} \mu(t)^* &= \arg\min_{\mu} \mathcal{F}(\tilde{s}(a), \mu) \\ a(t)^* &= \arg\min_{a} \mathcal{F}(\tilde{s}(a), \mu) \end{aligned} \tag{9.10}$$

Action reduces to sampling input that is expected under the recognition density (i.e., sampling selectively what one expects to experience). In other words, agents must necessarily (if implicitly) make inferences about the causes of their sensory signals and sample signals that are consistent with those inferences. In summary, the free-energy principle requires the internal states of an agent and its action to suppress

free-energy. This corresponds to optimising a probabilistic model of how sensations are caused, so that the resulting predictions can guide active sampling of sensory data. The requisite interplay between action and perception (i.e., active inference) ensures the agent's sensory states have low entropy. This recapitulates the notion that "perception and behaviour can interact synergistically, via the environment" to optimise behaviour (Verschure et al. 2003). Active inference is an example of *self-referenced* learning (Maturana and Varela 1980; Porr and Wörgötter 2003) in which "the actions of the learner influence its own learning without any valuation process" (Porr and Wörgötter 2003).

9.2.2.2 Summary

In conclusion, we have derived recognition dynamics for expected states (in generalised coordinates of motion) and parameters, which cause sensory samples. The solution to these equations minimise free-energy and therefore minimise a bound on sensory surprise or (negative) log-evidence. Optimisation of the expected states and parameters corresponds to perceptual inference and learning respectively. The precise form of the recognition dynamics depends on the energy $\mathcal{L} = -\ln p(\tilde{s}, \vartheta | m)$ associated with a particular generative model. In what follows, we consider dynamic models of the world.

9.2.3 Dynamic Generative Models

We now look at hierarchal dynamic models (discussed in Friston 2008) and assume that any sensory data can be modelled with a special case of these models. Consider the state-space model

$$\begin{aligned} s &= f^{(v)}(x, v, \theta) + \omega^{(v)} : \omega^{(v)} \sim \mathcal{N}\big(0, \Sigma^{(v)}(x, v, \gamma)\big) \\ \dot{x} &= f^{(x)}(x, v, \theta) + \omega^{(x)} : \omega^{(x)} \sim \mathcal{N}\big(0, \Sigma^{(x)}(x, v, \gamma)\big) \end{aligned} \tag{9.11}$$

The nonlinear functions $f^{(u)} : u = v, x$ represent a sensory mapping and equations of motion respectively and are parameterised by $\theta \subset \varphi$. The states $v \subset u$ are referred to as sources or causes, while hidden states $x \subset u$ meditate the influence of the causes on sensory data and endow the system with memory. We assume the random fluctuations $\omega^{(u)} \in \Omega$ are analytic, such that the covariance of $\tilde{\omega}^{(u)}$ is well defined. This model allows for state-dependent changes in the amplitude of random fluctuations, which speaks to a key distinction between the effect of states on first and second-order sensory dynamics. These effects are meditated by the vector and matrix functions $f^{(u)} \in \Re^{\dim(u)}$ and $\Sigma^{(u)} \in \Re^{\dim(u)\times\dim(u)}$ respectively, which are parameterised by first and second-order parameters $\theta, \gamma \subset \varphi$. Under local linearity assumptions, the generalised motion of the sensory response and hidden states can be expressed compactly as

$$\begin{aligned}\tilde{s} &= \tilde{f}^{(v)} + \tilde{\omega}^{(v)} \\ \mathcal{D}\tilde{x} &= \tilde{f}^{(x)} + \tilde{\omega}^{(x)}\end{aligned} \tag{9.12}$$

Where the generalised predictions are

$$\tilde{f}^{(u)} = \begin{bmatrix} f^{(u)} = f^{(u)} \\ f'^{(u)} = f_x^{(u)} x' + f_v^{(u)} v' \\ f''^{(u)} = f_x^{(u)} x'' + f_v^{(u)} v'' \\ \vdots \end{bmatrix} \tag{9.13}$$

Equation (9.12) means that Gaussian assumptions about the random fluctuations specify a generative model in terms of a likelihood and empirical priors on the motion of hidden states

$$\begin{aligned} p\big(\tilde{s}|\tilde{x}, \tilde{v}, \theta, m\big) &= \mathcal{N}\big(\tilde{f}^{(v)}, \tilde{\Sigma}^{(v)}\big) \\ p\big(\mathcal{D}\tilde{x}|x, \tilde{v}, \theta, m\big) &= \mathcal{N}\big(\tilde{f}^{(x)}, \tilde{\Sigma}^{(x)}\big) \end{aligned} \tag{9.14}$$

These probability densities are encoded by their covariances $\tilde{\Sigma}^{(u)}$ or precisions $\tilde{\Pi}^{(u)} := \tilde{\Pi}^{(u)}(x, v, \gamma)$ with precision parameters $\gamma \subset \varphi$ that control the amplitude and smoothness of the random fluctuations. Generally, the covariances factorise; $\tilde{\Sigma}^{(u)} = V^{(u)} \otimes \Sigma^{(u)}$ into a covariance proper and a matrix of correlations $V^{(u)}$ among generalised fluctuations that encodes their smoothness. Given this generative model, we can now write down the energy as a function of the conditional means, which has a simple quadratic form (ignoring constants)

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\tilde{\varepsilon}^{(v)T}\tilde{\Pi}^{(v)}\tilde{\varepsilon}^{(v)} - \frac{1}{2}\ln\big|\tilde{\Pi}^{(v)}\big| \\ &\quad + \frac{1}{2}\tilde{\varepsilon}^{(x)T}\tilde{\Pi}^{(x)}\tilde{\varepsilon}^{(x)} - \frac{1}{2}\ln\big|\tilde{\Pi}^{(x)}\big| \\ &\quad + \frac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \frac{1}{2}\ln\big|\tilde{\Pi}^{(\varphi)}\big| \\ \tilde{\varepsilon}^{(v)} &= \tilde{s} - \tilde{f}^{(v)} \\ \tilde{\varepsilon}^{(x)} &= \mathcal{D}\tilde{\mu}^{(x)} - \tilde{f}^{(x)} \\ \tilde{\varepsilon}^{(\varphi)} &= \tilde{\mu}^{(\varphi)} - \tilde{\eta}^{(\varphi)} \end{aligned} \tag{9.15}$$

Here, the auxiliary variables $\tilde{\varepsilon}^{(j)} : j = v, x, \varphi$ are prediction errors for sensory data, the motion of hidden states and parameters respectively. The predictions for the states are $\tilde{f}^{(u)}(\mu)$ and the predictions for the parameters are the prior expectations $\tilde{\eta}^{(\varphi)}$. Equation (9.16) assumes flat priors on the states and that priors $p(\varphi|m) = \mathcal{N}(\tilde{\eta}^{(\varphi)}, \tilde{\Sigma}^{(\varphi)})$ on the parameters are Gaussian, where κ is the precision on the motion of the parameter (see Eq. (9.9)).

9.2.3.1 Perceptual Inference and Predictive Coding

Usually, these models are cast in hierarchical form to make certain conditional independences explicit. Hierarchical forms may look more complicated but they are

simpler than the general form above. They are useful because they provide an empirical Bayesian perspective on inference and learning that may be exploited by the brain. Hierarchical dynamic models have the following form

$$\begin{aligned} s &= f^{(1,v)}\left(x^{(1)}, v^{(1)}, \theta\right) + \omega^{(1,v)} \\ \dot{x}^{(1)} &= f^{(1,x)}\left(x^{(1)}, v^{(1)}, \theta\right) + \omega^{(1,x)} \\ &\vdots \\ v^{(i-1)} &= f^{(i,v)}\left(x^{(i)}, v^{(i)}, \theta\right) + \omega^{(i,v)} \\ \dot{x}^{(i)} &= f^{(i,x)}\left(x^{(i)}, v^{(i)}, \theta\right) + \omega^{(i,x)} \\ &\vdots \end{aligned} \tag{9.16}$$

The random terms $\omega^{(i,u)}$ are conditionally independent and enter each level of the hierarchy. They play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causes $v = v^{(1)} \oplus v^{(2)} \oplus \cdots$ link levels, whereas the hidden states $x = x^{(1)} \oplus x^{(2)} \oplus \cdots$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. This input can enter nonlinearly to produce quite complicated generalised convolutions with deep (hierarchical) structure. If we substitute Eq. (9.16) into the recognition dynamics of Eq. (9.8) (ignoring the derivatives of curvatures and state-dependent noise), we get the following hierarchical message passing scheme

$$\begin{aligned} \dot{\tilde{\mu}}^{(i,v)} &= \mathcal{D}\tilde{\mu}^{(i,v)} + \tilde{f}_{\tilde{v}}^{(i,v)T}\xi^{(i,v)} + \tilde{f}_{\tilde{v}}^{(i,x)T}\xi^{(i,x)} - \xi^{(i+1,v)} \\ \dot{\tilde{\mu}}^{(i,x)} &= \mathcal{D}\tilde{\mu}^{(i,x)} + \tilde{f}_{\tilde{x}}^{(i,v)T}\xi^{(i,v)} + \tilde{f}_{\tilde{x}}^{(i,x)T}\xi^{(i,x)} - \mathcal{D}^{T}\xi^{(i,x)} \\ \xi^{(i,v)} &= \tilde{\Pi}^{(i,v)}\tilde{\varepsilon}^{(i,v)} \\ \xi^{(i,x)} &= \tilde{\Pi}^{(i,x)}\tilde{\varepsilon}^{(i,x)} \\ \tilde{\varepsilon}^{(i,v)} &= \tilde{\mu}^{(i-1,v)} - \tilde{f}^{(i,v)} \\ \tilde{\varepsilon}^{(i,x)} &= \mathcal{D}\tilde{\mu}^{(i,x)} - \tilde{f}^{(i,x)} \end{aligned} \tag{9.17}$$

In neural network terms, Eq. (9.17) suggests that error-units receive messages from the states in the same level and the level above. Conversely, state-units are driven by error-units in the same level and the level below, were $\tilde{f}_{w}^{(i,u)} : u = v, x$ are the forward connection strengths to the state unit representing $w \in \tilde{v}, \tilde{x}$. Critically, recognition requires only the (precision-weighted) prediction error from the lower level $\xi^{(i,v)}$ and the level in question, $\xi^{(i,x)}$ and $\xi^{(i+1,v)}$ (see Fig. 9.1 and Mumford 1992). These constitute bottom-up and lateral messages that drive conditional expectations $\tilde{\mu}^{(i,u)}$ towards a better prediction, which reduces the prediction error in the level below. These top-down and lateral predictions correspond to $\tilde{f}^{(i,u)}$. This is the essence of recurrent message passing between hierarchical levels to optimise free-energy or suppress prediction error (see Friston 2008 for a more detailed discussion). This scheme can be regarded as generalisation of linear predictive coding (Rao and Ballard 1999).

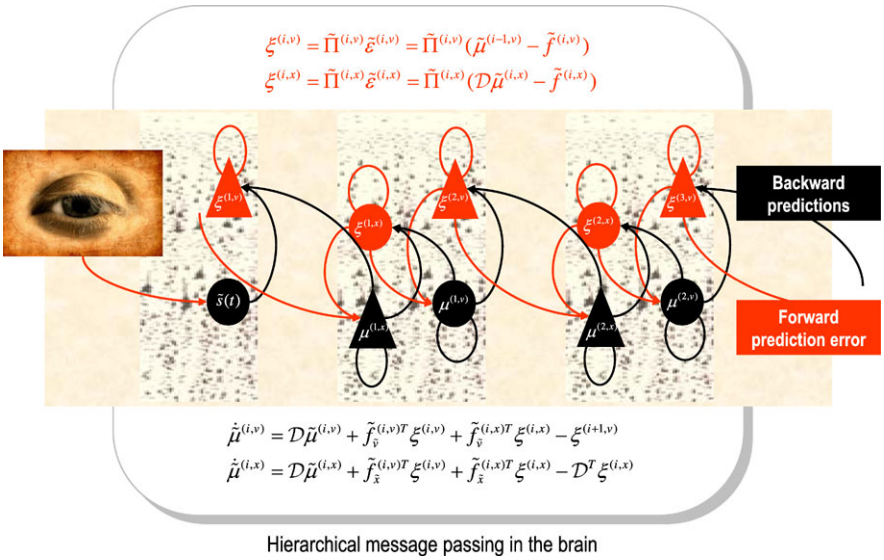

Fig. 9.1 Schematic detailing the neuronal architectures that could encode conditional expectations about the states and parameters of (three levels of) a hierarchical model of the world. This schematic shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and nonlinear backward connections that are used to construct predictions. These predictions try to explain input from lower areas by suppressing prediction error. In this scheme, the sources of forward connections are superficial pyramidal cells and the sources of backward connections are deep pyramidal cells. 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 causal states 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, hidden states. This induces a hierarchy over levels and, within each level, a hierarchical relationship between states, where causes predict the motion of hidden states

Equation (9.17) shows that precision effectively sets the synaptic gain of error-units to their top-down and lateral inputs. Therefore, changes in precision $\tilde{\Pi}^{(i,u)}$ correspond to neuromodulation of error-units encoding precision-weighted prediction error $\xi^{(i,u)}$. This translates as an optimisation of synaptic gain of principal (superficial pyramidal) cells that elaborate prediction error (see Mumford 1992; Friston 2008) and fits comfortably with (among other things) the modulatory effects of dopaminergic and cholinergic neurotransmission. We will exploit this interpretation in the final section. We next consider learning.

9.2.3.2 Perceptual Learning and Associative Plasticity

Perceptual learning corresponds to optimising the first-order parameters $\theta \subset \varphi$. Equation (9.9) describes a process that is remarkably similar to models of associative plasticity based on correlated pre and post-synaptic activity. This can be seen

most easily by assuming an explicit form for the generating functions; for example (for a single parameter and ignoring high-order derivatives)

$$\begin{aligned} f_j^{(i,x)} &= \theta x_k^{(i)} \quad \Rightarrow \\ \dot{\mu}^{(\theta)} &= \mu'^{(\theta)} \\ \dot{\mu'}^{(\theta)} &= -\tilde{\mu}_k^{(i,x)T} \xi_j^{(i,x)} - \Pi^{(\theta)} \mu^{(\theta)} - \kappa \mu'^{(\theta)} \end{aligned} \tag{9.18}$$

Here $\mu^{(\theta)}$ is the connection strength mediating the influence of the k-th hidden state on the motion of the j-th, at hierarchical level $i = 1, 2, \ldots$. This strength changes in proportion to a 'synaptic tag' $\mu'^{(\theta)}$ that accumulates in proportion to the product of the k-th pre-synaptic input $\tilde{\mu}_k^{(i,x)}$ and post-synaptic response $\xi_j^{(i,x)}$ of the j-th error unit (first term of Eq. (9.18)). The tag is auto-regulated by the synaptic strength and decays with first-order kinetics (second and third terms respectively). Crucially, this activity-dependent plasticity rests on (precise) prediction errors that are accumulated by the 'tag'. This highlights the fact that learning (optimising synaptic efficacy) depends on an optimal level of precision encoded by the synaptic gain of error units. Similar equations can be derived for the optimisation of the gain or precision parameters $\gamma \subset \varphi$. However, in this work we will use fixed values and change them to simulate pathology. We conclude this section by examining the dynamics prescribing optimal action.

