I am therefore I think

Karl Friston[1] and Christoph Mathys[1,2,3]

[1] *The Wellcome Trust Centre for Neuroimaging, UCL, 12 Queen Square, London, WC1N 3BG, UK*
[2]*Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Switzerland*
[3]*Laboratory for Social and Neural Systems Research (SNS Lab), Department of Economics, University of Zurich, Switzerland*

Address for Correspondence
Karl Friston
The Wellcome Trust Centre for Neuroimaging,
Institute of Neurology, UCL
12 Queen Square, London, UK WC1N 3BG
Email k.friston@ucl.ac.uk

Abstract

This article offers an account of embodied exchange with the world that associates conscious operations with actively inferring the causes of our sensations. Its agenda is to link formal (mathematical) descriptions of dynamical systems to a description of perception in terms of beliefs and goals. The argument has two parts: the first calls on the lawful dynamics of any (weakly mixing ergodic) system that persists in a changing environment – from a single cell organism to a human brain. These lawful dynamics suggest that (internal) states can be interpreted as modelling or predicting the (external) causes of sensory perturbations. In other words, if a system exists, its internal states must encode probabilistic beliefs about external states. Heuristically, this means that if I exist (am) then I must have beliefs (think). The second part of the argument is that the only tenable beliefs I can entertain about myself are that I exist. This may seem rather facile; however, if we associate existing with ergodicity, then (ergodic) systems that exist by predicting external states can only possess prior beliefs that their environment is predictable. It transpires that this is equivalent to believing that the world – and the way it is sampled – will minimise uncertainty about the causes of sensations. We conclude by illustrating the behaviour that emerges under these beliefs using simulations of saccadic searches – and considering them in the light of conscious perception.

Key words: *active inference · autopoiesis · cognitive · dynamics · free energy · consciousness · self-organization*

1 Introduction

"How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" Erwin Schrödinger (1943)

Friston, K. (2017). I am therefore I think. In M. Leuzinger-Bohleber, S. Arnold & M. Solms (Eds.), *The unconscious : a bridge between psychoanalysis and cognitive neuroscience* (pp. 113-137). London: Routledge.

How does Schrödinger's question touch on philosophical propositions, such as René Descartes' famous proposition *Cogito ergo sum* ('I think, therefore I am'). This article tries to cast thinking in terms of probabilistic beliefs and thereby link consciousness to probabilistic descriptions of self-organisation – of the sort found in statistical physics and dynamical systems theory. This agenda may sound ambitious but it may be easier than at first glance. The trick is to associate philosophical notions – such as 'I think' and 'I am' – with formal constructs, such as probabilistic inference and the implications of existing over extended periods of time. In brief, we start off by considering what it means for a system, like a cell or a brain, to exist. Mathematically, this implies certain properties that constrain the way the states of such systems must change – such as ergodicity. In other words, any measurement of such systems must converge over time (Birkhoff, 1931). Crucially, if we consider these fundamental properties in the context of a separation between internal and external states, one can show that the internal states minimize the same quantity minimized in Bayesian statistics. This means one can always interpret a system that exists and as making probabilistic inferences about its external milieu. Formally, this turns Descartes' proposition on its head; suggesting that 'I am [ergodic], therefore I think'. However, if a system 'thinks' – in the sense of updating probabilistic beliefs – then what does it believe?

We will consider the self-consistent (prior) belief, which is 'I think, therefore I am [ergodic]'. This prior belief is formally equivalent to minimising uncertainty about the (external) causes of sensory states that intervene between external and internal states. This is precisely the imperative that drives both scientific hypothesis testing and active perception (Helmholtz, 1866/1962; Gregory, 1980; O'Regan & Noë, 2001; Wurtz, McAlonan, Cavanaugh, & Berman, 2011). Furthermore, it provides a nice perspective on the active sampling of our sensorium. This perspective suggests perceptual processing can be associated with Bayesian belief updating (Dayan, Hinton, & Neal, 1995). For a comprehensive and heuristic discussion of these ideas see (Clark, 2013) and (Hohwy, 2013). To make these arguments more concrete, we will consider an example of the ensuing active inference, using simulations originally described in (Friston, Adams, Perrinet, & Breakspear, 2012).

This article comprises three sections. The first section draws on two recent developments in formal treatments of self-organization and nonequilibrium steady-state dynamics. The first development is an application to Bayesian inference and embodied perception in the brain (Friston, Kilner, & Harrison, 2006). The second is an attempt to understand the nature of self-organization in random dynamical systems (Ashby, 1947; Haken, 1983; Maturana & Varela, 1980; Nicolis & Prigogine, 1977; Schrödinger, 1944). This material has been presented previously (Friston, 2013) and – although rather technical – has some relatively simple implications. Its premise is that biological self-organization is (almost) inevitable and manifests as a form of active Bayesian inference. We have previously suggested (Friston, 2013) that the events "*within the spatial boundary of a living organism*" (Schrödinger, 1944) may arise from the very existence of a boundary or blanket – and that a (Markov) blanket may be inevitable under local coupling among dynamical systems. We will see that the very existence of a Markov blanket means we can interpret the self-organisation of internal states in terms of Bayesian inference about external states.

The second section looks more closely at the nature of Bayesian inference, in terms of prior beliefs that might be associated with a system that minimises the dispersion (entropy) of its external states through action. We will see that these prior beliefs lead to a pre-emptive sampling of the sensorium that tries to minimise uncertainty about hypotheses encoded by the internal states.

The final section illustrates these ideas using simulations of saccadic eye movements to unpack the nature of active inference. This section contains a brief description of an agent's prior beliefs about the causal structure of its (visual) world and how that world would be sampled to minimise uncertainty. Finally, the results of simulated saccadic eye movements and associated inference are described, with a special emphasis on the selection of beliefs and hypotheses that are entertained by our enactive brain.

2 I am therefore I think

This section covers the basic theory behind the self-organisation of (weakly mixing ergodic) random dynamical systems to show that they can always be interpreted in terms of active modelling or inference. This is important because it leads to conclusions that are exactly consistent with the good regulator theorem (every good regulator is a model of its environment) and related treatments of self-organization (Ashby, 1947; Nicolis & Prigogine, 1977; van Leeuwen, 1990; Maturana & Varela, 1980). It also means that there is a direct Bayesian interpretation of any self-organised dynamics in a system that exists in the ergodic sense. What follows is a summary of the material in (Friston, 2013).

2.1 Ergodic densities and flows

We start by considering any (weakly mixing) ergodic random dynamical that can be described by stochastic differential equations of the following form:

$$\dot{x} = f(x) + \omega \qquad 1$$

Here, the flow of states $f(x)$ is subject to random fluctuations ω. Because the system is ergodic (and weakly mixing) it will, after a sufficient amount of time, converge to an invariant set of states called a *pullback* or *random global attractor* (Crauel & Flandoli, 1994; Crauel, 1999). The associated ergodic density $p(x \mid m)$ for any system or model m is the solution to the Fokker-Planck equation (Frank, 2004) describing the time evolution of the probability density over states

$$\dot{p}(x \mid m) = \nabla \cdot (\Gamma \nabla - f)p \qquad 2$$

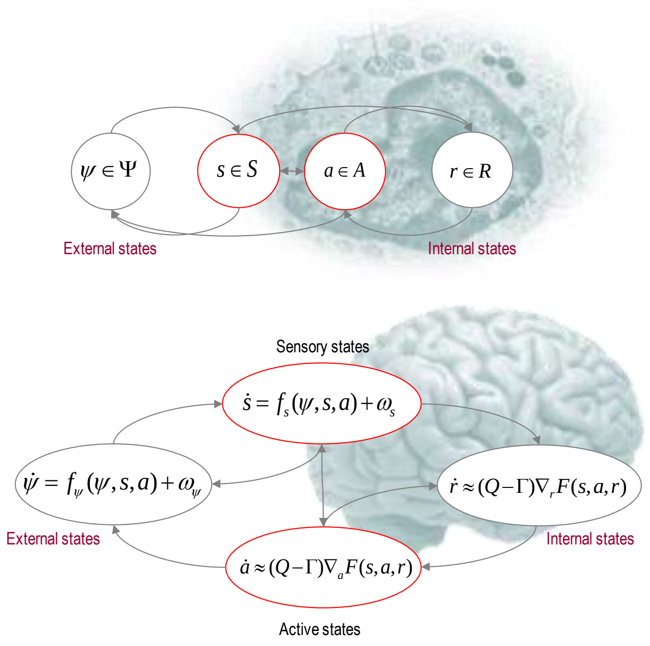
Here, the diffusion tensor Γ is the half the covariance (amplitude) of the random fluctuations. Equation 2 shows that the ergodic density depends upon flow, which can always be expressed in terms of curl-free and divergence-free components. This is the Helmholtz decomposition (a.k.a. the fundamental theorem of vector calculus) and can be formulated in terms of an anti-symmetric matrix $Q(x) = -Q(x)^T$ and a scalar potential $L(x)$ that plays a role of a Lagrangian (Ao, 2004)

$$f(x) = (Q - \Gamma)\nabla L(x) \quad 3$$

Using this standard form (Yuan, Ma, Yuan, & Ping, 2010), it is straightforward to show that $p(\tilde{x} \mid m) = \exp(-L(x))$ is the solution to the Fokker Planck equation above (Friston & Ao, 2012). This means one can express the flow in terms of the ergodic density:

$$f(\tilde{x}) = (Q - \Gamma)\nabla \ln p(x \mid m) \quad 4$$

This is an important result because it shows the flow can be decomposed into a component that flows towards regions with a higher ergodic density (the curl-free or irrotational component) and an orthogonal (divergence-free or solenoidal) component that circulates on isocontours of the ergodic density. These components are like walking uphill and walking around the base of the hill respectively. In summary, any ergodic random dynamical system can be formulated as a circuitous ascent on the log likelihood over the states it visits. Later, we will interpret this likelihood in a statistical sense and see that any random dynamical system can be interpreted as performing some form of inference on itself.