9.2.3.3 Action

Because action can only affect the free-energy through the sensory data, it can only affect sensory prediction error. If we assume that action performs a gradient descent on free-energy, it is prescribed by:

$$\begin{aligned} \dot{a} &= -\mathcal{F}_a \\ &= -\tilde{\varepsilon}_a^{(v)T} \xi^{(v)} \\ \tilde{\varepsilon}_a^{(v)} &= f_{\tilde{x}}^{(v)} \sum_i \mathcal{D}^{-i} \left(f_{\tilde{x}}^{(x)} \right)^{i-1} f_a^{(x)} \end{aligned} \tag{9.19}$$

The partial derivative of the error with respect to action is the partial derivative of the sensory samples with respect to action. In biologically plausible instances of this scheme, this partial derivative would have to be computed on the basis of a mapping from action to sensory consequences, which are usually quite simple; for example, activating an intrafusal muscle fibre elicits stretch receptor activity in the corresponding spindle (see Friston et al. 2010 for discussion).

9.2.3.4 Summary

In conclusion, we have established some simple dynamics for active inference that implement recognition or perceptual inference, learning and behaviour. However,

we have said nothing about the form of models biological agents might call upon. In the next section, we turn to some fundamental questions about the nature of generative models underlying active inference and, in particular, the role of $f^{(x)}(x, v, \theta)$ in furnishing formal priors on the motion of hidden states in the world.

9.3 Priors and Policies

In this section, we focus on the equations of motion that constitute an agent's generative model of its world. In the previous section, we saw that every agent or phenotype can be regarded as a model of its environment (econiche and internal milieu). Mathematically, this model corresponds to the form of the equations of motion describing hidden states. If these forms are subject to selective pressure, we can regard evolution as optimising formal priors on the environmental dynamics to which each phenotype is exposed. Because these dynamics describe a flow through different states (i.e., state-transitions), they correspond to policies. This section tries to establish the different sorts of priors or policies that might have emerged at an evolutionary scale. It also tries to relate existing formulations (such as optimal control theory, dynamic programming and reinforcement learning) to the dynamical framework that ensues. Briefly, we will see that there are two fundamentally different sorts of policies one could entertain. The first class of (fixed-point) policies can be derived from vector calculus and equilibrium arguments about ensemble densities on the states agents occupy (Birkhoff 1931; Moore 1966; McKelvey and Palfrey 1995; Haile et al. 2008; see Eldredge and Gould 1972 for an evolutionary take on equilibria). These equilibria arguments suggest that the states that are most likely to be occupied (peaks of the ensemble density) require the local policy (flow) to have negative divergence. We will refer to this as the *divergence-constraint*. Mathematically, divergence measures the rate at which flow disperses or dispels a density at any particular point in state-space. This somewhat abstract treatment (and in particular the divergence-constraint) leads to putative policies that ensure attractive states are occupied with the greatest probability. Important examples of these value-based policies are considered in optimal control (Bellman 1952; Sutton and Barto 1981; Todorov 2006) and reinforcement learning (Rescorla and Wagner 1972; Watkins and Dayan 1992; Friston et al. 1994; Montague et al. 1995; Daw and Doya 2006; Daw et al. 2006; Dayan and Daw 2008; Niv and Schoenbaum 2008). This class of policies rests on assuming that all hidden states are equipped with a particular cost, which has the important implication that the optimal flow (prior or policy) has fixed-point attractors. These attract states to low-cost invariant sets; more formally global random attractors $\mathcal{A}(\omega)$, when considering random fluctuations $\omega \in \Omega$ on the states (Matheron 1975; Crauel and Flandoli 1994; Crauel 1999). One of the main purposes of this section is to suggest that although fixed-point policies may provide useful heuristics, they are not necessarily optimal or indeed tenable in a general (dynamical) setting. This is because the external and internal milieu is changing constantly and does not support fixed-point attractors in the state-space of any phenotype. Put simply, any

agent that aspires to a fixed state is doomed, both ethologically and physiologically. To accommodate this, we introduce the notion of itinerant policies, whose implicit attractors are space filling and support wondering (possibly chaotic) trajectories or orbits (e.g., Maturana and Varela 1980; Haken 1983; Freeman 1994; Tsuda 2001; Tyukin et al. 2003; Tschacher and Haken 2007; Tyukin et al. 2009; Rabinovich et al. 2008). Put simply, this means an agent will move through its state-space, sampling different weakly attracting states (attractors in the Milnor sense; Tyukin et al. 2009; Colliaux et al. 2009 or attractor ruins; Rabinovich et al. 2008; Gros 2009) in an itinerant fashion.

The basic idea behind the construction of these itinerant policies (priors) rests on the destruction or vitiation of (weakly) attracting sets. We will focus on attractors that destroy themselves (autovitiate), when they have been occupied too long or other imperatives come into play. This sort of policy will be illustrated with a simple simulation of active inference that leads to exploration and exploitation, under physiologically plausible constraints. The associated model (agent) will be used in the next section to see what would happen if we confound its ability to infer and learn optimally.

9.3.1 Set-up and Preliminaries

The distinction between fixed-point and itinerant policies arises from the following distinction among different subsets of hidden states: $x \supseteq \{x^{(a)}, x^{(p)}, x^{(q)}\}$. This partition acknowledges the fact that, from the agent's perspective, there are two proper disjoint subsets of states. The first comprises those states that can be affected by action $x^{(a)} \subset x$; namely states that support the motion of effectors (e.g., motor plant) and causal (e.g., Newtonian) mechanics in the external milieu. We will call these *physical states*. The other subset $x^{(p)} \subset x \backslash x^{(a)}$ represents states in the internal milieu, which must be maintained within certain bounds (e.g., physiological states that determine interoceptive signals; Davidson 1993). To help remember what these refer to, we will call them *physiological states* and represent the bounds with an indicator or cost-function $c(x^{(p)}) = 0 : x^{(p)} \in \mathcal{A}^{(p)}$ that is zero on the interior of some bounded (attractive or low cost) set $\mathcal{A}^{(p)}$ and one otherwise. Note that the cost-function is defined only on the physiological states. Indeed, one could define the physiological states as the domain of the cost-function. We will use the notion of an indicator or cost-function extensively below for two reasons. First, it is the sort of constraint that can be specified epigenetically and is therefore consistent with the evolutionary perspective above (cf., Traulsen et al. 2006; Maynard Smith 1992). For example, it is not inconceivable that natural selection has equipped us with indicator functions that register when (inferred) blood sugar falls outside the normal 3.6 and 5.8 mM range. Second, utility, loss, or cost-functions are an integral part of optimal control in reinforcement learning and optimal decision (game) theory in economics (e.g., Shreve and Soner 1994; Camerer 2003; Coricelli et al. 2007; Johnson et al. 2007). The remaining hidden states will be called *manifold-states* $x^{(q)} = x \backslash \{x^{(a)}, x^{(p)}\} \subset x$ for reasons that will become clear later.

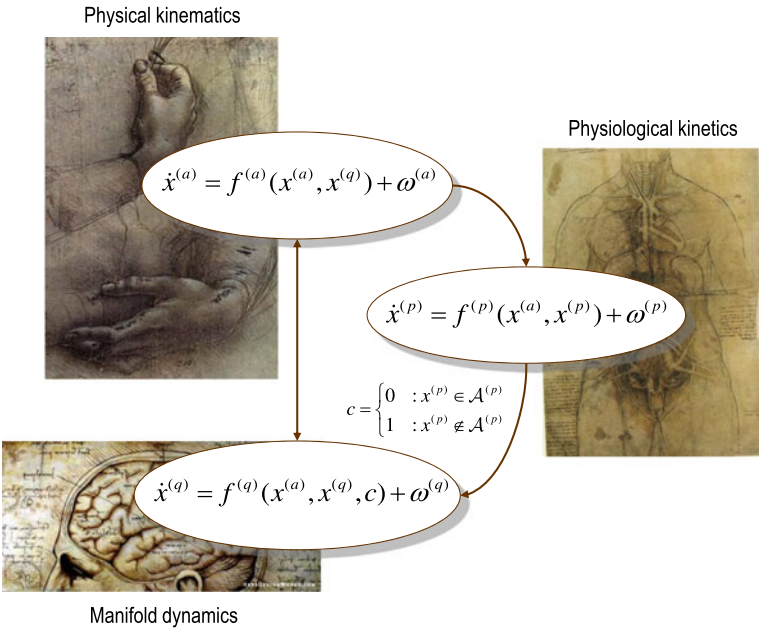

Fig. 9.2 Schematic showing the partition of hidden states into physical states, physiological states, and manifold-states. Physical states correspond, heuristically, to mechanics of the physical world, such as the movement of the motor plant and physical objects. The physiological states pertain to the internal milieu and exhibit kinetics that depend upon physical states. The manifold-states represent the remaining hidden states that govern causal dynamics in the sensorium. These affect (and can be affected by) the physical states but are only affected by the physiological states through indicator or cost-functions reporting whether the physiological states occupy a particular subset: $\mathcal{A}^{(p)}$. The stochastic differential equations describing each partition are a probabilistic summary of their dynamics. The *arrows* represent conditional dependencies and the schematic can be regarded as a Bayesian dependency graph

With this partition in place, we can now consider the conditional dependencies among the subsets. We will assume that physiological states depend on and only on themselves and physical states (e.g., changes in blood sugar after ingestion). The physical states depend upon themselves and manifold-states that shape the manifold that contains the flow of physical states (e.g., forces on manipulanda in the immediate environment). Finally, the manifold-states per se can be influenced by the physical states and physiological states, where the latter influence is mediated by a cost-function. The partition into physical and physiological states means that action cannot affect physiological states directly. This is important and respects the constraints biological agents evolve under. For example, no amount of voluntary (striatal) muscle activity can directly increase blood sugar, it can only do so vicariously by changing physical states that affect physiology. We can summarise these dependencies mathematically with the following equations of motion, which are shown as a dependency graph in Fig. 9.2.

$$f^{(x)} = \begin{bmatrix} f^{(a)}(x^{(a)}, x^{(q)}) \\ f^{(q)}(x^{(a)}, x^{(q)}, c) \\ f^{(p)}(x^{(a)}, x^{(p)}) \end{bmatrix}$$
$$c = \begin{cases} 0 & : x^{(p)} \in \mathcal{A}^{(p)} \\ 1 & : x^{(p)} \notin \mathcal{A}^{(p)} \end{cases} \tag{9.20}$$

These equations of motion are part of the agent's generative model and induce formal priors on state-transitions (i.e., a policy). Our objective now is to find constraints on their form that disclose the nature of implicit policies. Clearly, the only explicit constraint we have is the indicator or cost-function on physiological states. This defines the physiological states the agent expects to be in a priori. In what follows, we will use this cost-function in two distinct ways. First, we will use it to define low-cost attractors in state-space using equilibrium arguments. This requires a rather abstract formulation of the problem, which ignores the distinction between physical and physiological states and leads to conventional (fixed-point) policies. We then reinstate the partition and use indicator or cost-functions to engender flow in the physical space that destroys costly fixed-points in the physiological space. This leads to itinerant policies, which we will use to examine pathological policies in the last section.

9.3.2 Fixed-Point Policies: The Equilibrium Perspective

In this subsection, we will consider policies as prior expectations on flow that lead to low-cost equilibrium densities. This perspective provides a fundamental (divergence) constraint on local flow that can be exploited directly (or is met implicitly) in schemes based upon value; the path-integral of cost. However, to pursue this analysis we need to make a rather severe and implausible assumption. Namely, that we can ignore the conditional dependencies implicit in the partition above and assume that all states can be treated equally. This means the policy reduces to $f := f^{(x)}(x, v, \theta)$. With this simplifying assumption, one can appeal to standard results in vector calculus that describe the evolution of the probability density on the states the agent could occupy as a function of time. This is the ensemble density of the previous section. It can be regarded as either the probability distribution of an infinite number of copies of the agent, observed simultaneously. Alternatively, under ergodic assumptions, this is the same as the probability that an agent will be found in a particular state when observed at different times. This probability is also called the sojourn time and reflects the relative amount of time each state is occupied. The evolution of the ensemble density over time is described by the Fokker-Planck equation

$$\begin{aligned} \dot{p}(x|m) &:= \Lambda p \\ &= \nabla \cdot (\Gamma \nabla - f) p \\ &= \nabla \cdot \Gamma \nabla p - \nabla \cdot (pf) \\ &= \nabla \cdot \Gamma \nabla p - p \nabla \cdot f - f \cdot \nabla p \end{aligned} \tag{9.21}$$

Here Γ is half the amplitude (variance) of random fluctuations on the states. At equilibrium, $\dot{p}(\tilde{x}|m) = 0$ and

$$p(x|m) := p = \frac{\nabla \cdot \Gamma \nabla p - f \cdot \nabla p}{\nabla \cdot f} \tag{9.22}$$

Notice that as the divergence $\nabla \cdot f$ increases, the sojourn time (i.e., the proportion of time a state is occupied) falls. Crucially, at the peaks of the ensemble density, the gradient is zero and its curvature is negative, which means the divergence must be negative (from Eq. (9.22))

$$\left.\begin{aligned} p > 0 \\ \nabla p = 0 \\ \nabla \cdot \nabla p < 0 \end{aligned}\right\} \quad \Rightarrow \quad \nabla \cdot f < 0 \tag{9.23}$$

This divergence-constraint simply says that any policy or flow must have negative divergence at (low cost) maxima of the equilibrium density. One can exploit this constraint by ensuring that all costly fixed-points have positive divergence. Essentially, this destroys any fixed-points in the environment by making them unstable. These policies are easy to construct. For example, the following (Newtonian) policy can be made to satisfy the divergence-constraint very simply by ensuring $\chi(c) \leq 0$, where

$$f = \begin{bmatrix} x' \\ -c\varphi_x(x) + \chi(c)x' \end{bmatrix} \quad \Rightarrow \quad \nabla \cdot f = \chi(c) \tag{9.24}$$

This flow (policy) describes the Newtonian motion of a unit mass in a potential energy well $\varphi(x, \theta)$, where cost plays the role of negative dissipation or friction (and vitiates fixed points in costly regions). Crucially, under this policy, divergence is a function of, and only of, cost. This means the associated ensemble density can only have maxima in regions, where $\chi(c) \leq 0$. Put simply, this ensures that agents are expelled from high-cost regions of state-space and get 'stuck' in attractive (flat) regions. We can illustrate this sort of policy by revisiting a benchmark problem in optimal control:

9.3.2.1 The Mountain-Car Problem

The mountain-car problem can be envisaged as follows: one has to move a car from the bottom of valley and keep it there. However, the car is too heavy to simply drive up the hill. This means that the target can only be accessed by starting on the opposite side of the valley to gain enough momentum to carry it up the other side. This represents an interesting problem, when considered in the state-space of position and velocity, $x, x' \in \tilde{x}$; the agent has to move *away* from the target location ($x = 1$) to attain its goal and execute a very circuitous movement (cf., avoiding obstacles). This problem can be specified with the following equations

$$
\begin{aligned}
\mathbf{g} &= \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{x}}' \end{bmatrix} \\
\mathbf{f} &= \begin{bmatrix} \mathbf{x}' \\ -\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) - \frac{1}{4}\mathbf{x}' + \sigma(a) \end{bmatrix} \\
\boldsymbol{\varphi}_{\mathbf{x}} &= \begin{cases} 2\mathbf{x} + 1 & :\mathbf{x} \leq 0 \\ \mathbf{x}^2(1 + 5\mathbf{x}^2)^{-3/2} + \mathbf{x}^4/16 & :\mathbf{x} > 0 \end{cases}
\end{aligned} \tag{9.25}
$$

We have used bold to highlight the fact that the states and functions are the true values generating sensory data (as distinct from any hidden states assumed by a generative model of these data). Crucially, at $\mathbf{x} = 0$ the force on the car cannot be overcome by the agent, because a squashing function $-1 \leq \sigma(a) \leq 1$ is applied to action to prevent it being greater than one. Divergence-based policies provide a remarkably simple and effective solution to problems of this sort and can be implemented under active inference using policies with the form of Eq. (9.24) (see Friston et al. 2010 for more details). These policies are entailed by the agent's generative model of its sensory inputs. For example,

$$
\begin{aligned}
f^{(v)} &= \begin{bmatrix} x \\ x' \end{bmatrix} \\
f^{(x)} &= \begin{bmatrix} x' \\ -c\varphi_x(x) + \chi(c)x' \end{bmatrix} \\
\varphi_x &= \theta_1(x - \theta_2) \\
\chi &= \frac{1}{4} - 32(1 - c) \\
c &= \begin{cases} 0 & :|x - 1| \leq \Delta \\ 1 & :|x - 1| > \Delta \end{cases}
\end{aligned} \tag{9.26}
$$

Figure 9.3 shows how paradoxical but adaptive behaviour (e.g. moving away from a target to ensure it is secured later) emerges from these simple priors on the motion of hidden states. This example used $\Delta = \frac{1}{16}, \theta_1 \approx 0.6$ and $\theta_2 \approx -0.2$. These simulations of active inference involve integrating the states in the environments (e.g., Eq. (9.25)) and the agent (Eqs. (9.17) and (9.19)) simultaneously as described in Appendix B.

Clearly, the construction of policies that use divergence to vitiate costly fixed-points rests on knowing the form of the policy. In principle, this is no problem, because we are talking about the agent's prior expectations or model of its environment. At no point do we assume that any of the states in the generative model actually exist. For example, the true landscape that exerts forces on a mountain car (Eq. (9.25) and Fig. 9.3) is much more complicated than the agent's model of this landscape, which is a simple quadratic approximation (Eq. (9.26)). This highlights the fact that our expectations about the world and its actual causal structure do not have to be formally equivalent to support adaptive policies. However, it is clearly important that there is a sufficient homology between modelled and experienced causal structure, otherwise the agent will be perpetually surprised by 'obstructions' to its path. This begs the question as to whether there is any universal form of policy

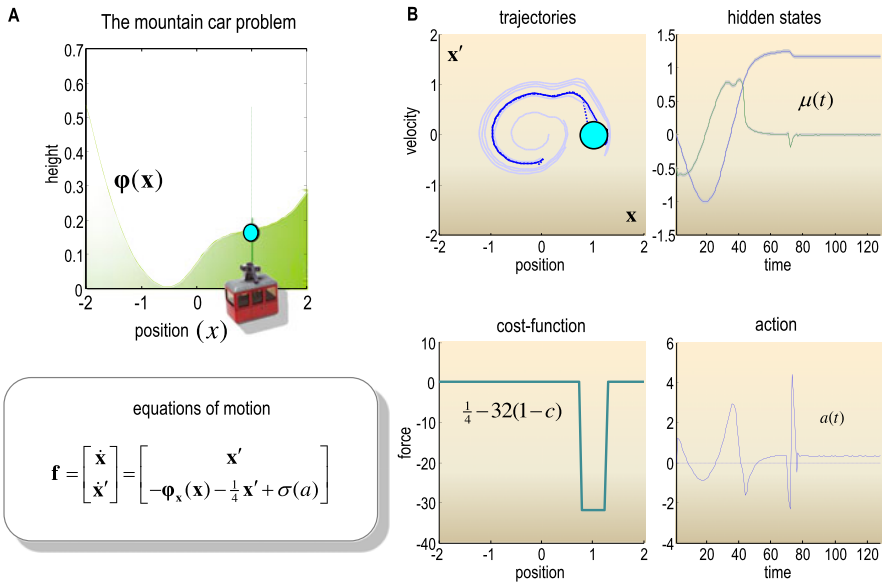

Fig. 9.3 This figure shows how paradoxical but adaptive behaviour (e.g., moving away from a target to ensure it is secured later) emerges from simple priors on the (Newtonian) motion of hidden states in the world. **A**: The *upper panel* shows the landscape or potential energy function (with a minimum at position $x = -0.5$) that exerts forces on a mountain car. The car is shown at the target position on the hill at $x = 1$, indicated by the *cyan ball*. The equations of motion of the car are shown below the figure. Crucially, at $x = 0$ the agent cannot overcome the force on the car because a squashing function $-1 \leq \sigma(a) \leq 1$ is applied to action to prevent it being greater than one. This means that the agent can only access the target by starting halfway up the left hill to gain enough momentum to carry it up the other side. **B**: The results of active inference under priors that destabilise fixed-points outside the target domain. The priors are encoded in a cost-function $c(x)$ (*lower left*), which acts like negative friction. When 'friction' is negative the car expects to go faster. The inferred hidden states (*upper right*: position in *blue* and velocity in *green*) show that the car explores its landscape until it encounters the target. At this point, friction increases (i.e., cost decreases) dramatically to prevent the car from escaping the target (by falling down the hill). The ensuing trajectory is shown in *blue* (*upper left*) in the phase-space of position and velocity. The *paler lines* provide exemplar trajectories from other trials with different starting positions. In the real world, friction is constant. However, the car 'expects' friction to change with its position, enforcing exploration or exploitation. These expectations are fulfilled by action (*lower right*)

that would comply with the divergence-constraint. An example of a universal form is afforded by policies based upon value.

9.3.2.2 Value-Based Policies

In what follows, we consider the key notion of value $V(x)$ as a function of state-space that reports the relative probability or sojourn time a state is occupied at equilibrium. Let flow be decomposed into the gradient of value and an orthogonal com-

ponent $f = \nabla V + \zeta$, such that $\nabla V \cdot \zeta = 0$, where the value of a state is proportional to its log-sojourn time or density at equilibrium

$$\begin{aligned} V = \Gamma \ln p \quad &\Rightarrow \quad p\nabla V = \Gamma \nabla p \\ p := p(x|m) &= \exp(V/\Gamma) \end{aligned} \tag{9.27}$$

Equation (9.27) implies (intuitively) that if ζ is orthogonal to value (log-density) gradients, it must also be orthogonal to the density gradients per se: $\nabla V \cdot \zeta = 0 \Rightarrow \nabla p \cdot \zeta = 0$. If we now substitute Eq. (9.27) into the Fokker-Planck equation (9.21) and solve for the equilibrium density that satisfies $\Lambda p = 0$, we obtain (using standard results from vector calculus)

$$\begin{aligned} \Lambda p = \nabla \cdot (p\nabla V) - \nabla \cdot (p\nabla V) - \nabla \cdot p\zeta = 0 \quad \Rightarrow \\ \nabla \cdot p\zeta = p\nabla \cdot \zeta - \zeta \cdot \nabla p = 0 \quad \Rightarrow \quad \nabla \cdot \zeta = 0 \end{aligned} \tag{9.28}$$

This means that the orthogonal flow $\zeta = \nabla \times W$ is divergence-free and can be expressed in terms of a vector-potential $W(x)$. This is just an example of the Helmholtz decomposition (also known as the fundamental theorem of vector calculus). It means we can express any policy as the sum of irrotational (curl-free) ∇V and solenoidal (divergence-free) $\nabla \times W$ components. If the two components are orthogonal, then the scalar-potential $V(x)$ defines the equilibrium density and its attracting states; that is, the scalar-potential is value. This equivalence rests on the orthogonality condition $\nabla V \cdot \zeta = 0$, which we will call the *curl-constraint*. Under this constraint, curl-free flow prescribed by value counters the change in the equilibrium density due to random fluctuations. Conversely, divergence-free flow follows isoprobability contours and does not change the equilibrium density. Finally, it is easy to show that value is a Lyapunov function for policies that conform to the curl-constraint

$$\begin{aligned} f &= \nabla V + \zeta : \nabla V \cdot \zeta = 0 \\ &= \nabla V + \nabla \times W \\ \dot{V}(x(t)) &= \nabla V \cdot f = \nabla V \cdot \nabla V + \nabla V \cdot \zeta = \nabla V \cdot \nabla V \geq 0 \end{aligned} \tag{9.29}$$

Lyapunov functions increase (or decrease) with time and are used to prove the stability of fixed-points in dynamical systems. This means every policy that satisfies the curl-constraint increases its value as a function of time. The notion of a Lyapunov function is introduced here, because of its relationship to value or attraction in optimal control and decision (game) theory, respectively:

9.3.2.3 Optimal Control and Reinforcement Learning

In optimal control theory and its ethological variants (i.e., reinforcement learning), adaptive behaviour is formulated in terms how agents navigate state-space to access sparse rewards and avoid costly regimes. The aim is to find a (proximal) policy that attains long-term (distal) rewards. In terms of the above, a policy $f = \nabla V + \zeta$ is specified via the scalar-potential or value $V(x)$ also known as (negative) cost-to-go.

In this sense, value is sometimes called a navigation function. The value-function is chosen to minimise expected cost. More formally, the cost-to-go of a state is the cost expected over future states. In the deterministic limit $\Gamma \to 0$, this is just the path integral of cost

$$\begin{aligned} V(x) &= -\int_t^\infty d\tau c(x(\tau)) \quad \Rightarrow \\ \dot{V}(x(t)) &= c(x) = \nabla V \cdot f \geq 0 \end{aligned} \tag{9.30}$$

This says that cost is the rate of increase in value (the Lyapunov function). Crucially, Eq. (9.30) shows that the maxima of the equilibrium density can only exist where cost is zero, at which point value stops increasing and the divergence-constraint is satisfied

$$\begin{aligned} \nabla V(x) = 0 \quad &\Rightarrow \quad c(x) = 0 \\ \nabla \cdot \nabla V(x) < 0 \quad &\Rightarrow \quad \nabla \cdot f < 0 \\ \nabla \cdot f &= \nabla \cdot \nabla V + \nabla \cdot \zeta \\ &= \nabla \cdot \nabla V \end{aligned} \tag{9.31}$$

Heuristically, we can regard value as guiding flow towards points where there is no cost (i.e., no gradients). This means that, in principle, we have a way to prescribe equilibria with maxima (attracting fixed-points) that are specified with a cost-function. Equation (9.30) shows that the cost-function can be derived easily, given the policy and implicit value-function. However, to specify a policy with cost, we have to derive the value-function from the cost-function; that is, solve Eq. (9.30) for value. This is the difficult problem optimal control and value-learning deal with:

In the deterministic limit, the equilibrium density becomes a point mass at the maximum of the value function (see Eq. (9.27)). This is the fixed-point to which all trajectories are attracted. Value-based policies represent universal solutions that do not require any knowledge about the form of the equations of motion generating sensory contingencies. However, this is also their weakness, because we require the solution of Eq. (9.30) under unknown constraints. This leads to the celebrated Hamilton-Jacobi-Bellman equation in optimal control theory (Bellman 1952), for which there is no general solution. However, there is a vast literature on approximate solutions based upon dynamic programming and stochastic iteration. Variants of these schemes appear as temporal difference models (Sutton and Barto 1981) and Q-learning (Watkins and Dayan 1992) in machine learning, and as heuristics in psychological studies of reinforcement learning (Rescorla and Wagner 1972). Almost invariably, these approximate solutions rest on updating explicit representations of the value-function using a prediction error on cost (or reward). This is called a reward prediction error, which we will return to in the discussion. We will not pursue this enormous field here for one simple reason: fixed-point policies are not solutions to real-world problems. This is because there are no valuable fixed-points in dynamical systems: an organism can only occupy a fixed-point when it is frozen or petrified (i.e., dead).

Furthermore, from a technical point of view, value-based (fixed-point) policies are incomplete. This is because real-world (non-abstract) systems do not satisfy

the curl-constraint: Although, the Helmholtz decomposition provides a universal form for policies, with curl and divergence-free components, there is no fundamental lemma or requirement for these components to be orthogonal. This means the scalar-potential is not necessarily a Lyapunov function (i.e., a value-function) or a useful navigation function (see Eq. (9.29)). The interactions among states that violate the curl-constraint are implicit in the conditional dependencies in Eq. (9.20) (for nonlinear equations of motion). In the next subsection, we relax the simplifying assumptions necessary for the abstract formulations used in economics and reinforcement learning and turn to itinerant policies.

9.3.3 Itinerant Policies

In this subsection, we look at functional forms for policies using the (non-abstract) set up that distinguishes between physical, physiological and other hidden states. Here, we consider attractive states that are not fixed-points but bounded sets that arise from itinerant (wandering or searching) dynamics. This is sensible, given the nature of the environment, and speaks to optimising space-filling attractors that ensure low cost equilibria.

The importance of itinerancy has been articulated many times in the past (see Nara 2003), particularly from the perspective of computation and autonomy (see van Leeuwen 2008; with a focus on Milnor attractors). It has also been considered formally in relation to cognition (e.g., Gros 2009, with a focus on attractor relics, ghosts or ruins) and implicitly in ethology (e.g., Panksepp et al. 1984). The ethological perspective is useful here because it suggests that some species are equipped with prior expectations that they will engage in exploratory or social play, For example, 'rough and tumble play' may be a fundamental form of play comprising a unique set of behaviours that can be distinguished from aggression and other childhood activities. Indeed, there is growing interest in understanding brain dynamics per se in terms of itinerancy and metastability (e.g., Jirsa et al. 1994; Breakspear and Stam 2005; Bressler and Tognoli 2006). Tani et al. (2004) consider itinerant dynamics in terms of bifurcation parameters that generate multiple goal-directed actions on the behavioural side, and optimisation of the same parameters when recognising actions. They provide a series of elegant robotic simulations to show generalisation by learning with this scheme. See also Herrmann et al. (1999) for interesting simulations of itinerant exploration, using just prediction errors on sensory samples over time.

We will see below that it is fairly easy to construct itinerant policies. Furthermore, they can have constant (negative) divergence at all points in state-space. This means that their equilibria depend on the divergence-free component of flow (i.e., the component that is discounted by the curl-constraint in fixed-point policies). Although there may not be a universal form for itinerant policies, the principles upon which they are based may be universal.

One universal principle (which we exploit here) is the vitiation or destruction of costly attractors. A key difference between general vitiative mechanisms and the

divergence-based vitiation above is that the destruction of costly attractors can be state and time-dependent. This idea appears in several guises and has found important applications in a number of domains. For example, it is closely related to the notion of autopoiesis and self-organisation in situated (embodied) cognition (Maturana and Varela 1980). It is formally related to the destruction of gradients in synergetic treatments of intentionality (Tschacher and Haken 2007). Mathematically, it is finding a powerful application to universal optimisation schemes (Tyukin et al. 2003) and, indeed, as models of perceptual categorisation (Tyukin et al. 2009). The dynamical phenomena, upon which these schemes rest, involve an itinerant wandering through state-space along heteroclinic channels (orbits connecting different fixed-points). Crucially, these attracting sets are weak (Milnor) attractors or attractor ruins that expel the state until it finds the next weak attractor or ruin. The result is a sequence of transitions through state-space that, in some instances, can be stable and repeating. The resulting stable heteroclinic channels have already been proposed as a metaphor for neuronal dynamics and underlying cognitive processing (Rabinovich et al. 2008). Furthermore, the notion of Milnor or ruined attractors underlies much of the technical and cognitive literature on itinerant dynamics. For example, Tyukin et al. (2009) can explain "a range of phenomena in biological vision, such as mental rotation, visual search, and the presence of multiple time scales in adaptation" using the concept of weakly attracting sets. It is this sort of policy we exploit in the final part of this section.