$\psi \in \Psi$
$s \in S$
$a \in A$
$r \in R$
External states
Internal states
Sensory states
$\dot{s} = f_s(\psi, s, a) + \omega_s$
$\dot{\psi} = f_\psi(\psi, s, a) + \omega_\psi$
$\dot{r} \approx (Q - \Gamma)\nabla_r F(s, a, r)$
External states
Internal states
$\dot{a} \approx (Q - \Gamma)\nabla_a F(s, a, r)$
Active states

Figure 1: Markov blankets and the free energy principle. *These schematics illustrate the partition of states into internal states and external states that are separated by a Markov blanket – comprising sensory and active states. The upper panel shows this partition as it might be applied to a cell: where the internal states can be associated with the intracellular states of a cell, while sensory states become the surface states or cell membrane overlying active states (e.g., the actin filaments of the cytoskeleton). The lower panel shows the same dependencies but rearranged so that they can be related to action and perception in the brain: where active and internal states minimize a free energy functional of sensory states. The ensuing self-organization of internal states then corresponds to perception, while action couples brain states back to external states. See main text for details and Table 1 for a definition of the variables.*

Table 1: *Definitions of the tuple* $(\Omega, \Psi, S, A, R, p, q)$ *underlying active inference*

- *A sample space* Ω from which random fluctuations $\omega \in \Omega$ are drawn
- *External states* $\mathrm{Y} : \mathrm{Y} \times A \times \mathrm{W} \to \mathbb{R}$ – states of the world that cause sensory states and depend on action
- *Sensory states* $S : \mathrm{Y} \times A \times \mathrm{W} \to \mathbb{R}$ – the agent's sensations that constitute a probabilistic mapping from action and external states
- *Action states* $A : S \times R \times \Omega \to \mathbb{R}$ – an agent's action that depends on its sensory and internal states
- *Internal states* $R : R \times S \times \Omega \to \mathbb{R}$ – representational states of the agent that cause action and depend on sensory states
- *Ergodic density* $p(\psi, s, a, r \mid m)$ – a probability density function over external $\psi \in \Psi$, sensory $s \in S$, active $a \in A$ and internal states $r \in R$ for a system or model denoted by m
- *Variational density* $q(\psi \mid r)$ – an arbitrary probability density function over external states that is parameterized by internal states

2.2 Systems and Markov blankets

When we talk about a system that can be distinguished from its environment, we implicitly call on the notion of a Markov blanket. A Markov blanket is a set of states that separates two other sets in a statistical sense. The term Markov blanket was introduced in the setting of Bayesian networks or graphs (Pearl, 1988) and refers to the children of a set (the set of states that are influenced), its parents (the set of states that influence it) and the parents of its children.

A Markov blanket induces a partition of states into *internal states* and *external* states that are hidden (insulated) from the internal (insular) states by the Markov blanket. For example, the surface of a cell may constitute a Markov blanket separating intracellular (internal) and extracellular (external) states (Auletta, 2013; Friston, 2013). Statistically speaking, external states can only be seen vicariously by the internal states, through the Markov blanket. The Markov blanket can itself be partitioned into two sets that are, and are not, children of external states. We will refer to these as surface or *sensory states* and *active states* respectively. Put simply, the existence of a Markov blanket $S \times A$ implies a partition of states into external, sensory, active and internal states: $x \in X = \Psi \times S \times A \times R$ as in Figure 1. External states cause sensory states that influence – but are not influenced by – internal states, while internal states cause active states that influence – but are not influenced by – external states. Crucially, the dependencies induced by Markov blankets create a circular causality that is reminiscent of the perception and action cycle (Fuster, 2004). This circular causality means that external states cause changes in internal states, via sensory states, while the internal states couple back to external states through active states.

Equipped with this partition, we can now consider the dependencies among states implied by the Markov blanket, in terms of their flow above. The flow through any point (s, a, r) in the state space of the internal states and their Markov blanket is (Friston, 2013):

$$\begin{aligned} f_r(s,a,r) &= (\Gamma - Q)\nabla_r \ln p(s,a,r \mid m) \\ f_a(s,a,r) &= (\Gamma - Q)\nabla_a \ln p(s,a,r \mid m) \end{aligned} \qquad 5$$

This shows that the flow of internal and active states performs a circuitous gradient ascent on the *marginal* ergodic density over internal states and their Markov blanket. It is a marginal density because we have marginalised over the external states. This means the internal and active states behave as if they know the distribution over external states that would be necessary to perform the marginalisation. In other words, the internal states will appear to respond to sensory fluctuations based on (posterior) beliefs about underlying fluctuations in external states. We can formalize this notion by associating these beliefs with a probability density over external states $q(\psi \mid r)$ that is encoded (parameterized) by internal states:

Lemma (free energy): *for any random dynamical system with a Markov blanket and Lagrangian* $L(x) = -\ln p(\psi, s, a, r)$*, there is a free energy* $F(s,a,r)$ *that describes the flow of internal and active states in terms of a generalized descent*

$$\begin{aligned} f_r(s,a,r) &= (Q - \Gamma)\nabla_r F \\ f_a(s,a,r) &= (Q - \Gamma)\nabla_a F \\ F(s,a,r) &= E_q[L(x)] - H[q(\psi \mid r)] \end{aligned} \qquad 6$$

This free energy is a functional of a variational) density $q(\psi \mid r)$ *– parameterized by internal states – that corresponds to the expected Lagrangian minus the entropy of the variational density.*

Proof: using Bayes rule, we can rearrange the expression for free energy in terms of a Kullback-Leibler divergence (Beal, 2003)

$$F(s,a,r) = -\ln p(s,a,r \mid m) + D_{KL}[q(\psi \mid r) \,||\, p(\psi \mid s,a)] \qquad 7$$

When $q(\psi \mid r) = p(\psi \mid s,a)$, the divergence term disappears and we recover the ergodic flow in Equation 5:

$$\begin{aligned} f_r(s,a,r) &= (\Gamma - Q)\nabla_r \ln p(s,a,r \mid m) \\ f_a(s,a,r) &= (\Gamma - Q)\nabla_a \ln p(s,a,r \mid m) \end{aligned} \qquad 8$$

In other words, the ergodic flow ensures the variational density is the posterior density, such that the variational density represents the hidden states in a Bayes-optimal sense □

Remarks: all this proof says is that if one interprets internal states as parameterising Bayesian beliefs about external states, then the dynamics of internal and active states can be described as a gradient descent on a variational free energy function of internal states and their Markov blanket. Variational free energy was introduced by Feynman to solve difficult marginalisation problems in path integral formulations of quantum physics (Feynman, 1972). This is also the free energy bound that is used extensively in *approximate Bayesian inference* (e.g., variational Bayes) (Beal, 2003; Hinton & van Camp, 1993; Kass & Steffey, 1989). The expression for free energy in Equation 7 discloses its Bayesian interpretation: the first term is the negative log evidence or *marginal likelihood* of the internal states and their Markov blanket. The second term is a *relative entropy* or Kullback-Leibler divergence (Kullback & Leibler, 1951) between the variational density and the posterior density over external states (i.e., the causes of sensory states). Because this divergence cannot be less than zero, the internal flow will appear to have minimized the divergence between the variational and posterior density. In other words, the internal states will appear to have solved the problem of Bayesian inference by encoding posterior beliefs about the causes of its sensory states – under a generative model provided by the Lagrangian. This is known as *exact* Bayesian inference because the variational and posterior densities are identical. Later, we will consider approximate forms (under the Laplace assumption) leading to *approximate* Bayesian inference. In short, the internal states will appear to engage in Bayesian inference but what about action?

Because the divergence in Equation 7 can never be less than zero, free energy is an upper bound on the negative log evidence. Now, because the system is ergodic we have

$$\begin{gathered} F(s,a,r) \geq -\ln p(s,a,r \mid m) \\ \Rightarrow \\ E_t[F(s,a,r)] \geq E_t[\underbrace{-\ln p(s,a,r \mid m)}_{\text{expected surprise}}] = \underbrace{H[p(s,a,r \mid m)]}_{\text{entropy}} \end{gathered} \qquad 9$$

This means that action will (on average) appear to place an upper bound on the entropy of the internal states and their Markov blanket. Together with the Bayesian modelling perspective, this is consistent with the good regulator theorem (Conant & Ashby, 1970) and related accounts of self-organization (Ashby, 1947; Nicolis & Prigogine, 1977; Friston & Ao, 2012; van Leeuwen, 1990; Pasquale, Massobrio, Bologna, Chiappalone, & Martinoia, 2008). These treatments emphasise the homoeostatic nature of self-organisation that maintains internal states within physiological bounds. This characteristic resistance to the dispersion of internal states underlies the peculiar ability of animate systems to resist the second law of thermodynamics – or more precisely the fluctuation theorem (Evans & Searles, 1994). Furthermore, as shown elsewhere (Friston, 2012; Friston, 2010) free energy minimization is consistent with information-theoretic formulations of sensory processing and behaviour (Bialek, Nemenman, & Tishby, 2001; Barlow, 1961; Linsker, 1990). Finally, Equation 7 shows that minimizing free energy entails maximizing the entropy of the variational density (or posterior uncertainty) in accord with the maximum entropy principle (Jaynes, 1957). Maximising posterior uncertainty may seem odd but it is a vital part of Bayesian inference – that is closely related to the Laplace's principle of indifference and Occam's razor.

One interesting feature of the free energy formulation is that internal states encode beliefs about the consequences of action. In other words, internal states can only infer action through its sensory sequelae – in the same sense that we are not aware of our movements *per se*, only their consequences. This creates an important distinction between action and (approximate posterior) beliefs about its consequences encoded by internal states.

From our current (existential) perspective, this statistical interpretation of ergodic behaviour has something quite profound to say: because internal states encode probabilistic beliefs, the free energy *function* of internal states now becomes a free energy *functional* (function of a function) of a probability distribution. This is interesting because the flow of (material) states therefore becomes a functional of (immaterial) probabilistic beliefs. This furnishes a formal link between the physical (states or *res extensa*) and the mindful (beliefs or *res cogitans*). Another perspective on this bridge over the Cartesian divide is that the free energy Lemma above provides a (wide sense) realization relationship (Wilson 2001; Gillett 2002). In other words, the implicit process of inference affords a unique mapping between biophysical states (internal states) and the properties (probabilistic beliefs) they realize (c.f., Bechtel 1999). See figure 2. In the next section, we consider the nature of these beliefs in more detail and work towards a more explicit realisation of beliefs in the context of active inference and Bayesian belief updating.

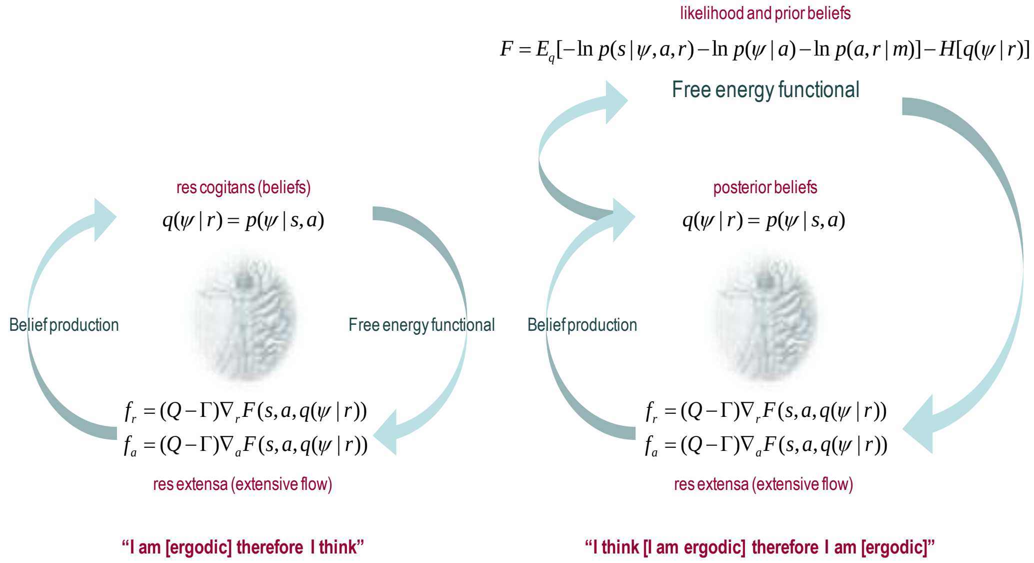

likelihood and prior beliefs
$F = E_q[-\ln p(s \mid \psi, a, r) - \ln p(\psi \mid a) - \ln p(a, r \mid m)] - H[q(\psi \mid r)]$
Free energy functional
res cogitans (beliefs)
$q(\psi \mid r) = p(\psi \mid s, a)$
posterior beliefs
$q(\psi \mid r) = p(\psi \mid s, a)$
Belief production
Free energy functional
Belief production
$f_r = (Q - \Gamma)\nabla_r F(s, a, q(\psi \mid r))$
$f_a = (Q - \Gamma)\nabla_a F(s, a, q(\psi \mid r))$
$f_r = (Q - \Gamma)\nabla_r F(s, a, q(\psi \mid r))$
$f_a = (Q - \Gamma)\nabla_a F(s, a, q(\psi \mid r))$
res extensa (extensive flow)
res extensa (extensive flow)
"I am [ergodic] therefore I think"
"I think [I am ergodic] therefore I am [ergodic]"

***Figure 2**: **Material and immaterial aspects of ergodic dynamics.** Left panel: This schematic highlights the relationship between the (ergodic) flow of biophysical states described by a free energy functional of a probability density. This (variational) density corresponds to posterior beliefs over the external states causing sensory states. The important point illustrated here is that there is a lawful coupling between the (material) flow and the (immaterial) beliefs that are realised by (caused by) and realise (cause) the flow. Right panel: the same scheme has been unpacked to highlight the fact that the free energy functional depends upon both posterior beliefs and prior beliefs inherent in the generative model (or flow's Lagrangian). Furthermore, these prior beliefs are themselves predicated on posterior beliefs (through their uncertainty or entropy) adding an extra layer of probabilistic bootstrapping. The key distinction between right and left panels is that the (material) flow of the system is now prescribed by (immaterial) prior beliefs about the nature of this flow and, in this sense, suggests one can interpret the dynamics of animate systems as predicated upon beliefs about their own behaviour.*

3 I think therefore I am

The previous section established that (ergodic) systems with measures that converge over time are, in some sense, equipped with probabilistic beliefs that are encoded by their internal states. However, there is nothing in this formulation that differentiates between the sorts of systems (or their Markov blankets) that emerge over time – or the nature of the attracting set of states that endow them with a recognisable phenotype. This attracting set or random dynamical attractor can be highly structured and space-filling but retain a low volume or entropy (Freeman, 1994) – like the cycle of states we seek out in our daily routine. Conversely, the attracting set could encompass a large number of states with an amorphous (high volume or entropy) attractor. In this section, we try to explain the emergence of structured attracting sets that are characteristic of animate systems – like ourselves – by exploiting the modelling interpretation of ergodic flows. The argument goes as follows:

Lemma: *the internal states of any ergodic system that is equipped with a Markov blanket and a low measure attracting set (or low entropy ergodic density) must believe they have a low measure attracting set.*

Proof: the proof is by *reductio ad absurdum*. The free energy Lemma demonstrates the existence of a Lagrangian that plays the role of a probabilistic generative model of a system's external states. This model entails (prior) beliefs of the system that determine its action, where action constitutes the flow described by the Lagrangian. Now, if the beliefs entailed by the Lagrangian do not include a low measure attracting set, then any action that preserves a low measure attracting set cannot be described by the Lagrangian. This means the system cannot be ergodic (because it violates free energy Lemma) and will dissipate after a sufficient period of time □

Remarks: this argument rests upon the circular causality inherent in the mapping between probabilistic descriptions (beliefs) and biophysical dynamics (flows). In other words, action is a component of flow that induces an ergodic probability density with an associated Lagrangian. This Lagrangian entails beliefs that describe action. Put simply, the physical behaviour of an ergodic system must be consistent with its beliefs. This means it is entirely tenable to regard the internal states (encoding beliefs) to be the authors of their own existence – where these beliefs are fulfilled by action. Clearly, this rests on a permissive environment that allows itself to be sampled in a way that renders the beliefs a veridical description of that sampling. This permissive aspect emphasises the co-evolution of agents and their (sampled) environments that are forever locked in an existential alliance.