9.3.3.1 Itinerant Control and Autovitiation

The basic idea is to construct a policy (equations of motion) in which costly states in the physiological subspace change the manifold on which the physical states are evolving. In principle, the only ergodic solution, under this sort of policy, is one in which an attractor (manifold) in the physical subspace induces a low-cost attractor in the physiological subspace. Clearly, this rests upon the existence of such solutions. The mathematical treatment of the existence of these solutions is not necessarily simple. Indeed, it is only recently that the conditions for the existence of stable heteroclinic channels have been established (Rabinovich et al. 2008). Furthermore, even the existence of weakly attracting (Milnor) sets presents some deep challenges (see Tyukin et al. 2003). Generally, attractors are invariant sets that attract states from their neighbourhood, known as a basin of attraction (like a pudding basin that collects its contents at its base). Milnor attractors generalise this notion so that the basin of attraction is not required to be in the neighbourhood of the attractor (like a pudding basin or sieve 'riddled' with holes). This allows the states to escape the attractor when subject to small random fluctuations (like shaking the pudding basin). Attractor ruins result from changing the manifold to destroy an attractor but preserve its characteristic ability to attract trajectories (like a basin with a hole at the base, from which its contents can escape slowly). A key distinction between different sorts of itinerancy is based on whether the manifold supporting itinerant flow is fixed or changing. Milnor attractors and attractor ruins support itinerant dynamics with Type I complexity (Friston 2000); that is, the manifold is invariant. Conversely,

when dynamical systems are coupled to each other, the states of one system can change the manifold (topology or shape of the pudding basin) of another, leading to Type II complexity (Friston 2000). This sort of itinerancy rests on the construction (autopoiesis) and destruction (autovitiation) of attractors in one subspace by changes in the states of another. This is the mechanism we will pursue, given the partition in Eq. (9.20).

We will forego further mathematical discussion and try to illustrate the basic idea with a simple example. This example has been chosen because it embodies autovitiation using intuitive constructs from neurobiology. Consider the following policy

$$\begin{aligned} f^{(x)} &= \begin{bmatrix} f^{(a)} \\ f^{(q)} \end{bmatrix} \\ f^{(a)} &= f^{(a,k)}\left(x^{(a)}\right) : k = \arg\max_i x_i^{(q)} \\ f_i^{(q)} &= h\left(x^{(a)}, x^{(q)}\right) : x^{(a)} \notin \mathcal{A}_i^{(a)} \\ f_i^{(q)} &< h\left(x^{(a)}, x^{(q)}\right) : x^{(a)} \in \mathcal{A}_i^{(a)} \end{aligned} \tag{9.32}$$

This policy describes coupled nonlinear systems in physical $x^{(a)}$ and manifold-subspaces $x^{(q)} = [x_1^{(q)}, \ldots, x_K^{(q)}]$. Physical flow is 'selected' by the (k-th) manifold-state with the highest value, where each alternative flow $f^{(a,k)}(x)$ has a unique attractor $\mathcal{A}_k^{(a)}$. More formally, for all real $t > T$ there exists a time $T \in \Re^+$ for which $x(t)^{(a)} \in \mathcal{A}_i^{(a)}$, under $f^{(a,i)}(x) : i \in 1, \ldots, K$. For each attractor there is a corresponding manifold-state. These change according to some arbitrary function $h(x^{(a)}, x^{(q)})$. Crucially, all the manifold-states experience the same change unless the physical-state occupies the attractor selected by the manifold-state. In this instance, the manifold states decreases, relative to its competitors. The attractor is vitiated when its manifold-state ceases to be the largest and another physical flow supervenes. This is a simple and fairly universal scheme that ensures all the attractors are visited at some point. The key aspect of these schemes is that attractors are destroyed when occupied.

There are clearly many ways that we could have constructed itinerant schemes to illustrate this sort of policy. We elected to use competition among attractors in the physical state-space for several reasons. First, dynamics of this sort can be cast in the abstract form required for conventional value-based policies. This is because the system will visit a discrete number of attractive states $\mathcal{A}_i^{(a)} : i \in 1, \ldots, K$ with well defined probabilities. This will be pursued in a later communication using model-based reinforcement learning. Second, the, saltatory migration from one attractor (pattern) to the next is a ubiquitous phenomenon in neuronal dynamics; manifest as synfire chains (Abeles et al. 2004), reproducible patterns in neuronal avalanches (Pasquale et al. 2008) and 'loss-less' saltatory transitions observed in local field potentials (Thiagarajan et al. 2010). Functionally, the use of attractors with associated basins of attraction, provides a generic way of 'tiling' any space and bears a formal resemblance to classical receptive fields in vision or, indeed, place-cells in spatial

navigation (O'Keefe and Dostrovsky 1971; Sheynikhovich et al. 2009; Robbe and Buzsáki 2009). This means that itinerant policies may furnish a model of saccadic eye movements during exploration of visual scenes (Chen and Zelinsky 2006) or in the context of foraging and spatial exploration. In what follows, we will adopt the second heuristic and associate the attractors $\mathcal{A}_i^{(a)}$ with $i \in 1, \ldots, K$ locations in something like a Morris water-maze (Morris 1984). To emulate conditioned place-preference (e.g., Seip et al. 2008), we have to augment the itinerant scheme above (Eq. (9.32)) with physiological states that can moderate the vitiation of rewarding attractors. For simplicity, we will deal with just four locations and two physiological states.

9.3.3.2 The Generative Model

The particular policy we will focus on for the remainder of this paper is part of the following generative model

$$
\begin{aligned}
s &= f^{(v)} + \omega^{(v)} \\
\dot{x} &= f^{(x)} + \omega^{(x)} \\
f^{(v)} &= \begin{bmatrix} x^{(a)} \\ x'^{(a)} \\ x^{(p)} \end{bmatrix} \\
f^{(x)} &= \begin{bmatrix} f^{(a)} \\ f'^{(a)} \\ f^{(p)} \\ f^{(q)} \end{bmatrix} \\
&= \begin{bmatrix} x'^{(a)} \\ 8(\alpha_k - x^{(a)}) - 4x'^{(a)} \\ \theta^T \beta(x^{(a)}) - x^{(p)} \\ \theta c(x^{(p)}) - 4\beta(x^{(a)}) - \sum_i x_i^{(q)} \end{bmatrix} \quad \Rightarrow \quad \nabla \cdot f = -4 - 1 - K \\
\beta_i &= \begin{cases} 0 & :|\alpha_i - x^{(a)}| \geq \Delta \\ 1 & :|\alpha_i - x^{(a)}| < \Delta \end{cases} \qquad c_j = \begin{cases} 0 & :x_j^{(p)} \geq \tau \\ 1 & :x_j^{(p)} < \tau \end{cases} \\
i &\in 1, \ldots, K, \quad j \in 1, \ldots, J, \quad k = \arg\max_i x_i^{(q)}
\end{aligned}
\tag{9.33}
$$

To complete the specification of this model, we will use the following values (unless otherwise stated): A sensory log-precision of eight $\Pi^{(v)} = 8 \Leftrightarrow \omega_i^{(v)} \sim N(0, e^{-8})$, a log-precision of four or six on the motion of hidden states: $\Pi^{(a)} = 4$, $\Pi^{(p)} = 6$, $\Pi^{(q)} = 4$, a spatial threshold of $\Delta = \frac{1}{8}$ and a physiological threshold of $\tau = \frac{1}{8}$.

The sensory mapping $f^{(v)}$ means that the agent has access to its position and velocity and (in this example) two physiological states $x^{(p)} = [x_1^{(p)}, x_2^{(p)}]^{\mathrm{T}}$ (e.g., blood sugar and osmolarity). The second line describes the policy in terms of formal expectations about the generalised motion of hidden states: The agent assumes

that it pulled to the location, α_k under a degree of friction. This location is the point attractor $\alpha_k \subseteq \mathcal{A}_k^{(a)}$ associated with the highest manifold-state. This means we can regard $x_i^{(q)}$ as the attractiveness of its corresponding location. The manifold-states are subject to three influences, the third is just a non-specific return to zero (mediated by the sum over physiological states). The second mediates itinerancy by vitiating the attractiveness of fixed-points when the agent is in their neighbourhood; i.e., $|\alpha_i - x^{(a)}| < \Delta$. The first makes some locations progressively more attractive, when the cost-function $c(x^{(p)})$ reports that a physiological state has fallen below threshold, $\tau = \frac{1}{8}$. This cost-dependent attractiveness depends on parameters θ_{ij} that encode an association between the j-th physiological-state and the i-th location. These parameters also mediate an increase in the physiological-state—a reward—when the location is occupied, as reported by the indicator function $\beta(x^{(a)})$. In the absence of any reward, the physiological states simply decay with first-order kinetics.

These dynamics mean that when a physiological state falls below threshold this costly state is reported by a (vector) cost-function. This increases the attraction of locations in proportion to a parameterised association between each location and the costly physiological state (cf., Drive Reduction Theory; Hull 1943). The attractiveness of the appropriate location increases until it supervenes over remaining locations, at which point it draws the agent towards it. When the agent is sufficiently close, the physiological state is replenished and the agent is rewarded. This construction of interdependent physical, physiological and manifold dynamics ensures that no physiological state will remain below threshold for long. The ensuing physiological homeostasis depends on physiological imperatives vitiating (non-rewarding) physical attractors. In the absence of any cost (i.e., all physiological states are above some lower bound) all locations will compete with each other, until they are all visited in turn. This is a simple example of a system that shows cost-dependent heteroclinic channels which, in ethological terms includes both exploration and exploitation (e.g., Nowak and Sigmund 1993). Note that the divergence of this policy is a negative constant (see Eq. (9.33)). This means that the self-organising dynamics conform to the divergence constraint but are mediated by changes in divergence-free flow.

9.3.3.3 The Generative Process

Hitherto, we have described the policy as if it were a description of a real environment. However, the policy is just the agent's fantasy about an unknown environment. Crucially, this model is can be much more structured than the environment in which the agent is immersed. The actual generative process we will use can be written as follows.

$$\begin{aligned} \mathbf{f}^{(v)} &= \begin{bmatrix} \mathbf{x}^{(a)} \\ \mathbf{x}'^{(a)} \\ \mathbf{x}^{(p)} \end{bmatrix} \\ \mathbf{f}^{(x)} &= \begin{bmatrix} \mathbf{x}'^{(a)} \\ a - 2\mathbf{x}^{(a)} - 4\mathbf{x}'^{(a)} \\ \boldsymbol{\theta}\beta(\mathbf{x}^{(a)}) - \mathbf{x}^{(p)} \end{bmatrix} \end{aligned} \tag{9.34}$$

Where $\omega_i^{(u)} \sim N(0, e^{-16}) : u = v, x$. Here, the only forces acting upon the agent are those that it generates itself with action. In other words, although the agent has a concept of fixed-points to which it is variously attracted, the environment per se has no such attractors (other than a fixed-point at $\mathbf{x} = 0$). However, a number of the locations do deliver rewards. The mapping between these locations and the rewards is encoded by the (unknown) parameters $\boldsymbol{\theta}_{ij} \in \{0, 1\}$. These play the same role as the parameters of the agent's generative model. If the true parameters and those used by the agent are the same, then the agent will happily navigate its environment alternately visiting rewarding locations to replenish its physiology (e.g., eating and drinking at different locations). However, to achieve this it has to learn the correct parameters. Crucially, this learning is purely perceptual and driven by the prediction errors established by conditional expectations about physiological rewards at every location. This is a key attribute of the current scheme and highlights the critical role of perceptual learning (parameter optimisation) in acquiring and maintaining appropriate policies (cf., conditioned place-preference in animal studies; Seip et al. 2008). We will return to this in the last section.

In summary, we have described an itinerant policy in terms of a generative model that prescribes the motion of physical and physiological states and how they couple to each other. Under active inference, this policy will enslave action to fulfil implicit prior expectations, under the constraints afforded by the real generative process in the environment. To illustrate this, we integrated the differential equations describing active inference from the first section, using the generative process and model above (Eqs. (9.33) and (9.34)). In this example, we used the correct mapping between rewards and locations ($\theta_{ij} = \boldsymbol{\theta}_{ij}$) such that the first location (upper right) replenished the first physiological state and the second location (lower left) replenished the second physiological state. The resulting behaviour is shown in Fig. 9.4. The upper left panel shows the predicted sensory input and its associated prediction errors (dotted red lines). This sensory input corresponds to the position and motion of the agent in two dimensions and the two physiological states. The underlying conditional expectations of these hidden states, which include the manifold-states, are shown on the upper right. The corresponding physical trajectory is shown on the lower left superimposed on the four attractor locations (cyan circles). This trajectory was driven purely by active inference, with the action controlling forces in two dimensions (shown in the lower right panel). The trajectory here shows that the two rewarding locations (upper right and lower left) are visited most frequently, with occasional excursions to the remaining two locations. The numbers by each location represent the percentage of time spent within $\Delta = \frac{1}{8}$ of the location.

Figure 9.5 provides a more detailed description of the conditional expectations about the physiological and manifold (internal) states in the upper panel and the true physiological states in the lower panels. The upper panel shows the expected physiological states (solid lines) and the manifold-states (broken lines). The key thing to take from these time courses is the recurrent build-up and self-destruction of manifold-states, as each attracting fixed-point is visited and consequently rendered less attractive. Crucially, the attractors delivering rewards become more attractive

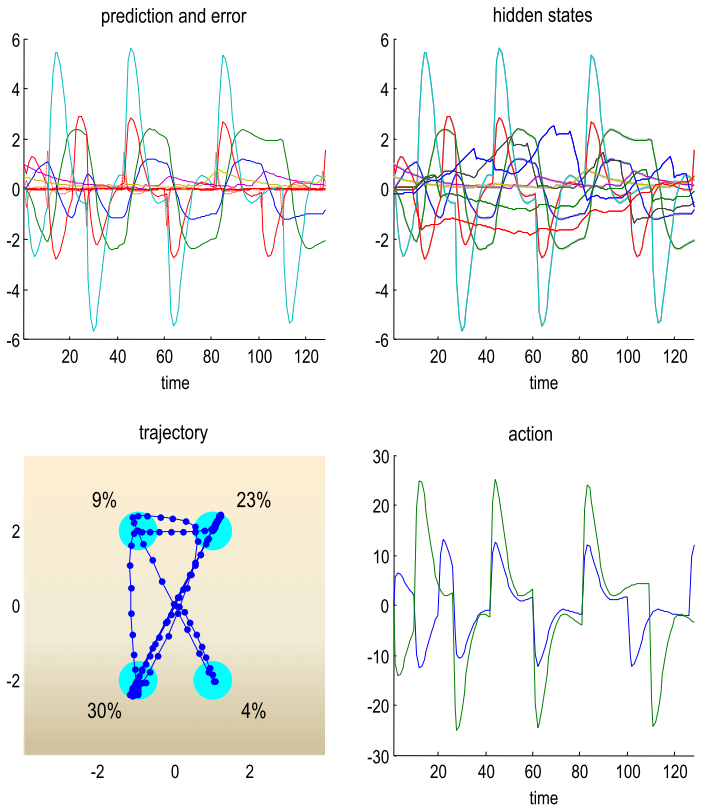

Fig. 9.4 Conditional expectations and behaviour under an itinerant policy. The *upper right panel* shows the conditional expectations of hidden states, while the *upper left panel* shows the corresponding predictions of sensory input (*solid lines*) and prediction errors (*dotted red lines*). Action tries to suppress these prediction errors and is shown on the *lower right*. These action variables exert forces in two orthogonal directions to produce the movements shown on the *lower left*. The ensuing path is shown as a continuous *blue line*, where each dot represents a single time bin in the simulations. The *cyan circles* represent the four attractors used in this itinerant policy. It can be seen that most of the time is spent at the two locations that supply physiological rewards: 23% for the first (*upper right*) and 30% for the second (*lower left*)

after the physiological state falls below some threshold (red dotted line in all panels). This ensures that the physiological states are lower bounded as seen in the lower left panel. This shows the first (blue) and second (green) levels of the physiological variable as a function of time. It can be seen that whenever the level falls below threshold, the values are replenished rapidly by a visit to the appropriate attractor. The same data are shown on the lower right. Here the two physiological states have been plotted against each other to show how they are always (jointly) above or near threshold.

These simulations were integrated as described in Appendix B and (Friston et al. 2010) using log-precisions of eight and four on the sensory input and motion of

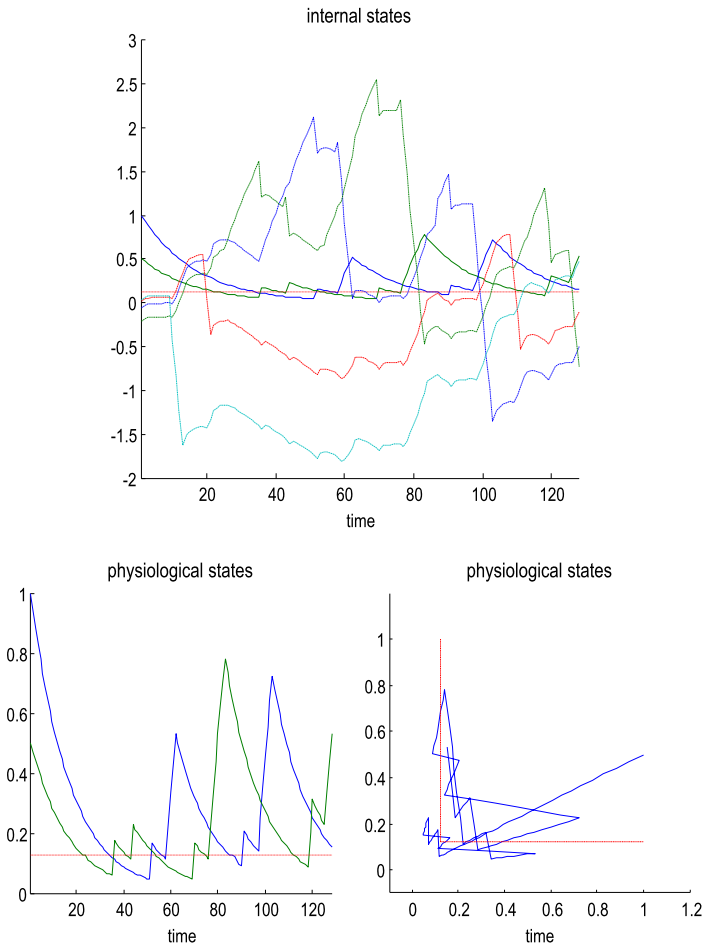

Fig. 9.5 This figure provides a more detailed description of the conditional expectations of the physiological (and manifold) states in the *upper panel* and the true physiological states in the *lower panels*. The *upper panel* shows the expected physiological states (*solid lines*) and the manifold-states (*broken lines*). The key thing to take from these dynamics is the recurrent build up and autovitiation of manifold-states, as each attracting fixed-point is visited and consequently rendered unattractive. Crucially, the attractors delivering rewards become more attractive after the physiological state falls below some threshold (*red dotted lines in all panels*). This ensures that the physiological states are lower bounded, as shown in the *lower left panel*. This shows the levels of first (*blue*) and second (*green*) physiological variables as functions of time. It can be seen that whenever the level falls below threshold, the values are rapidly replenished by a visit to the appropriate attractor. The same data are shown on the *lower right*. Here, the two physiological states have been plotted against each other to show how they are always (jointly) above or near threshold

physical states, respectively. These values are crucial for implementing any policy, as we will see in the next section, where we use low and high precisions to simulate pathological behaviour.

9.3.4 Summary

This section has focused on plausible forms for the motion of hidden states in generative models of the world. These forms correspond to formal priors or policies, which themselves have been optimised (by evolution in a biological setting). We have introduced the distinction between fixed-point and itinerant policies. Fixed-point policies (from optimal control theory and reinforcement learning) can be elaborated using an equilibrium perspective on abstract models of state-space, under constraints on divergence-free flow (the curl-constraint). When this constraint is satisfied, the scalar-potential guiding flow becomes a Lyapunov function and the (log of the) equilibrium density; that is, value. Conversely, itinerant policies are called for when one partitions hidden states into those that can be controlled directly and those which cannot. Both fixed-point and itinerant policies must conform to a divergence-constraint, in that the flow at low-cost points of the equilibrium density must have negative divergence. Furthermore, both sorts of policies rest upon the destruction or vitiation of costly fixed-points (either directly by making divergence or value depend on cost or indirectly using cost-dependent autovitiation). The notion of vitiating attractors to create itinerant dynamics along heteroclinic channels can be exploited in itinerant policies using fairly simple schemes. We have seen an example of one such scheme that will be used in the next section to study some of its key modes of failure.

If you have got this far through the arguments then you must either be very interested, or an editor (or both). Furthermore, you may be thinking "this is all plausible but its just common sense dressed up in the rhetoric of dynamical systems". In one sense this is true; however, it is worth reflecting on what has been achieved: We now have a model of exploratory behaviour and conditioned place-preference that is detailed to the level of forces, friction and physiology, using neurobiologically tenable computations. Furthermore, at no point did we need to invoke any (abstract) reinforcement learning scheme: the only learning required is conventional associative plasticity that is an integral part of perception. In the final section, we will use this model to see how abnormal perceptual inference and learning can have profound effects on behaviour.

9.4 Pathological Policies

In this section, we provide some simple case studies, using simulations to show how behaviour breaks down when perception is suboptimal. Specifically, we will look at the effect of changing the precision of random fluctuations on the hidden states. This may seem a rather arbitrary target for simulated lesions; however, there are some key reasons for starting here. Up until now, we have treated the precisions as known quantities. In more general treatments they are optimised using update or recognition schemes that are not dissimilar to those used for perceptual learning (see Friston 2008). This optimisation of the precisions corresponds to optimising uncertainty

about prediction errors and the consequent predictions. As noted in the first section, precision may be encoded in the post-synaptic gain of prediction error units. The most likely candidates for these prediction error units are the principal (superficial pyramidal) cells originating forward connections in the cortex (see Friston 2008). In the present context, an important determinant of post-synaptic gain is classical neuromodulation. For example, changes in post-synaptic sensitivity due to the effect of dopaminergic or cholinergic neurotransmission on slow conductances following depolarisation. This premise is important in terms of clinical neuroscience because the vast majority of neuropsychiatric disorders are associated with abnormalities in neuromodulatory neurotransmission at one level or another (e.g., Liss and Roeper 2008; Goto et al. 2010). Indeed, the very fact that most psychotropic treatments target these systems testifies to this fact. Furthermore, the drugs most commonly associated with addictive behaviour affect dopaminergic and related classical neuromodulatory systems:

The mesocorticolimbic dopamine (DA) system comprises DA producing cells in the ventral tegmental area (VTA) of the midbrain and projects to forebrain structures including the nucleus accumbens (NAcc), medial prefrontal cortex (mPFC) and amygdala. It is generally thought that this system evolved to mediate behaviours essential for survival (Kelley and Berridge 2002; Panksepp et al. 2002) and that it plays an essential role in mediating biological incentives. Acute exposure to all drugs of abuse directly or indirectly increases DA neurotransmission in the NAcc and repeated drug exposure results in enduring changes in mesocorticolimbic brain regions (Berke and Hyman 2000; Henry and White 1995; Nestler 2005; Pierce and Kalivas 1997). These drugs include psychostimulants (e.g., cocaine, amphetamine and its derivatives methamphetamine and methlyenedioxy methamphetamine), opiates (e.g., heroin and morphine) and other common drugs of abuse (e.g., alcohol and nicotine). Psychostimulants act directly on dopaminergic terminals in the NAcc (Khoshbouei et al. 2003), while opiates act indirectly by inhibiting GABAergic neurons in the VTA with disinhibition of DA neurons.

In what follows, we will repeat the simulations of the previous section but using suboptimal low and high levels of precision on the motion of hidden states. This produces two characteristic failures of behaviour and learning that map, roughly, onto the psychomotor poverty and bradykinesia associated with Parkinson's disease on the one hand and stereotyped perseverative behaviours that are reminiscent of addiction on the other. We first consider the affect of reducing precision.

9.4.1 Simulating Parkinsonism

In the first simulations, we will look at the effects of reducing precision on the motion of hidden states. This can be seen as a crude model of neurodegeneration in ascending dopaminergic systems, which would reduce synaptic gain and precision $\tilde{\Pi}^{(i,u)}$ in Eq. (9.17). To simulate this reduction, we repeated the foraging simulations above, using progressively lower levels of precision on the motion of physical states:

$\Pi^{(a)} \in 4, 2, 0$. The results of these simulations are shown in Fig. 9.6, in terms of the trajectories in physical subspace (left panels) and the physiological subspace (right panels). It is immediately obvious that the accuracy and speed of locomotion is impaired, with a progressive failure to hit the targets and pronounced over-shooting. The physiological sequelae of this impaired behaviour are shown in terms of a progressive failure to keep the physiological states above threshold. Indeed, in the lower right panel, the physiological states are sometimes close to zero.

The reason for this loss of control is simple. Action is driven by sensory prediction errors (see Eq. (9.19)). These prediction errors depend upon precise predictions. If the precision or certainty about the inferred motion of hidden states falls, more weight is placed on sensory evidence. Heuristically, a low precision on the empirical priors afforded by the motion of hidden states means that conditional predictions are based upon sensory evidence. Because action tries to reduce prediction errors it now depends more on what is sensed, as opposed to what is predicted. In the absence of precise predictions, the agent will simply stop moving. We can see the beginnings of this motor poverty in Fig. 9.6 (lower panels), where the forces exerted by action are attenuated, resulting in trajectories with a much lower curvature. If we continued reducing the level of precision (cf., dopamine), the agent would ultimately become akinetic. We have illustrated this behaviour in a variety of simulations previously, for example, the same behaviour can be elicited using the mountain car example in Fig. 9.3, as shown in Friston et al. (2010).

Figure 9.7 shows the action and underlying sensory prediction errors associated with the trajectories in Fig. 9.6. The action (in both directions) is shown as a function of time in the left panels. The right panels show the corresponding prediction error on the four physical states (position and velocity in two directions). The key thing to take from these results is the progressive reduction in the amplitude of action due to an underlying fall in the amplitude of sensory prediction errors. This leads to smaller forces on the physical motion of the agent and the bradykinesia seen in Fig. 9.6. The progressive reduction in sensory prediction errors reflects a loss of confidence (precision) in top-down prior expectations about movement, which would normally subtend itinerant behaviour. This example is used to highlight the key role of precision, especially the precision of predictions about the motion of hidden states. If these predictions become less precise, they have less influence, relative to sensory information and consequently exert less influence over action. In this view, pathologies that involve a loss of neuromodulation can be regarded as subverting the potency of empirical prior expectations that maintain adaptive behaviour.

9.4.1.1 Summary

In summary, we have seen how perceptual synthesis plays a crucial role in providing predictions that action can fulfil. However, if these predictions are under confident, they will fail to elicit sufficient sensory prediction errors to engage behaviour. A key mechanism, by which conditional confidence can be undermined, is false inference about the amplitude of random fluctuations on hidden states. This leads to

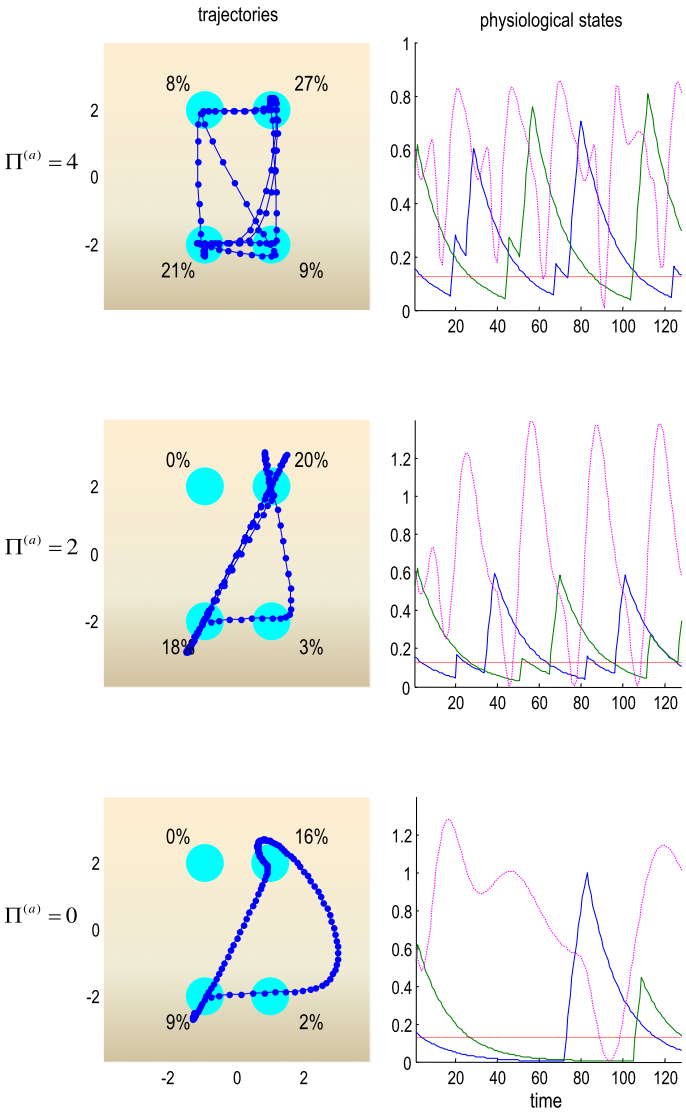

Fig. 9.6 This figure shows the (true) trajectories and resulting physiological states (using the same format as Figs. 9.4 and 9.5) for different levels of precision on the motion of physical states (i.e., position and velocity). The *top row* shows normal behaviour elicited with a log-precision of four. The *remaining two rows* show progressive pathology in behaviour, when using log-precisions of two and zero, respectively. The *left panels* show deterioration of the trajectories, with a generalised slowing of movements and a loss of accuracy, when locating the target (attracting fixed-points). This slowing is reflected in the number of times a target is visited. This is indicated in the *right panels* by the *dotted lines*, which report the distance from the centre. In an extreme case (log-precision of zero), only one definite movement has been emitted in the 128 second simulated exposure. These simulations are meant to reproduce the characteristic psychomotor slowing, bradykinesia and loss of fine movement control associated with Parkinsonism due to neurodegeneration or psycholytic therapy