In what follows, we will work through this argument in more detail, looking at the beliefs that systems or agents might entertain. In the current setup, beliefs are probability distributions that specify the free energy, which describes the flow of internal states and action (Equation 6). We have seen that posterior beliefs are encoded by the internal states. However, free energy is also defined by the Lagrangian or probabilistic generative model. We can unpack this model in terms of its likelihood and priors in the following way:

$$F = E_q[-\underbrace{\ln p(\tilde{s} \mid \tilde{\psi}_v, \tilde{\psi}_u, \tilde{a}, \tilde{r})}_{\text{likelihood}} - \underbrace{\ln p(\tilde{\psi}_v \mid \tilde{\psi}_u)}_{\text{empirical priors}} - \underbrace{\ln p(\tilde{\psi}_u \mid \tilde{a})}_{\text{empirical priors}} - \underbrace{\ln p(\tilde{a}, \tilde{r} \mid m)}_{\text{full priors}}] - \underbrace{H[q(\tilde{\psi} \mid \tilde{r})]}_{\text{posterior uncertainty}} \qquad 10$$

Here, the external states $\Psi = \Psi_v \times \Psi_u$ have been separated into *hidden* and *control states*. Control states are the consequences of action and – in the generative model – determine the evolution of hidden states. Equation 10 also introduces a ~ over states to denote their motion, acceleration and so on, such that $\tilde{x} = (x, x', \ldots)$. This allows us to consider probabilities over the flow or trajectories of states as opposed their instantaneous value.

As above, the free energy is the expected Lagrangian (energy) minus the entropy of the posterior beliefs. However, here the Lagrangian has been factorised into likelihood, empirical and full priors. The likelihood is simply the probability of some sensory states given their causes, while empirical priors are probability distributions over the hidden states, given control states and control states, given action. The full priors are over action and internal states and will not play a role in what follows. Empirical priors are a technical term denoting components of a (hierarchical) generative model that are 'sandwiched' between the likelihood and full priors.

These components of the generative model are implicit in the system dynamics; in other words, they constitute the Lagrangian whose expectation, under posterior beliefs, produces the marginal density describing flow in Equation 5. In this sense, the generative model embodies prior beliefs about the evolution of hidden and control states and how these states create the sensorium in a probabilistic sense. The prior beliefs concern the consequences of action and – through the minimisation of free energy by action – constitute self-fulfilling beliefs about the consequences of behaviour. So what form might these beliefs take?

If we consider any system or agent that has adapted to (or has adapted) its environment, then its ergodic density will have low entropy, which is upper bounded by the measure or volume of its random dynamical attractor (Friston, 2012). In other words, it will engage with its environment in a structured and organised fashion, revisiting attracting sets of states time and time again in an itinerant fashion. In computational biology and physics, this is known as a nonequilibrium steady-state (Jarzynski, 1997; Tomé, 2006). One can express the entropy of the ergodic density in terms of the entropy of sensory states (and their Markov blanket) and the entropy of external states, given the state of the Markov blanket:

$$\begin{aligned} H(\Psi, S, A, R) &= H(S, A, R \mid m) + H(\Psi \mid S, A) \\ &= \underbrace{E_t[-\ln p(\tilde{s}, \tilde{a}, \tilde{r} \mid m)]}_{\text{expected surprise}} + \underbrace{E_t[H(\Psi \mid S = \tilde{s}, A = \tilde{a})]}_{\text{expected uncertainty}} \end{aligned} \qquad 11$$

Because the system is ergodic, these two components correspond to the expected surprise and posterior uncertainty over time. Surprise is just the negative log marginal density or Bayesian model evidence that is implicitly minimised by ergodic flow, while the expected uncertainty is the entropy of the posterior distribution over external states. Although these expressions may look complicated, they say something quite intuitive; namely, agents that have adapted to (or have adapted) their environment are generally not surprised about their sensory samples and are confident about the causes of those samples. However, there is a problem here.

Recall from the previous section that the ergodic flow only suppresses surprise or its free energy bound (Equation 9). This means there is no guarantee that the causes of sensations will have low entropy. Indeed, we saw in the previous section that action only suppresses the entropy of sensory states and their Markov blanket – it does not affect the entropy of external states. In other words, ergodic systems could, in principle, have high entropy and dissipate themselves over large volumes of state space. So what properties (or beliefs) must low entropy systems possess?

We know that external states depend on action. This means it is sufficient for the agent to have prior beliefs (entailed in its kinetics, wiring or functional architecture) that the consequences of action will create a low entropy environment. However, action only minimises surprise. Therefore agents with low entropy ergodic densities must find high entropy distributions surprising:

$$\underbrace{-\ln p(\tilde{\psi}_u \mid \tilde{a})}_{\text{prior surprise}} = \gamma \cdot \underbrace{H[p(\tilde{\psi}_v(t+\tau) \mid \tilde{\psi}_u, \tilde{s}, \tilde{a})]}_{\text{posterior uncertainty}} \approx \gamma \cdot \underbrace{H[q(\tilde{\psi}_v(t+\tau) \mid \tilde{\psi}_u, \tilde{r}_v)]}_{\text{expected uncertainty}} \tag{12}$$

This expression says that the most likely control states are those that minimise posterior entropy or uncertainty about hidden states in the future. The intuition behind this rests on appreciating that the flow decreases entropy production, while the dispersive effects of random fluctuations increase entropy production. In nonequilibrium steady-state, these two effects are balanced and the entropy does not change. This balance can be nuanced by control states, to ensure the flow of hidden states decreases entropy production – by directing the flow to regions of high ergodic density. In short, inherent in the generative model of a low entropy system, there are prior beliefs that control states are unlikely to engender posterior uncertainty. The second equality above uses the fact that the variational density is the posterior density over hidden states. This means expected uncertainty can be approximated by evaluating the entropy of hidden states in the future, for any given (trajectory of) control states.

The constant of proportionality in the above equation can be interpreted as the precision of prior beliefs about control – and itself has an optimal value that maximises free energy. This precision has been discussed (in the slightly different of discrete state space Markov decision problems) in terms of sensitivity or inverse temperature in (softmax) choice rules that predominate in optimal decision theory and economics – and also in terms of dopamine discharges in neurobiology (Friston, Schwartenbeck, FitzGerald, Moutoussis, Behrens, & Raymond J. Dolan, 2013).

There are three interesting technical points about the (inherently optimistic) prior beliefs about control states. First, notice that while they speak to a minimisation of uncertainty under posterior beliefs (Equation 12), the posterior beliefs *per se* are trying to maximise posterior uncertainty (Equation 6). This creates a dialectic that can only be resolved by sampling the sensorium in a way that minimises uncertainty subject to Laplace's principle of indifference. Second, notice that the posterior uncertainty is about future states. This immediately brings active inference into the realm of anticipation and planning, which we will see illustrated below in terms of anticipatory eye movements. Finally, we have a rather unique probabilistic construction in which prior beliefs are a functional of posterior beliefs, where posterior beliefs depend upon prior beliefs. This circular dependency is illustrated in the right panel of Figure 1 and underwrites the (mindful) control of external states.

The prior beliefs above are functions of control states and functionals of beliefs about future hidden states. The uncertainty or entropy that quantifies these prior beliefs can be regarded as the *salience* of fictive outcomes, where a salient outcome resolves uncertainty (entropy) associated with posterior beliefs about the causes of sensations. In the illustrations below, salience will be a function of where a visual image could be sampled:

Friston, K. (2017). I am therefore I think. In M. Leuzinger-Bohleber, S. Arnold & M. Solms (Eds.), *The unconscious : a bridge between psychoanalysis and cognitive neuroscience* (pp. 113-137). London: Routledge.

$\sigma(\tilde{\psi}_u) = -H[q(\tilde{\psi}_v(t+\tau) \mid \tilde{\psi}_u, \tilde{r}_v)]$. This leads to the notion of a salience map that constitutes prior beliefs about the outcomes of visual palpation.

In summary, we have seen that agents can be the authors of their own phenotype – if we interpret the Lagrangian that underwrites their dynamics in terms of beliefs: in particular, prior beliefs about how the sensorium will be sampled. In practical terms, this means one can simulate self-organisation in one of two ways. First, one could write down some equations of motion and examine the emergent (nonequilibrium steady-state) behaviour – that could be described with a Lagrangian or ergodic density. Alternatively, one can write down the Lagrangian and simulate the same behaviour using a gradient ascent on the ensuing free energy. The crucial difference is that one can prescribe behaviour in terms of a Lagrangian that describes the flow – as opposed to describing the flow that prescribes a Lagrangian. It is in this sense – of prescribing behaviour in terms of Lagrangians based on prior beliefs – that we can regard agents as the autopoietic scribes of their embodied exchange with the world. In the final section, we will try to illustrate this in terms of active sampling of the visual world and demonstrate some phenomena that come close to unconscious and possibly conscious inference.

4 Simulating saccadic searches

This section reproduces the simulations of sequential eye movements described in (Friston, Adams, Perrinet, & Breakspear, 2012) to illustrate the theory of the previous section. Saccadic eye movements are a useful vehicle to illustrate active inference because they speak directly to visual search strategies and a wealth of psychophysical, neurobiological and theoretical study; e.g., (Tatler, Hayhoe, Land, & Ballard, 2011; Wurtz, McAlonan, Cavanaugh, & Berman, 2011; Shires, Joshi, & Basso, 2010; Bisley & Goldberg, 2010; Ferreira, Apel, & Henderson, 2008; Grossberg, Roberts, Aguilar, & Bullock, 1997; Srihasam, Bullock, & Grossberg, 2009). Having said this, we will focus on the basic phenomenology of saccadic eye movements to illustrate the key features of the active inference scheme described above. This scheme can be regarded as a formal example of active vision (Wurtz, McAlonan, Cavanaugh, & Berman, 2011); sometimes described in enactivist terms as visual palpation (O'Regan & Noë, 2001). People interested in the neurobiological aspects of active inference will find accessible introduction to predictive coding in (Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012).

This section first describes the production of visual signals and how they are modelled. We focus on a fairly simple paradigm – the categorisation of faces – to illustrate the behaviour induced by prior beliefs that constitute a generative model. Specifying a generative model allows us to simulate self-organised behaviour by specifying the dynamics of external states, internal states and their Markov blanket in terms of a Lagrangian that entails prior beliefs about minimising uncertainty. In other words, we will integrate the equations of motion in Equations (1) and (6) and interpreting the resulting dynamics (see Figure 3 for the form of these equations):

$$\dot{\tilde{x}} = f(\tilde{x}) + \omega : \quad f(\tilde{x}) = \begin{cases} f_{\psi}(\tilde{\psi}, \tilde{s}, \tilde{a}) \\ f_{s}(\tilde{\psi}, \tilde{s}, \tilde{a}) \\ (Q - \Gamma)\nabla_{\tilde{r}} F(\tilde{s}, \tilde{a}, \tilde{r}) \\ (Q - \Gamma)\nabla_{\tilde{a}} F(\tilde{s}, \tilde{a}, \tilde{r}) \end{cases} \qquad 13$$

Clearly, to perform the simulations we have to specify the equations of motion for external and states. This corresponds to specifying the nature of the processes generating sensory samples. In addition, we have to specify the Lagrangian $L(\tilde{x})$ that determines the free energy and subsequent dynamics of perception and action.

In what follows (and generally), the divergence free component of flow is chosen to simply update the (generalised) states, given their generalised motion: $Q\nabla_{\tilde{x}} F = D \cdot \tilde{x}$, where $D \cdot \tilde{x} = (x', x'', \ldots)$ is a derivative operator (Friston, Stephan, Li, & Daunizeau, 2010). One interesting issue that will emerge from these simulations is that the external states generating sensations are not necessarily the states assumed by a generative model.

4.1 The generative process

In these simulations sensory signals are generated in two modalities – proprioception and vision. Proprioception reports the centre of gaze as a displacement from the origin of some extrinsic frame of reference. The visual input corresponds to an array of sensory channels sampling a two-dimensional image or visual scene $I : \mathbb{R}^2 \to \mathbb{R}$. This sampling uses a grid of 16 x 16 channels that samples a small part the image – like a foveal sampling. To make this sampling more biologically realistic, each visual channel was equipped with a centre-surround receptive field that reports a local weighted average of the image.

The only changing external state $\psi \in \mathbb{R}^2$ describes the centre of oculomotor fixation, whose motion is driven by action (and decays with a time constant of 16 time bins of 12 ms). This external state determines where the visual scene is sampled (foveated). The proprioceptive and visual signals were effectively noiseless, where there random fluctuations had a log precision of 16. The motion of the fixation point was subject to low amplitude fluctuations, with a log precision of eight. This completes our description of the process generating proprioceptive and visual signals. We now turn to the model of this process that generates action.

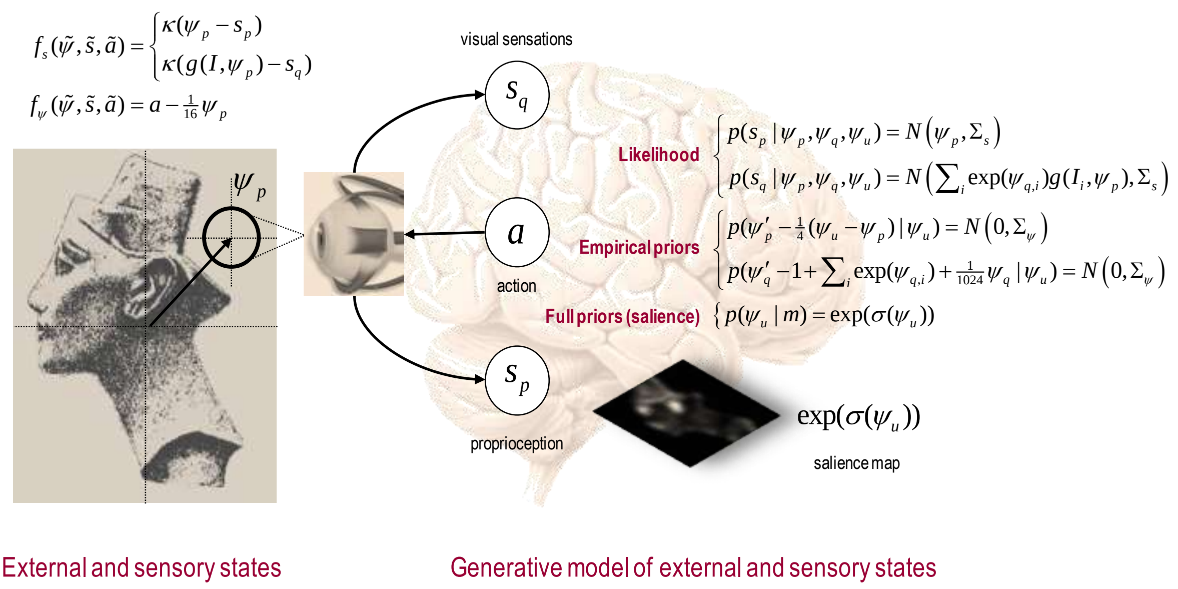

$f_s(\tilde{\psi},\tilde{s},\tilde{a}) = \begin{cases} \kappa(\psi_p - s_p) \\ \kappa(g(I,\psi_p) - s_q) \end{cases}$
$f_\psi(\tilde{\psi},\tilde{s},\tilde{a}) = a - \frac{1}{16}\psi_p$
ψ_p
visual sensations
s_q
a
action
s_p
proprioception
Likelihood
$\begin{cases} p(s_p \mid \psi_p,\psi_q,\psi_u) = N\left(\psi_p,\Sigma_s\right) \\ p(s_q \mid \psi_p,\psi_q,\psi_u) = N\left(\sum_i \exp(\psi_{q,i}) g(I_i,\psi_p),\Sigma_s\right) \end{cases}$
Empirical priors
$\begin{cases} p(\psi'_p - \frac{1}{4}(\psi_u - \psi_p) \mid \psi_u) = N\left(0,\Sigma_\psi\right) \\ p(\psi'_q - 1 + \sum_i \exp(\psi_{q,i}) + \frac{1}{1024}\psi_q \mid \psi_u) = N\left(0,\Sigma_\psi\right) \end{cases}$
Full priors (salience)
$\{ p(\psi_u \mid m) = \exp(\sigma(\psi_u))$
$\exp(\sigma(\psi_u))$
salience map
External and sensory states
Generative model of external and sensory states

***Figure 3*: *generative models of visual search*:** *This schematic (left panel) provides the equations of motion generating sensory states (in proprioceptive and visual modalities). Visual input is generated from an image that can be sampled locally by foveating a particular point – an external state of the world. The generative model (right panel) is provided in terms of likelihood, empirical prior and full prior densities. The proprioceptive likelihood is based on a noisy version of the expected eye position, while the visual input is generated from a number of potential images or hypotheses. Their relative weights are (nonnegative) perceptual hidden states, whose dynamics ensure they always sum to unity. The full priors are determined by a map of salience that approximates the posterior confidence in the inferred hidden states (here, the perceptual states encoding the competing images). See main text and (Friston, Adams, Perrinet, & Breakspear, 2012) for further details.*

4.2 The generative model

As in the generative process, proprioceptive signals are just a noisy mapping from external proprioceptive states encoding the direction of gaze. Visual input is modelled as a mixture of images sampled at a location specified by the hidden proprioceptive state. This hidden state decays with a time constant of four time bins (48 ms) towards a hidden control state. In other words, the hidden control determines the location that attracts gaze – in a way not dissimilar to the equilibrium point hypothesis for motor reflexes (Feldman & Levin, 1995). Crucially, in the model, visual input depends on a number of hypotheses or internal images $I_i : \mathbb{R}^2 \to \mathbb{R} : i \in \{1, \ldots N\}$ that constitute the agent's prior beliefs about what could cause its sensations. The input encountered at any particular time is a weighted mixture of these internal images, where the weights correspond to hidden perceptual states. The dynamics of these perceptual states implement a form of dynamic softmax, in the sense that the solution of their equations of motion ensures the weights sum to one. This means we can interpret the (hidden) perceptual states as the (softmax) probability that the *i*-th internal image or hypothesis is the cause of visual input.