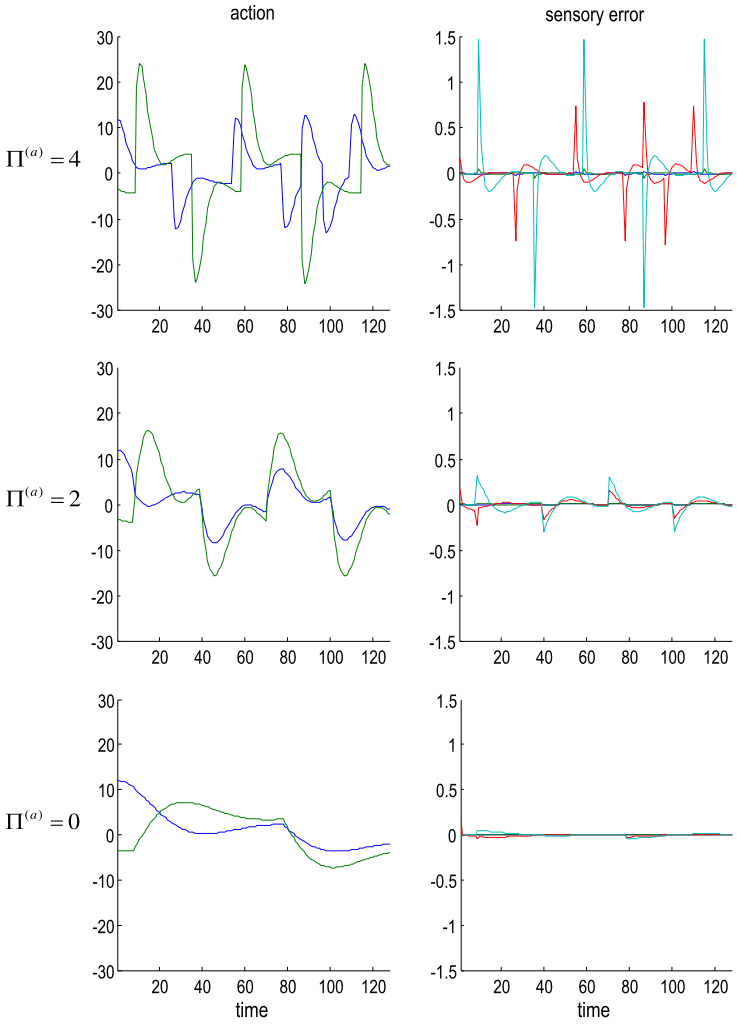

Fig. 9.7 This figure reports the action and underlying sensory prediction errors associated with the trajectories in the previous figure. The action (in both directions) is shown as a function of time in the left column, while the right column shows the corresponding prediction error on the four physical states (position and velocity in two directions). The key thing to take from these results is the progressive reduction in the amplitude of action due to an underlying fall in the amplitude of sensory prediction errors. This leads to smaller forces on the physical motion of the agent and the bradykinesia seen in the previous figure. The reduction in sensory prediction error reflects a loss of confidence (precision) in top–down prior expectations about movements, which would normally subtend itinerant activity

the adoption of pathologically low precision on internal prediction errors that may be associated with the failure of synaptic gain control associated with Parkinsonism (e.g., Zhao et al. 2001). In this context, impaired inference about proprioceptive states translates into a failure of motor intention. This mechanism also sits comfortably with the role of substantia nigra-amygdala connections in surprise-induced enhancement of attention in the perceptual domain: Lesion studies in rats (Lee et al. 2006) show that these connections are "critical to mechanisms by which the coding of prediction error by midbrain dopamine neurons is translated into enhancement of attention and learning modulated by the cholinergic system". Furthermore, low dose apomorphine, which is thought to inhibit DA release by activating pre-synaptic DA autoreceptors, decreases the frequency of itinerant behaviours (e.g., Niesink and Van Ree 1989). Interestingly, increasing precision has relatively little effect on perceptual inference and the attending behaviour; however, it can have a profound effect on perceptual learning. We consider this in the next section, where we ask what would happen if the precision or gain was too high? Here, the consequences are expressed less in terms of locomotion but more in terms of deleterious effects on perceptual learning that determines the organisation of behaviour.

9.4.2 Simulating Addiction

Hitherto, all our simulations have assumed the agent has learned the association between the locations in its environment and the physiological rewards available. These are encoded by the parameters $\theta_{ij} \in \varphi$ in the generative model. In the final simulations, we study how these associations can be acquired and the effects of increasing precision (e.g., dopamine) on this learning.

9.4.2.1 Normal Learning

To study the effects of learning, we changed the reward contingencies by moving the reward usually available at the second location (lower left) to the third location (upper right). This presents an interesting problem under active inference, because action fulfils expectations and the agent expects to be rewarded at the first and second location. It must now undo this association to discover something unexpected, while acting to fulfil its expectations. Itinerant policies meet this challenge easily because, by their construction, they explore all putative reward locations in an itinerant fashion. In brief, the itinerant policy means the agent expects to visit most states at some point and therefore its behaviour will follow suit. This ensures that new associations between the physical and physiological dynamics are encountered and remembered, through optimisation of the parameters encoded by connection strengths (synaptic efficacy). An illustration of perceptual learning under an itinerant policy is shown in Fig. 9.8. This summarises the results of perceptual learning after 128 seconds of exploration, following a switch in the location of the second

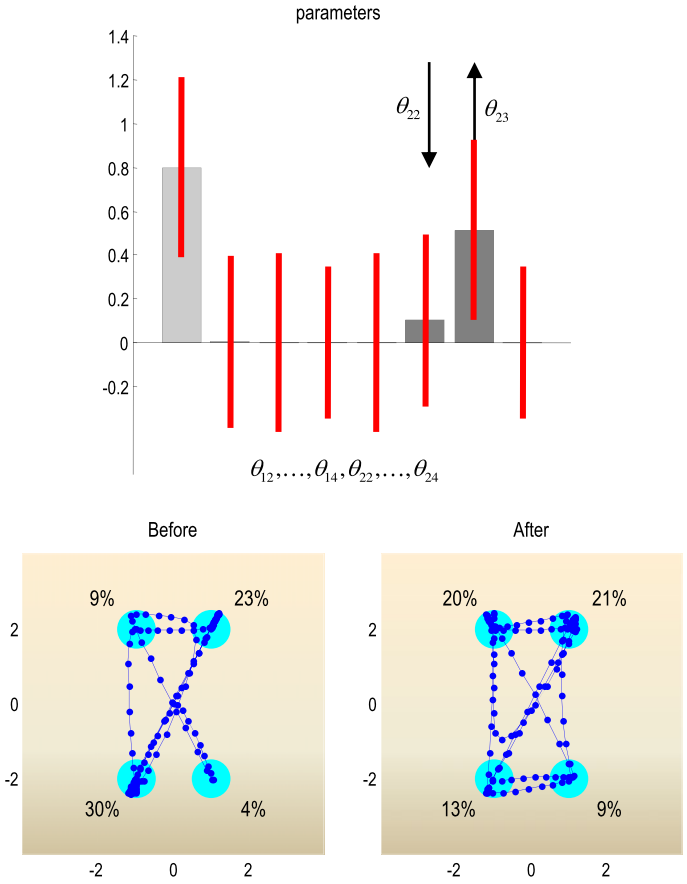

Fig. 9.8 This figure summarises the results of perceptual learning after 128 seconds of exploration, following a switch in the location of the second reward. This location was switched from the lower left to the upper left attractor. The *upper panel* shows the parameter expectations (*grey bars*) and the 90% conditional confidence intervals (*red lines*). The eight parameters constitute the matrix of coefficients θ that associate the two rewards with the four attracting locations. Before learning, rewards were available at the first and second locations (corresponding to parameters one and six). The switch of the location of the second reward corresponds to re-setting the sixth parameter from one to zero $\theta_{22} \rightarrow 0$ with a complimentary increase in the seventh parameter from zero to one $\theta_{23} \rightarrow 1$. The *top panel* shows that the true values are contained within the 90% confidence intervals and a degree of 'reversal learning' has occurred (*arrows* above the parameters in *dark gray*). The corresponding behaviour (before and after learning) is shown in the *lower panels* (*left* and *right* respectively), using the same format as in previous figures. Before learning, the old and new locations of the second reward were visited 30% and 9% of the time respectively. Conversely, after learning this ratio reversed, such that the newly rewarded location is now visited 20% of the time

reward. This can be regarded as a simulation of reversal learning, in the context of conditioned place-preference (McDonald et al. 2002). The reward location was switched from the lower left to the upper left. The upper panel shows the parameter

expectations (grey bars) and the 90% conditional confidence intervals (red bars). It should be noted that these confidence intervals (which are based upon the conditional precisions in Eq. (9.6)), are not represented explicitly by the agent. However, they provide a useful measure of the implicit certainty the agent has in its expectations about causal structure in its world. The eight parameters correspond to the matrix of coefficients $\theta \in \varphi$ that associate the two rewards with the four attracting locations. Before learning, rewards were available at the first and second locations (corresponding to parameters one and six). The switch of the location of the second reward to the third location corresponds to a reduction in the sixth parameter θ_{22} (from one to zero) and a complimentary increase in the seventh parameter θ_{23} (from zero to one). The top panel shows that the true values are contained within the 90% confidence intervals and a degree of reversal learning has occurred. The corresponding behaviour before and after learning is shown in the lower panels (left and right, respectively). Before learning, the old and new locations of the second reward were visited 30% and 9% of the time, respectively. After learning, this ratio has reversed, such that the newly rewarded location is now visited 20% of the time. Note that there is no imperative to spend all the time at a rewarding location; just to emit a sufficient number of visits to ensure the physiological states do not fall to very low levels (data not shown). This learning occurred with a log-precision on the motion of the physiological states of four; $\Pi^{(p)} = 4$. Next, we examine what happens with inappropriately high levels of precision on physiological kinetics.

9.4.2.2 Pathological Learning

We repeated the above simulations but using a pathologically high level of precision that can be thought of (roughly) as a hyper-dopaminergic state. The motivation for this is based on the fact that most addictive behaviours involve taking drugs that cross the blood/brain barrier and augment neuromodulatory transmission. For example, acute exposure to psychostimulants increases extracellular DA levels in the NAcc and this increase is significantly enhanced after repeated exposure; due to increased activity of DA neurons and alterations in DA axon terminals (Pierce and Kalivas 1997). Although a very simplistic interpretation of addiction, we can associate increases in extracellular DA levels with an increase in precision. Intuitively speaking, this means the agent becomes overly confident about its internal predictions, in relation to the sensory evidence encountered. So what effect will this have on learning?

Figure 9.9 reports the results of simulated learning under increasing levels of log-precision on the motion (kinetics) of the physiological states. The left panels show the corresponding behaviour using the same format as in previous figures. The right panels show the conditional expectations and confidence following a 128 second exposure to the environment, after the location of the second reward was switched. The first row reproduces the results of Fig. 9.8 showing veridical, if incomplete, reversal learning (a decrease in parameter six and an increase in parameter seven).

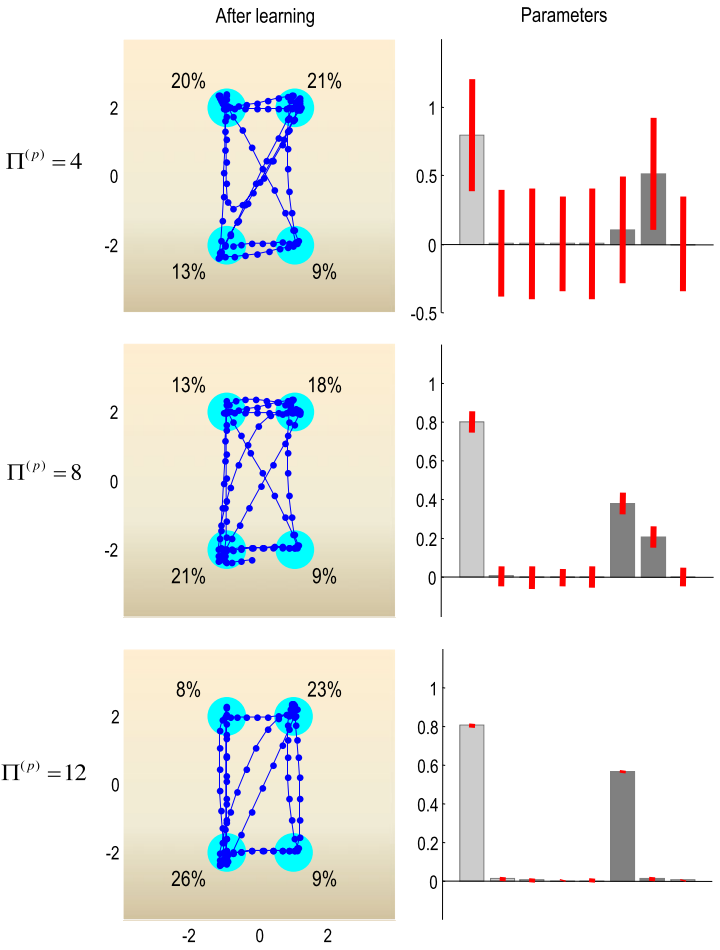

Fig. 9.9 This figure reports the results of simulated learning under increasing levels of log-precision on the motion or dynamics of the two physiological states. The *left column* shows the corresponding trajectories using the same format as in previous figure. The *right column* shows the conditional expectations and confidence intervals following 128 second exposure to the environment, after the location of the second reward had been switched. These use the same format as the *upper panel* of the previous figure. The *first row* reproduces the results of Fig. 9.8 showing veridical, if incomplete, learning of the switched locations (a decrease in parameter six and an increase in parameter seven). This reversal learning is partially (*middle row*) and completely (*lower row*) blocked as log-precision increases from four to eight and from eight to twelve. The failure to learn the change in the association between locations and rewards is reflected in the occupancy of the corresponding locations. For example, the newly rewarding location (*upper left*) is visited on 20%, 13% and 8% of the time as precision increases and learning fails

This learning is partially (middle row) and completely (lower row) blocked as the log-precision increases from four to eight and from eight to twelve. The failure to learn the change in the association between locations and rewards is reflected in the

occupancy of the corresponding locations. For example, the newly rewarding location (upper left) is visited on 20%, 13% and 8% of the time, as precision increases and learning fails. There is a concomitant retention of place-preference for the previously rewarded location (lower left). The reason for this failure of reversal learning and consequent failure to adaptively update place-preference is reflected in the conditional confidence intervals on the parameters. These reveal a progressive reduction in conditional uncertainty (increase in conditional precision), which interferes with learning. The mechanism of this interference is quite subtle but illuminating: Recall from Sect. 9.2 (Eq. (9.18)) that learning (associative plasticity) is driven by the appropriate prediction error, here prediction errors about the motion or changes in physiological states. These are extremely sensitive to the assumed precision about fluctuations in these states as shown in the next figure:

Figure 9.10 shows the conditional expectations or predictions about the motion of physiological states and their associated prediction errors (left and right columns, respectively). The upper rows correspond to a roughly optimal log-precision of four, while the middle and lower rows show the results for pathologically high log-precisions (cf. hyper-dopaminergic states) of 8 and 12, respectively. The corresponding increase in precision means that the conditional representations of changes in physiological state (here the second physiological variable) are over confident and, in extreme cases, a fantasy. This is shown in the left panels in terms of the conditional expectations (solid lines) and the true changes (dotted lines). These are in good agreement for appropriate levels of precision but not at high levels of precision (see lower row). When precision is very high, the agent expects to be rewarded when it visits the old location. This expectation is so precise that it completely ignores sensory evidence to the contrary. These false predictions are reflected in a progressive fall in prediction error (see right column); such that, at high levels of precision, there is no prediction error when there should be. For example, look at the prediction error at around 20 seconds, when the second reward is elicited for the first time. In summary, a high precision leads to over confident inference about the states of the world and their motion, which subverts appropriate prediction errors and their ability to drive associative plasticity. This leads to false expectations about exteroceptive and interoceptive signals and a consequent failure of active inference (behaviour). This example highlights the complicated but intuitive interplay between perceptual inference, learning and action.

9.5 Discussion

In summary, we have seen how inappropriately high levels of precision in generalised predictive coding schemes can lead to false, over confident, predictions that do not properly reflect the true state of the world. This leads to an inappropriately low expression of prediction errors signalled, presumably, by (superficial pyramidal) principal cells in the cortex and a concomitant failure of associative plasticity in their synaptic connections. This failure to learn causal contingencies or associations in the environment results in maladaptive ‘place-preferences’ as reflected

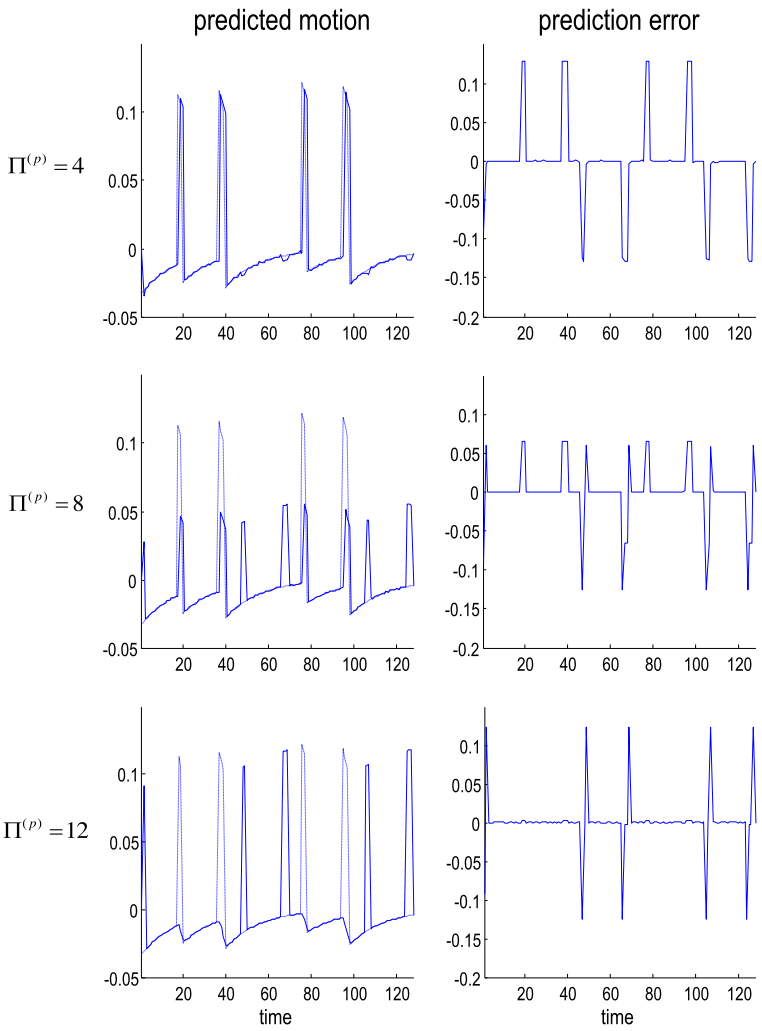

Fig. 9.10 This figure shows the conditional expectations or predictions about the motion of physiological states and their associated prediction errors (*left* and *right columns*, respectively). The *upper rows* correspond to a roughly optimal log-precision of four, while the *middle* and *lower rows* show the results for pathologically high log-precisions (cf., hyper-dopaminergic states) of eight and twelve, respectively. The increase in precision means that the conditional representations of changes in the physiological state (here the second physiological variable) are overconfident and, in extreme cases, illusory. This is shown in the *left panels* in terms of the conditional expectations (*solid lines*) and the true changes (*dotted lines*). These are in good agreement for appropriate levels of precision but represent a 'fantasy' at very high levels of precision (see *lower row*). These overconfident predictions are reflected in a progressive fall in prediction error (see *right column*), such that, at high levels of precision there is no prediction error when there should be. In short, a high precision leads to overconfident inference, which subverts appropriate prediction errors and their ability to drive associative plasticity

in the ensuing perseverative behaviour. This may represent one way in which addictive behaviour could be understood. The implicit explanation for why high levels of precision are maintained in addictive (preservative) behaviour rests upon the assumption that the behaviour per se results in the brain adopting inappropriately high levels of precision. Neurobiologically speaking, this translates into inappropriately high levels of post-synaptic gain in specific neuronal populations. This is consistent with the action of nearly all known drugs of abuse, which affect the mesocorticolimbic dopamine system. Clearly, there can be many ways in which to associate dopaminergic and other neuromodulatory mechanisms with the various parameters and states of predictive coding models. We have chosen to focus on the role of classical neuromodulators in optimising the sensitivity or gain of cells and have equated this with the brain's representation of the precision of random fluctuations in the environment: in other words, a representation of uncertainty. This is certainly consistent with some electrophysiological interpretations of dopaminergic firing, in which phasic dopamine release may represent reward prediction error per se and sustained or tonic firing represents the level of uncertainty (Fiorillo et al. 2003). For example, prediction error on the physiological states could be encoded by phasic discharges in the dopaminergic system, whereas the post-synaptic gain of DA error units may be influenced by (or cause) tonic discharge rates.

Traditionally, midbrain dopamine neurons in the substantia nigra and ventral tegmental area (VTA) are thought to encode reward prediction error (Montague et al. 1996; Schultz et al. 1997; Schultz 1998; Salzman et al. 2005). Activity in these neurons reflects a mismatch between expected and experienced reward that emulates the prediction errors used in (abstract) value-learning theories (Friston et al. 1994; Montague et al. 1996; Sutton and Barto 1981). Indeed, aberrant reward prediction error accounts have proposed for addictive behaviour (Lapish et al. 2006; Redish 2004) and the maintenance of maladaptive habits (Takahashi et al. 2008). However, recent studies suggest a diverse and multilateral role for dopamine that is more consistent with encoding the precision of generalised prediction errors in the predictive coding sense (as opposed to reward prediction errors in particular). For example, punishment prediction error signals (Matsumoto and Hikosaka 2009) and mismatches between expected and experienced *information* (Bromberg-Martin and Hikosaka 2009) may be encoded in distinct anatomical populations of midbrain dopamine neurons. Furthermore, the timing of reward-related signals in VTA precludes the calculation of a reward prediction error per se (Redgrave and Gurney 2006) and may report a change in the certainty about sensory events, via cholinergic input from the pedunculopontine tegmentum (Dommett et al. 2005). Similarly, violations of perceptual expectations engage hippocampal projections to the VTA, which modulate a broad population of dopamine neurons (Lodge and Grace 2006). Human studies with functional neuroimaging suggest that the ventral striatum responds to non-rewarding, unexpected stimuli in proportion to the salience of the stimulus (Zink et al. 2006), as well as to novel stimuli (Wittmann et al. 2007). One of the proposed functions of these striatal responses is to reallocate resources to unexpected stimuli in both reward and non-

reward contexts (Zink et al. 2006). Hsu et al. (2005) show that "the level of ambiguity in choices correlates positively with activation in the amygdala and orbitofrontal cortex, and negatively with a striatal system" and interpret their findings in terms of a "neural circuit responding to degrees of uncertainty, contrary to decision theory". These results suggest that rather than just coding reward prediction errors, the striatum may have a more general role in processing salient and unexpected events, under varying degrees of ambiguity or uncertainty (precision). In summary, the mesocorticolimbic dopamine system may encode numerous types of expectation violations associated with a change in the precision of top-down predictions and ensuing prediction errors (see also Schultz and Dickinson 2000; Fiorillo 2008).

Perhaps one thing to take from these considerations is the complex but intuitive interplay between the many variables that need to be encoded by the brain for optimal behaviour. This means that it may not be easy, given the present state of knowledge, to associate the algorithmic components of optimal schemes with specific neurotransmitter systems or their kinetics. Having said this, there are obvious commonalities between the dynamical simulations presented above and the more abstract formulations that rest on things like the Rescorla-Wagner model (Rescorla and Wagner 1972) and dynamic programming. All these formulations highlight the importance of prediction error on physiological states normally associated with reward. This has been nuanced in the current formulation by a focus on the precision of this prediction error as opposed to the prediction error per se. As we have noted previously, it may be that dopamine does not encode the prediction error on value but the value (precision) of prediction error. The motivation for this perspective rests on the empirical observations discussed above and, more theoretically, on symmetry arguments that place precision centre-stage in terms of amplifying expected actions and percepts. This bilateral role of neuromodulation to select actions and precepts maps nicely to a role for post-synaptic gain in intention and attention. In short, we may be looking at the same mechanism but implemented in different parts of the brain.

9.6 Conclusion

In this chapter, we have tried to cover the fundaments of adaptive behaviour starting from basic principles. We have used the imperative for biological systems to resist an increase in their entropy to motivate a free-energy principle that explains both action and perception. When this principle is unpacked, in the context of generative models the brain might use, we arrive at a fairly simple message-passing scheme based upon prediction errors and the optimisation of their precision by synaptic gain. We then considered generic forms that these models might possess, where the form itself entails prior expectations about the motion of hidden states in the world and, through active inference, behaviour. We considered fixed-point policies of the sort found in psychology and optimal control theory. We then proceeded to itinerant

policies that have a more dynamic and ethologically valid flavour. The notion of itinerant policies, when combined with active inference, provides a rich framework in which to understand many aspects of behaviour. We have focused on changes in behaviour following a down-regulation or up-regulation of the precision, under which perceptual inference and learning proceeds. This was motivated by the psychopharmacology of addiction, which almost invariably involves some change in dopaminergic neurotransmission and, from an algorithmic perspective, the optimisation of precision in the brain. The results of these simulations suggest plausible explanations for bradykinetic and addictive behaviour that rest upon impaired inference and learning respectively. Both the functionalist perspective afforded by this analysis and the putative neurobiological mechanisms fit comfortably with many known facts in addiction research. However, a specific mapping between functional architectures of the sort considered here and the neurobiology of addiction clearly requires more work. Although an awful condition from a clinical point of view, addiction may be nature's most unique and pervasive psychopharmacological experiment, in which complex behaviour confounds the elemental (synaptic) mechanisms upon which it rests.

Acknowledgements The Wellcome Trust funded this work and greatest thanks to Marcia Bennett for helping prepare this manuscript.

Appendix A: Parameter Optimisation and Newton's Method

There is a close connection between the updates implied by Eq. (9.9) and Newton's method for optimisation. Consider the update under a local linearisation, assuming $\mathcal{L}_\varphi \approx \mathcal{F}_\varphi$

$$\begin{aligned}
\Delta\tilde{\mu}^{(\varphi)} &= \left(\exp\left(t\Im^{(\varphi)}\right) - I\right)\Im^{(\varphi)-1}\dot{\tilde{\mu}}^{(\varphi)} \\
\dot{\tilde{\mu}}^{(\varphi)} &= \begin{bmatrix} \mu'^{(\varphi)} \\ -\mathcal{L}_\varphi - \kappa\mu'^{(\varphi)} \end{bmatrix} \\
\Im^{(\varphi)} &= \frac{\partial\dot{\tilde{\mu}}^{(\varphi)}}{\partial\tilde{\mu}^{(\varphi)}} = \begin{bmatrix} 0 & I \\ -\mathcal{L}_{\varphi\varphi} & -\kappa \end{bmatrix}
\end{aligned} \tag{A.1}$$

As time proceeds, the change in generalised mean becomes

$$\begin{aligned}
\lim_{t\to\infty}\Delta\tilde{\mu}^{(\varphi)} &= -\Im^{(\varphi)-1}\dot{\tilde{\mu}}^{(\varphi)} = \begin{bmatrix} \Delta\mu^{(\varphi)} \\ \Delta\mu'^{(\varphi)} \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_{\varphi\varphi}^{-1}\mathcal{L}_\varphi \\ \mu'^{(\varphi)} \end{bmatrix} \\
\Im^{(\varphi)-1} &= \begin{bmatrix} -\kappa\mathcal{L}_{\varphi\varphi}^{-1} & -\mathcal{L}_{\varphi\varphi}^{-1} \\ I & 0 \end{bmatrix}
\end{aligned} \tag{A.2}$$

The first line means the motion cancels itself and becomes zero, while the change in the conditional mean $\Delta\mu^{(\varphi)} = -\mathcal{L}_{\varphi\varphi}^{-1}\mathcal{L}_\varphi$ becomes a classical Newton update. The conditional expectations of the parameters were updated after every simulated exposure using this scheme, as described in Friston (2008).

Appendix B: Simulating Action and Perception

The simulations in this paper involve integrating time-varying states in the environment and the agent. This is the solution to the following ordinary differential equation

$$\dot{u} = \begin{bmatrix} \dot{\tilde{s}} \\ \dot{\tilde{\mathbf{x}}} \\ \dot{\tilde{\mathbf{v}}} \\ \dot{\tilde{\omega}}^{(x)} \\ \dot{\tilde{\omega}}^{(v)} \\ \dot{\tilde{\mu}}^{(x)} \\ \dot{\tilde{\mu}}^{(v)} \\ \dot{a} \end{bmatrix} = \begin{bmatrix} \mathcal{D}\mathbf{g} + \mathcal{D}\tilde{\omega}^{(v)} \\ \mathbf{f} + \tilde{\omega}^{(x)} \\ \mathcal{D}\tilde{\mathbf{v}} \\ \mathcal{D}\tilde{\omega}^{(x)} \\ \mathcal{D}\tilde{\omega}^{(v)} \\ \mathcal{D}\tilde{\mu}^{x} - \mathcal{F}_{\tilde{x}} \\ \mathcal{D}\tilde{\mu}^{v} - \mathcal{F}_{\tilde{v}} \\ -\mathcal{F}_{a} \end{bmatrix}$$

$$\Im = \begin{bmatrix} 0 & \mathcal{D}\mathbf{g}_{\tilde{\mathbf{x}}} & \mathcal{D}\mathbf{g}_{\tilde{\mathbf{v}}} & \mathcal{D} & 0 & 0 & \cdots & 0 \\ & \mathbf{f}_{\tilde{\mathbf{x}}} & \mathbf{f}_{\tilde{\mathbf{v}}} & & I & & & \mathbf{f}_{a} \\ \vdots & & \mathcal{D} & & \vdots & \vdots & & 0 \\ & & & \mathcal{D} & 0 & & & \\ 0 & & \cdots & 0 & \mathcal{D} & 0 & \cdots & \\ -\mathcal{F}_{\tilde{x}\tilde{y}} & & \cdots & & 0 & \mathcal{D} - \mathcal{F}_{\tilde{x}\tilde{x}} & -F_{\tilde{x}\tilde{v}} & -F_{\tilde{x}a} \\ -\mathcal{F}_{\tilde{v}\tilde{y}} & & & & & -\mathcal{F}_{\tilde{v}\tilde{x}} & \mathcal{D} - \mathcal{F}_{\tilde{v}\tilde{v}} & -F_{\tilde{v}a} \\ -\mathcal{F}_{a\tilde{y}} & & & & & -\mathcal{F}_{a\tilde{x}} & -\mathcal{F}_{a\tilde{v}} & -\mathcal{F}_{aa} \end{bmatrix} \tag{B.1}$$

To update these states we use a local linearisation; $\Delta u = (\exp(\Delta t \Im) - I)\Im(t)^{-1}\dot{u}$ over time steps of Δt, where $\Im = \partial \dot{u}/\partial u$ is evaluated at the current conditional expectation (Friston et al. 2010).