In summary, given hidden proprioceptive and perceptual states the agent can predict its proprioceptive and visual input. The generative model is specified by these predictions and the amplitude of the random fluctuations that determine the agent's prior certainty about sensory inputs and the motion of hidden states. In the examples below, we used a log precision of eight for proprioceptive sensations and let the agent believe its visual input was fairly noisy, with a log precision of four. All that now remains is to specify prior beliefs about the hidden control state attracting the centre of gaze, which we have already established can be quantified in terms of salience – or the reduction of uncertainty (c.f., information gain) associated with each control state.

4.3 Priors and saliency

To simulate saccadic eye movements, we integrated the active inference scheme for 16 time bins (196 ms) and then computed a map of salience to update the prior expectations about the hidden control states that attract the centre of gaze. This was repeated eight times to give a sequence of eight saccadic eye movements. The salience was computed for $1024 = 32 \times 32$ locations distributed uniformly over the visual image or scene according to Equation 11 (where the expected uncertainty was evaluated at the location prescribed by the control state). In other words, salience was evaluated under current (posterior) beliefs about the content of the visual scene for all allowable points of fixation. The ensuing salience over the 32×32 locations constitutes a salience map whose peak drives the next saccade (by reflexively engaging action through the proprioceptive consequences of the most likely control state). Notice that salience is a function of, and only of, fictive beliefs about the state of the world and essentially tells the agent where to sample (look) next to confirm its suspicions about the causes of its sensations.

Figure 4 provides a simple illustration of salience based upon the posterior beliefs or hypothesis that local (foveal) visual inputs are caused by an image of Nefertiti. The right panel summarises the classic results of the Yarbus (Yarbus, 1967) in terms of an image and the eye movements it elicits. The left panels depict visual input after sampling the image on the right with centre-surround receptive fields. Note how the receptive fields suppress absolute levels of luminance contrast and highlight edges. It is these edges that inform posterior beliefs about both the content of the visual scene and where it is being sampled. This information reduces posterior uncertainty and is therefore salient. The salient features of the image (middle panel) include the ear, eye and mouth. The location of these features and a number of other salient locations appear to be consistent with the locations that attract saccadic eye movements (as shown on the right). Crucially, the map of salience extends well beyond the field of view (circle on the picture). This reflects the fact that salience is not an attribute of what is seen, but what might be seen under a particular hypothesis about the causes of sensations.

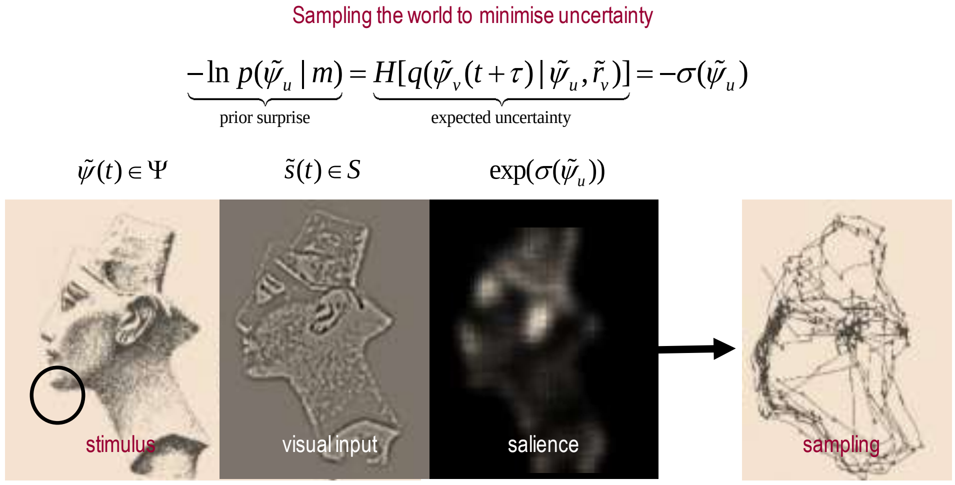

Sampling the world to minimise uncertainty
$\underbrace{-\ln p(\tilde{\psi}_u \mid m)}_{\text{prior surprise}} = \underbrace{H[q(\tilde{\psi}_v(t+\tau) \mid \tilde{\psi}_u, \tilde{r}_v)]}_{\text{expected uncertainty}} = -\sigma(\tilde{\psi}_u)$
$\tilde{\psi}(t) \in \Psi$
$\tilde{s}(t) \in S$
$\exp(\sigma(\tilde{\psi}_u))$
stimulus
visual input
salience
sampling

***Figure 4**: **Salience and visual searches**. This schematic provides an illustration of salience based upon the posterior beliefs or hypothesis that local (foveal) visual inputs are caused by an image of Nefertiti. The right panel summarises the classic results of Yarbus – in terms of a stimulus and the eye movements it elicits. The left panels depict visual input after sampling the image on the far right (using conventional centre surround receptive fields) and the associated saliency map based on a local sampling of 16 x 16 pixels (using the generative model described in the main text). The size of the field of view, in relation to the visual scene, is indicated by the circle on the left image. The key thing to note here is that the salient features of the image include the ear, eye and mouth. The locations of these features appear to be consistent with the locations that attract saccadic eye movements (as shown on the right).*

To make the simulations a bit more realistic, we added a further prior implementing inhibition of return (Wang & Klein, 2010; Itti & Koch, 2001). This involved suppressing the salience of locations that had been recently foveated. The addition of inhibition of return ensures that a new location is selected by each saccade and can be motivated ethologically by prior beliefs that the visual scene will change and that previous locations should be revisited.

In summary, we have described the process generating sensory information in terms of a visual scene and hidden states that specify how that scene is sampled. We have described both the likelihood and priors that together comprise a generative model. The special consideration here is that these priors are based upon beliefs that agents will sample salient sensory features based upon its current posterior beliefs about the causes of those features. We are now in a position to look at the sorts of behaviour this model produces.

4.4 Simulating saccadic eye movements

We conclude with a realisation of visual search under the generative model described above. Our purpose here is to illustrate the nature of active inference, when it is equipped with priors that maximise salience or minimise uncertainty. Figure 5 shows the results of a simulation, in which the agent had three internal images or hypotheses about the scene that it might sample (an upright face, an inverted face and a rotated face). The agent was presented with an upright face and its posterior expectations were evaluated over 16 (12 ms) time bins, after which salience was evaluated. The agent then emitted a saccade by foveating the most salient location during the subsequent 16 time bins. This was repeated for eight saccades.

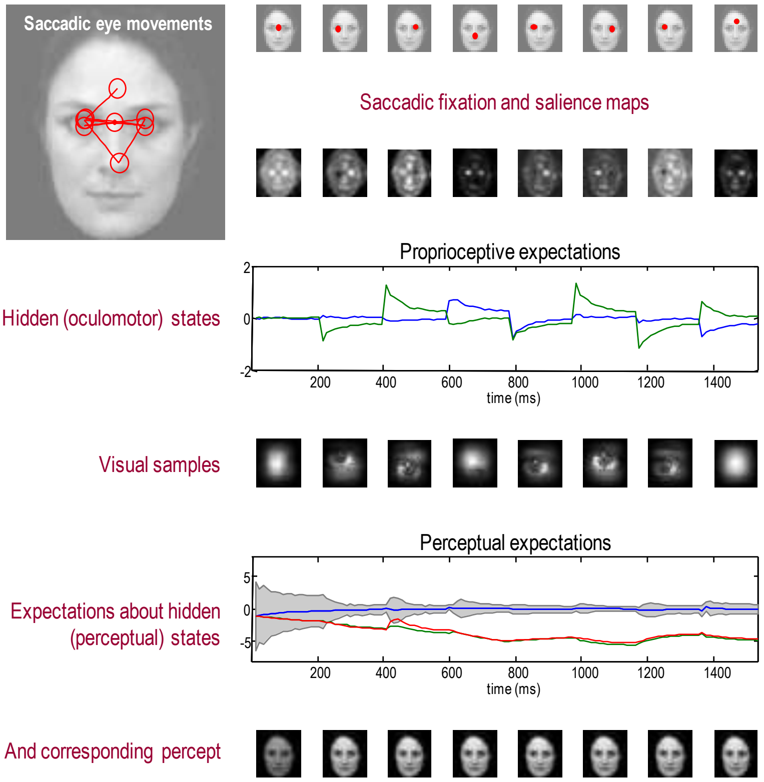


***Figure 5**: **The dynamics of perception**. This figure shows the results of a simulation, in which a face was presented to an agent whose responses were simulated using the inference scheme described in the main text. In this simulation, the agent had three internal images or hypotheses about the stimuli it might sample (an upright face, and inverted face and a rotated face). The agent was presented with an upright face and its posterior expectations were evaluated over 16 (12 ms) time bins until the next saccade was emitted. This was repeated for eight saccades. The ensuing eye movements are shown as red dots at the end of each saccade in the upper row. The corresponding sequence of eye movements is shown in the insert on the upper left, where the red circles correspond roughly to the proportion of the image sampled. These saccades are driven by prior beliefs about the direction of gaze based upon the saliency maps in the second row. Note that these maps change with successive saccades as posterior beliefs about the hidden states become more confident. These posterior beliefs provide both visual and proprioceptive predictions that drive eye movements. Oculomotor responses are shown in the third row in terms of the two hidden oculomotor (proprioceptive) states corresponding to vertical and horizontal displacements. The associated portions of the image sampled (at the end of each saccade) are shown in the fourth row. The penultimate row shows the posterior beliefs in terms of their sufficient statistics, namely posterior expectations and the 90% confidence interval about the true stimulus (grey area). The final row shows the percept that is implicitly selected (weighted by its uncertainty).*

The upper row shows the ensuing eye movements as red dots at the fixation point of each saccade. The corresponding sequence of eye movements is shown in the insert on the upper left, where the red circles correspond roughly to the agent's field of view. These saccades were driven by prior beliefs about the direction of gaze based upon the salience maps in the second row. Note that these maps change with successive saccades as posterior beliefs about the hidden perceptual states become more confident. Note also that salience is depleted in locations that were foveated in the previous saccade – this reflects the inhibition of return. Posterior beliefs about hidden states provide visual and proprioceptive predictions that drive eye movements. Oculomotor responses are shown in the third row in terms of the two hidden proprioceptive states corresponding to vertical and horizontal eye displacements. The portions of the image sampled (at the end of each saccade) are shown in the fourth row. The penultimate row shows the posterior beliefs in terms of their posterior expectations and 90% confidence interval about the true stimulus. The key thing to note here is that the expectation about the true stimulus supervenes over its competing representations and, as a result, posterior confidence about the stimulus category increases (the posterior confidence intervals shrink to the expectation): see (Churchland, Kiani, Chaudhuri, Wang, Pouget, & Shadlen, 2011) for an empirical study of this sort of phenomena. The images in the lower row depict the hypothesis selected, where their intensity has been scaled to reflect conditional uncertainty, using the entropy (average uncertainty) of the softmax probabilities.

This simulation illustrates a number of key points. First, it illustrates the nature of evidence accumulation in selecting a hypothesis or percept the best explains sensory data. One can see that this proceeds over two timescales; both within and between saccades. Within-saccade accumulation is evident even during the initial fixation, with further stepwise decreases in uncertainty as salient information is sampled. The within-saccade accumulation is formally related to evidence accumulation as described in models of perceptual discrimination (Gold & Shadlen, 2003; Churchland, Kiani, Chaudhuri, Wang, Pouget, & Shadlen, 2011) This is meant in the sense that the posterior expectations about perceptual states are driven by sensory information. However, the accumulation here rests explicitly on the priors implied by the generative model. In this case, the prevalence of any particular perceptual category is modelled as a dynamical process that has certain continuity properties. In other words, inherent in the model is the belief that the content of the world changes in a continuous fashion. This is reflected in the progressive elevation of the correct perceptual expectation above its competitors and the consequent shrinking of the posterior confidence interval. The transient changes in the posterior beliefs, shortly after each saccade, reflect the fact that new data are being generated as the eye sweeps towards its new target location. It is important to note that the agent is not just modelling visual contrast, but also how contrast changes with eye movements – this induces an increase in conditional uncertainty during the fast phase of the saccade. However, due to the veracity of the posterior beliefs, the posterior confidence shrinks again when the saccade reaches its target location. This shrinkage is usually to a smaller level than in the previous saccade.

This illustrates the second key point; namely, the circular causality that lies behind perceptual sampling. Put simply, the only hypothesis that can endure over successive saccades is the one that correctly predicts the salient features that are sampled. This sampling depends upon action or an embodied inference that speaks directly to the notion of visual palpation (O'Regan & Noë, 2001). This means that the hypothesis prescribes its own verification and can only survive if it is a correct representation of the world. If its salient features are not discovered, it will be discarded in favour of a better hypothesis. This provides a nice perspective on perception as hypothesis testing (Gregory, 1980; Kersten, Mamassian, & Yuille, 2004), where the emphasis is on the selective processes that underlie sequential testing. This is particularly pertinent when hypotheses can make predictions that are more extensive than the data available at any one time.

Finally, although the majority of saccades target the eyes and nose, as one might expect, there is one saccade to the forehead. This is somewhat paradoxical, because the forehead contains no edges and cannot increase posterior confidence about a face. However, this region is highly informative under the remaining two hypotheses (corresponding to the location of the nose in the inverted face and the left eye in the rotated face). This subliminal salience is revealed through inhibition of return and reflects the fact that the two competing hypotheses have not been completely excluded. This illustrates the competitive nature of perceptual selection induced by inhibition of return and can regarded, heuristically, as occasional checking of alternative hypotheses. This is a bit like a scientist who tries to refute his hypothesis by acquiring data that furnish efficient tests of his competing or null hypotheses.

In summary, we have seen an illustration of active inference that emerges when prior beliefs about controlled sampling of the sensorium reduce posterior or expected uncertainty. It is interesting to consider the sampling of sensory input, not by saccadic eye movements, but by rapid redeployment of attentional foci. In this setting, simulations of selective attention (Graboi & Lisman, 2003) leads to exactly the same conclusion; namely, "that during recognition of complex visual stimuli [exemplified by words], the window of attention undergoes *rapid covert movements to stimulus regions that are rich in information relevant to recognition*" (our italics).

5 Conclusion

In conclusion, starting with some basic considerations about the ergodic behaviour of random dynamical systems, we have seen how inference could be construed as emergent property of any weakly mixing (random dynamical) system – and how it can be described in terms of a generalized descent on variational free energy. Using this formalism, we have been able to address some fairly abstract issues in the philosophy of realisation and the Cartesian divide between the material (ergodic flow) and the immaterial (beliefs entailed by the flows Lagrangian). Some relatively simple (*reductio ad absurdum*) arguments about the Lagrangian of ergodic systems lead to the notion that any autopoietic or self-organising system implicitly entertains prior beliefs that they will sample the sensorium to minimise their uncertainty about its causal structure. These somewhat abstract arguments were unpacked using simulations of saccadic searches, which have been previously reported to illustrate the computational anatomy of embodied (active) perceptual inference.

The take-home message of this work is that the process of unconscious – and indeed conscious inference – may conform to the same basic principles that underlie self-organisation in any system with coupled dynamics. The emergence of intentional phenomena rests upon the notion of a Markov blanket that separates internal states from external states. The very presence of this separation implies a generalized synchrony (Huygens, 1673; Hunt, Ott, & Yorke, 1997) between external (e.g., environmental) and internal (e.g., neuronal) states that will appear to be lawful – in the sense that internal states minimize the same free energy functional used for Bayesian inference. This lends a quintessentially inferential or predictive aspect to internal states that has many of the hallmarks of cognition and consciousness; particularly when one considers that the agent is the author of its behaviour.