References

Abeles M, Hayon G, Lehmann D (2004) Modeling compositionality by dynamic binding of synfire chains. J Comput Neurosci 17(2):179–201

Ahmed SH, Graupner M, Gutkin B (2009) Computational approaches to the neurobiology of drug addiction. Pharmacopsychiatry 42(1):S144–S152 Suppl

Alcaro A, Huber R, Panksepp J (2007) Behavioral functions of the mesolimbic dopaminergic system: an affective neuroethological perspective. Brains Res Rev 56(2):283–321

Ballard DH, Hinton GE, Sejnowski TJ (1983) Parallel visual computation. Nature 306:21–26

Bellman R (1952) On the theory of dynamic programming. Proc Natl Acad Sci USA 38:716–719

Berke JD, Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. Neuron 25(3):515–532

Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656–660

Breakspear M, Stam CJ (2005) Dynamics of a neural system with a multiscale architecture. Philos Trans R Soc Lond B, Biol Sci 360(1457):1051–1074

Bressler SL, Tognoli E (2006) Operational principles of neurocognitive networks. Int J Psychophysiol 60(2):139–148

Bromberg-Martin ES, Hikosaka O (2009) Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron 63:119–126
Coricelli G, Dolan RJ, Sirigu A (2007) Brain, emotion and decision making: the paradigmatic example of regret. Trends Cogn Sci 11(6):258–265
Camerer CF (2003) Behavioural studies of strategic thinking in games. Trends Cogn Sci 7(5):225–231
Chen X, Zelinsky GJ (2006) Real-world visual search is dominated by top-down guidance. Vis Res 46(24):4118–4133
Colliaux D, Molter C, Yamaguchi Y (2009) Working memory dynamics and spontaneous activity in a flip-flop oscillations network model with a Milnor attractor. Cogn Neurodyn 3(2):141–151
Crauel H (1999) Global random attractors are uniquely determined by attracting deterministic compact sets. Ann Mat Pura Appl 176(4):57–72
Crauel H, Flandoli F (1994) Attractors for random dynamical systems. Probab Theory Relat Fields 100:365–393
Davidson TL (1993) The nature and function of interoceptive signals to feed: toward integration of physiological and learning perspectives. Psychol Rev 100(4):640–657
Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16(2):199–204
Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441(7095):876–879
Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8(4):429–453
Dayan P, Hinton GE, Neal RM (1995) The Helmholtz machine. Neural Comput 7:889–904
Dommett E, Coizet V, Blaha CD, Martindale J, Lefebvre V, Walton N, Mayhew JE, Overton PG, Redgrave P (2005) How visual stimuli activate dopaminergic neurons at short latency. Science 307:1476–1479
Eldredge N, Gould SJ (1972) Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf TJM (ed) Models in paleobiology. Freeman, San Francisco, pp 82–115
Evans DJ (2003) A non-equilibrium free energy theorem for deterministic systems. Mol Phys 101:15551–15554
Feynman RP (1972) Statistical mechanics. Benjamin, Reading
Freeman WJ (1994) Characterization of state transitions in spatially distributed, chaotic, nonlinear, dynamical systems in cerebral cortex. Integr Physiol Behav Sci 29(3):294–306
Friston KJ, Tononi G, Reeke GN Jr, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience 59(2):229–243
Friston KJ (2000) The labile brain. II. Transients, complexity and selection. Phil Trans Biol Sci 355(1394):237–252
Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B, Biol Sci 360(1456):815–836
Friston K (2008) Hierarchical models in the brain. PLoS Comput Biol 4(11):e1000211
Friston K, Kilner J, Harrison L (2006) A free energy principle for the brain. J Physiol Paris 100(1–3):70–87
Friston KJ, Daunizeau J, Kiebel SJ (2009) Reinforcement learning or active inference? PLoS ONE 29;4(7):e6421
Friston KJ, Daunizeau J, Kilner J, Kiebel SJ (2010) Action and behavior: a free-energy formulation. Biol Cybern [Epub ahead of print]
Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299(5614):1898–1902
Fiorillo CD (2008) Towards a general theory of neural computation based on prediction by single neurons. PLoS ONE 3:e3298
Goto Y, Yang CR, Otani S (2010) Functional and dysfunctional synaptic plasticity in prefrontal cortex: roles in psychiatric disorders. Biol Psychiatry 67(3):199–207
Gregory RL (1968) Perceptual illusions and brain models. Proc R Soc Lond B 171:179–196
Gregory RL (1980) Perceptions as hypotheses. Phil Trans R Soc Lond B 290:181–197

Gros C (2009) Cognitive computation with autonomously active neural networks: an emerging field. Cogn Comput 1:77–99
Haile PA, Hortaçsu A, Kosenok G (2008) On the empirical content of quantal response equilibrium. Am Econ Rev 98:180–200
Haken H (1983) Synergistics: an introduction. Non-equilibrium phase transition and self-organisation in physics, chemistry and biology, 3rd edn. Springer, Berlin
Herrmann JM, Pawelzik K, Geisel T (1999) Self-localization of autonomous robots by hidden representations. Auton Robots 7:31–40
Hinton GE, van Camp D (1993) Keeping neural networks simple by minimising the description length of weights. In: Proceedings of COLT-93, pp 5–13
von Helmholtz H (1866) Concerning the perceptions in general. In: Treatise on physiological optics, vol III, 3rd edn (translated by J.P.C. Southall 1925 Opt Soc Am Section 26, reprinted New York, Dover, 1962)
Henry DJ, White FJ (1995) The persistence of behavioral sensitization to cocaine parallels enhanced inhibition of nucleus accumbens neurons. J Neurosci 15(9):6287–6299
Hull C (1943) Principles of behavior. Appleton/Century-Crofts, New York
Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310(5754):1680–1683
Jirsa VK, Friedrich R, Haken H, Kelso JA (1994) A theoretical model of phase transitions in the human brain. Biol Cybern 71(1):27–35
Johnson A, van der Meer MA, Redish AD (2007) Integrating hippocampus and striatum in decision-making. Curr Opin Neurobiol 17(6):692–697
Kelley AE, Berridge KC (2002) The neuroscience of natural rewards: relevance to addictive drugs. J Neurosci 22(9):3306–3311
Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304
Khoshbouei H, Wang H, Lechleiter JD, Javitch JA, Galli A (2003) Amphetamine-induced dopamine efflux. A voltage-sensitive and intracellular Na^+-dependent mechanism. J Biol Chem 278(14):12070–12077
Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27(12):712–719
Lapish CC, Seamans JK, Chandler LJ (2006) Glutamate-dopamine cotransmission and reward processing in addiction. Alcohol Clin Exp Res 30:451–1465
Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A, Opt Image Sci Vis 20:1434–1448
Lee HJ, Youn JM, MJ O, Gallagher M, Holland PC (2006) Role of substantia nigra-amygdala connections in surprise-induced enhancement of attention. J Neurosci 26(22):6077–6081
Liss B, Roeper J (2008) Individual dopamine midbrain neurons: functional diversity and flexibility in health and disease. Brains Res Rev 58(2):314–321
Lodge DJ, Grace AA (2006) The hippocampus modulates dopamine neuron responsivity by regulating the intensity of phasic neuron activation. Neuropsychopharmacology 31:1356–1361
MacKay DM (1956) The epistemological problem for automata. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 235–251
MacKay DJC (1995) Free-energy minimisation algorithm for decoding and cryptoanalysis. Electron Lett 31:445–447
Matheron G (1975) Random sets and integral geometry. Wiley, New York
Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459:837–841
Maturana HR, Varela F (1980) De máquinas y seres vivos. Editorial Universitaria, Santiago. English version: *Autopoiesis: the organization of the living*, in Maturana, HR, and Varela, FG, Autopoiesis and Cognition. Dordrecht, Netherlands: Reidel
Maynard Smith J (1992) Byte-sized evolution. Nature 355:772–773
McDonald RJ, Ko CH, Hong NS (2002) Attenuation of context-specific inhibition on reversal learning of a stimulus-response task in rats with neurotoxic hippocampal damage. Behav Brain Res 136(1):113–126

McKelvey R, Palfrey T (1995) Quantal response equilibria for normal form games. Games Econ Behav 10:6–38
Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive Hebbian learning. Nature 377(6551):725–728
Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947
Moore CC (1966) Ergodicity of flows on homogeneous spaces. Am J Math 88:154–178
Morris R (1984) Developments of a water-maze procedure for studying spatial learning in the rat. J Neurosci Methods 11(1):47–60
Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66:241–251
Nara S (2003) Can potentially useful dynamics to solve complex problems emerge from constrained chaos and/or chaotic itinerancy? Chaos 13(3):1110–1121
Neisser U (1967) Cognitive psychology. Appleton/Century-Crofts, New York
Nestler EJ (2005) Is there a common molecular pathway for addiction? Nat Neurosci 8(11):1445–1449
Niesink RJ, Van Ree JM (1989) Involvement of opioid and dopaminergic systems in isolation-induced pinning and social grooming of young rats. Neuropharmacology 28(4):411–418
Niv Y, Schoenbaum G (2008) Dialogues on prediction errors. Trends Cogn Sci 12(7):265–272
Nowak M, Sigmund K (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's Dilemma game. Nature 364:56–58
O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res 34(1):171–175
Panksepp J, Siviy S, Normansell L (1984) The psychobiology of play: theoretical and methodological perspectives. Neurosci Biobehav Rev 8(4):465–492
Panksepp J, Knutson B, Burgdorf J (2002) The role of brain emotional systems in addictions: a neuro-evolutionary perspective and new 'self-report' animal model. Addiction 97(4):459–469
Pasquale V, Massobrio P, Bologna LL, Chiappalone M, Martinoia S (2008) Self-organization and neuronal avalanches in networks of dissociated cortical neurons. Neuroscience 153(4):1354–1369
Pierce RC, Kalivas PW (1997) A circuitry model of the expression of behavioural sensitization to amphetamine-like psychostimulants. Brain Res Brain Res Rev 25(2):192–216
Porr B, Wörgötter F (2003) Isotropic sequence order learning. Neural Comput 15(4):831–864
Rabinovich M, Huerta R, Laurent G (2008) Neuroscience. Transient dynamics for neural processing. Science 321(5885):48–50
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87
Redgrave P, Gurney K (2006) The short-latency dopamine signal: a role in discovering novel actions? Nat Rev, Neurosci 7(12):967–975
Redish AD (2004) Addiction as a computational process gone awry. Science 306:1944–1947
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF (eds) Classical conditioning II: current research and theory. Appleton/Century Crofts, New York, pp 64–99
Robbe D, Buzsáki G (2009) Alteration of theta timescale dynamics of hippocampal place cells by a cannabinoid is associated with memory impairment. J Neurosci 29(40):12597–12605
Salzman CD, Belova MA, Paton JJ (2005) Beetles, boxes and brain cells: neural mechanisms underlying valuation and learning. Curr Opin Neurobiol 15(6):721–729
Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80(1):1–27
Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annu Rev Neurosci 23:473–500
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599
Seip KM, Pereira M, Wansaw MP, Reiss JI, Dziopa EI, Morrell JI (2008) Incentive salience of cocaine across the postpartum period of the female rat. Psychopharmacology 199(1):119–130

Sheynikhovich D, Chavarriaga R, Strösslin T, Arleo A, Gerstner W (2009) Is there a geometric module for spatial orientation? Insights from a rodent navigation model. Psychol Rev 116(3):540–566

Shreve S, Soner HM (1994) Optimal investment and consumption with transaction costs. Ann Appl Probab 4:609–692

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88(2):135–170

Takahashi Y, Schoenbaum G, Niv Y (2008) Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front Neurosci 2:86–99

Tani J, Ito M, Sugita Y (2004) Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB. Neural Netw 17:1273–1289

Thiagarajan TC, Lebedev MA, Nicolelis MA, Plenz D (2010) Coherence potentials: loss-less all-or-none network events in the cortex. PLoS Biol 8(1):e1000278

Todorov E (2006) Linearly-solvable Markov decision problems. In: Scholkopf et al (ed) Advances in neural information processing systems, vol 19, pp 1369–1376. MIT Press, Cambridge

Traulsen A, Claussen JC, Hauert C (2006) Coevolutionary dynamics in large, but finite populations. Phys Rev E, Stat Nonlinear Soft Matter Phys 74(1 Pt 1):011901

Tschacher W, Haken H (2007) Intentionality in non-equilibrium systems? The functional aspects of self-organised pattern formation. New Ideas Psychol 25:1–15

Tsuda I (2001) Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. Behav Brain Sci 24(5):793–810

Tyukin I, van Leeuwen C, Prokhorov D (2003) Parameter estimation of sigmoid superpositions: dynamical system approach. Neural Comput 15(10):2419–2455

Tyukin I, Tyukina T, van Leeuwen C (2009) Invariant template matching in systems with spatiotemporal coding: a matter of instability. Neural Netw 22(4):425–449

van Leeuwen C (2008) Chaos breeds autonomy: connectionist design between bias and baby-sitting. Cogn Process 9(2):83–92

Verschure PF, Voegtlin T, Douglas RJ (2003) Environmentally mediated synergy between perception and behavior in mobile robots. Nature 425:620–624

Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8:279–292

Wittmann BC, Bunzeck N, Dolan RJ, Duzel E (2007) Anticipation of novelty recruits reward system and hippocampus while promoting recollection. Neuroimage 38:194–202

Zack M, Poulos CX (2009) Parallel roles for dopamine in pathological gambling and psychostimulant addiction. Curr Drug Abus Rev 2(1):11–25

Zhao Y, Kerscher N, Eysel U, Funke K (2001) Changes of contrast gain in cat dorsal lateral geniculate nucleus by dopamine receptor agonists. Neuroreport 12(13):2939–2945

Zink CF, Pagnoni G, Chappelow J, Martin-Skurski M, Berns GS (2006) Human striatal activation reflects degree of stimulus saliency. Neuroimage 29:977–983