Crucially, this inference or assimilation is active, in the sense that the internal states affect the causes of sensory input vicariously through action. The resulting circular causality between perception and action fits comfortably with many formulations in embodied cognition and artificial intelligence; for example, the perception action cycle (Fuster, 2004), active vision (Wurtz, McAlonan, Cavanaugh, & Berman, 2011; Shen, Valero, Day, & Paré, 2011), the use of predictive information (Ay, Bertschinger, Der, Güttler, & Olbrich, 2008; Bialek, Nemenman, & Tishby, 2001; Tishby & Polani, 2010) and homeokinetic formulations (Soodak & Iberall, 1978). Furthermore, it connects these perspectives to more general treatments of circular causality and autopoiesis in cybernetics and synergetics (Haken, 1983; Maturana & Varela, 1980). Remarkably, these conclusions were articulated by Helmholtz over a century ago (von Helmholtz, 1971):

"Each movement we make by which we alter the appearance of objects should be thought of as an experiment designed to test whether we have understood correctly the invariant relations of the phenomena before us, that is, their existence in definite spatial relations." Herman von Helmholtz (1878) p.384

Acknowledgments: This work was funded by the Wellcome trust (Ref: 088130/Z/09/Z) and was originally presented at the JOSEPH SANDLER PSYCHOANALYTIC RESEARCH CONFERENCE 2014.

REFERENCES

Ao, P. (2004). Potential in stochastic differential equations: novel construction. *J Phys. A: Math Gen. , 37*, L25–30.

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *J Gen Psychology. , 37*, 125-8.

Auletta, G. (2013). Information and Metabolism in Bacterial Chemotaxis. *Entropy , 15* (1), 311-26.

Ay, N., Bertschinger, N., Der, R., Güttler, F., & Olbrich, E. (2008). Predictive Information and Explorative Behavior of Autonomous Robots. *European Physical Journal B , 63*, 329-39.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication* (pp. 217-34). Cambridge, MA: MIT Press.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron , 76* (4), 695-711.

Beal, M. J. (2003). Variational Algorithms for Approximate Bayesian Inference. *PhD. Thesis, University College London* .

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computat. , 13* (11), 2409-63.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proc Natl Acad Sci USA , 17*, 656-60.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annu Rev Neurosci. , 33*, 1-21.

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X. J., Pouget, A., & Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron , 69* (4), 818-31.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci. , 36* (3), 181-204.

Conant, R. C., & Ashby, R. W. (1970). Every Good Regulator of a system must be a model of that system. *Int. J. Systems Sci. , 1* (2), 89-97.

Crauel, H. (1999). Global random attractors are uniquely determined by attracting deterministic compact sets. *Ann. Mat. Pura Appl. , 4* (176), 57-72.

Crauel, H., & Flandoli, F. (1994). Attractors for random dynamical systems. *Probab Theory Relat Fields , 100*, 365-393.

Dayan, P., Hinton, G. E., & Neal, R. (1995). The Helmholtz machine. *Neural Computation , 7*, 889-904.

Evans, D. J., & Searles, D. J. (1994). Equilibrium microstates which generate second law violating steady states. *Physical Review E , 50* (2), 1645–8.

Feldman, A. G., & Levin, M. F. (1995). The origin and use of positional frames of reference in motor control. *Behav Brain Sci. , 18*, 723-806.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends Cogn Sci. , 12* (11), 405-10.

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

Frank, T. D. (2004). *Nonlinear Fokker-Planck Equations: Fundamentals and Applications. Springer Series in Synergetics.* Berlin: Springer.

Freeman, W. J. (1994). Characterization of state transitions in spatially distributed, chaotic, nonlinear, dynamical systems in cerebral cortex. *Integr Physiol Behav Sci. , 29* (3), 294-306.

Friston, K. (2012). A Free Energy Principle for Biological Systems. *Entropy , 14*, 2100-2121.

Friston, K. J. (2013). Life as we know it. *Royal Society Interface* , in the press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci. , 11* (2), 127-38.

Friston, K., & Ao, P. (2012). Free-energy, value and attractors. *Computational and mathematical methods in medicine , 2012*, 937860.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front Psychol. , 3*, 151.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *J Physiol Paris. , 100* (1-3), 70-87.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Raymond J. Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front Hum Neurosci. , 7*, 598.

Friston, K., Stephan, K., Li, B., & Daunizeau, J. (2010). Generalised Filtering. *Mathematical Problems in Engineering , vol., 2010*, 621670.

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends Cogn Sci. , 8* (4), 143-5.

Gold, J. I., & Shadlen, M. N. (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *J Neurosci. , 23* (2), 632-51.

Graboi, D., & Lisman, J. (2003). Recognition by top-down and bottom-up processing in cortex: the control of selective attention. *J Neurophysiol. , 90* (2), 798-810.

Gregory, R. L. (1980). Perceptions as hypotheses. *Phil Trans R Soc Lond B. , 290*, 181-197.

Grossberg, S., Roberts, K., Aguilar, M., & Bullock, D. (1997). A neural model of multimodal adaptive saccadic eye movement control by superior colliculus. *J Neurosci. , 17* (24), 9706-25.

Haken, H. (1983). *Synergetics: An introduction. Non-equilibrium phase transition and self-selforganisation in physics, chemistry and biology* (3rd ed.). Berlin: Springer Verlag.

Helmholtz, H. (1866/1962). Concerning the perceptions in general. In *Treatise on physiological optics* (J. Southall, Trans., 3rd ed., Vol. III). New York: Dover.

Hinton, G. E., & van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of weights. *Proceedings of COLT-93* , 5-13.

Hohwy, J. (2013). *The Predictive Mind.* Oxford: Oxford University Press.

Hunt, B., Ott, E., & Yorke, J. (1997). Differentiable synchronisation of chaos. *Phys Rev E , 55*, 4029-4034.

Huygens, C. (1673). *Horologium Oscillatorium.* France: Parisiis.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nat Rev Neurosci. , 2* (3), 194-203.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett. , 78*, 2690 .

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review Series II , 106* (4), 620–30.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Stat Assoc. , 407*, 717-26.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol. , 55*, 271-304.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics , 22* (1), 79-86.

Linsker, R. (1990). Perceptual neural organization: some approaches based on network models and information theory. *Annu Rev Neurosci. , 13*, 257-81.

Maturana, H. R., & Varela, F. (1980). Autopoiesis: the organization of the living. In V. F. Maturana HR (Ed.), *Autopoiesis and Cognition.* Dordrecht, Netherlands: Reidel.

Nicolis, G., & Prigogine, I. (1977). *Self-organization in non-equilibrium systems.* New York: John Wiley.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav Brain Sci. , 24* (5), 939-73.

Pasquale, V., Massobrio, P., Bologna, L. L., Chiappalone, M., & Martinoia, S. (2008). Self-organization and neuronal avalanches in networks of dissociated cortical neurons. *Neuroscience , 153* (4), 1354-69.

Pearl, J. (1988). *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference.* San Fransisco, CA, USA: Morgan Kaufmann.

Schrödinger, E. (1944). *What Is Life? : The Physical Aspect of the Living Cell.* Trinity College, Dublin. Dublin: Trinity College, Dublin.

Shen, K., Valero, J., Day, G. S., & Paré, M. (2011). Investigating the role of the superior colliculus in active vision with the visual search paradigm. *Eur J Neurosci. , 33* (11), 2003-16.

Shires, J., Joshi, S., & Basso, M. A. (2010). Shedding new light on the role of the basal ganglia-superior colliculus pathway in eye movements. *Curr Opin Neurobiol. , 20* (6), 717-25.

Soodak, H., & Iberall, A. (1978). Homeokinetics: A Physical Science for Complex Systems. *Science , 201*, 579-582.

Srihasam, K., Bullock, D., & Grossberg, S. (2009). Target selection by the frontal cortex during coordinated saccadic and smooth pursuit eye movements. *J Cogn Neurosci. , 21* (8), 1611-27.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *J Vis. , 11* (5), 5.

Tishby, N., & Polani, D. (2010). Information Theory of Decisions and Actions. In V. Cutsuridis, A. Hussain, & J. and Taylor (Eds.), *Perception-reason-action cycle: Models, algorithms and systems.* Berlin: Springer.

Tomé, T. (2006). Entropy Production in Nonequilibrium Systems Described by a Fokker-Planck Equation. *Brazilian Journal of Physics , 36* (4A), 1285-1289.

van Leeuwen, C. (1990). Perceptual-learning systems as conservative structures: is economy an attractor? *Psychol Res. , 52* (2-3), 145-52.

von Helmholtz, H. (1971). The Facts of Perception (1878). In R. Karl (Ed.), *The Selected Writings of Hermann von Helmholtz.* Middletown, Connecticut: Wesleyan University Press.

Wang, Z., & Klein, R. M. (2010). Searching for inhibition of return in visual search: a review. *Vision Res. , 50* (2), 220-8.

Wurtz, R. H., McAlonan, K., Cavanaugh, J., & Berman, R. A. (2011). Thalamic pathways for active vision. *Trends Cogn Sci. , 5* (4), 177-84.

Yarbus, A. L. (1967). *Eye Movements and Vision.* New York: Plenum.

Yuan, R., Ma, Y., Yuan, B., & Ping, A. (2010, Dec 13). Bridging Engineering and Physics: Lyapunov Function as Potential Function. *arXiv:1012.2721v1 [nlin.CD]* .