Free Energy, Bistable Transitions, and Binary Disambiguation

Matt Rounds

August 2015

Abstract

We build a system that samples an image by acting to minimise Free Energy, and by doing so minimises the expected entropy of future world states. We test firstly how the system acts to disambiguate between competing prior expectations of possible world states, and secondly to see if the system can replicate a well known result in bistable perception. In the first case, the system was sufficiently convinced by Kanizsa's triangle to conclude it was looking at a triangle. In the second, the system failed to reproduce the expected distribution of percepts. We suggest that a full deep learning hierarchy might succeed where our approximation did not, and make minor hypotheses about a suitably primed subject's likely points of foveation when presented with Kanizsa's triangle.

Acknowledgements

I would like to thank Professor Anil Seth and Dr Chris Buckley for their unceasing enthusiasm and invaluable support over the last several months.

I would also like to thank Mr Jan Metzger for his generosity in funding the Metzger Scholarship, and the University of Sussex for funding the Chancellor's Masters Scholarship, both of which helped greatly in making this Master's degree possible.

Contents

Introduction **2**

1 Background: Free Energy **4**
1.1 Introduction 4
1.2 Perception as Inference 4
1.3 Introducing Free Energy 5
1.4 How the Brain Encodes World States W 8
1.5 Minimising Free Energy 10
1.6 Action as a Result of Minimisation 15
1.7 Decomposing Error Signals 17
1.8 The Dark Room Problem 18
1.9 Attention and Consciousness 19

2 System: Implementation **21**
2.1 Introduction 21
2.2 The System 22
2.3 Minimising the Free Energy of the System 24
2.4 Hidden Controls and Salience 28

3 System: Evaluation and Discussion **30**
3.1 Introduction 30
3.2 Disambiguation of Priors 30
3.3 Disambiguation Discussion 33
3.4 Distribution of rate of percept alternation 34
3.5 Bistability discussion 36

Conclusion **38**

Bibliography **39**

Appendix: Code **42**

Introduction

Intelligence is a consequence of the laws of physics. This assumption drives much of the modern endeavour to understand the brain and, by extension, intelligent behaviour. It is one of the reasons that accounts of intelligence that purport to offer a single explanation for a wide range of phenomena - perception, action, consciousness - are so appealing. However, this appeal is why they should be approached with a commensurate amount of caution.

This dissertation engages with one such account; Karl Friston's Free Energy Principle (**FEP**) [20][19][17]. The FEP has gained a great deal of traction over the last decade, having grown out of the work of Hinton and colleagues in the 1990s [14][27]. It can be thought of a particular instance of Predictive Processing (PP); the idea that brains are essentially prediction machines engaged in inference on the causes of their sensory input [29]. This idea of perception as inference can be traced back to Helmholtz [26], but Friston provides an ontological reworking. He argues that the minimisation of Free Energy, whilst equivalent to PP, is a result of the necessary minimisation of sensory entropy demanded of every organised entity by virtue of its existence. The process thus emerges from an imperative towards homeostasis, as described in the cybernetics of W. Ross Ashby and others [2][3].

Friston's account ties action into the same framework [23]. By minimising Free Energy, organisms reduce the disparity between their expected sensory input and what they actually experience. We will refer to this as **surprisal**, to distinguish it from the propositional attitude of surprise. Organisms minimise surprisal by either adjusting their expectations about the world to be more in line with their sensory experience, or acting to adjust their sensory experience so that it aligns with their expectations. This latter is called **active inference** [23].

Active inference might also take the form of the organism seeking out sensory data which will better disambiguate between competing gestalt expectations, rather than that which directly confirms a current expectation. 'Gestalt' because we assume that there are certain sets of brain states that combine, over multiple levels in the brain, the expectations connected with the various structural characteristics of what would be experienced as a unified object. Acting effectively relies on the predictive models encoding **counterfactual** expectations about the relationship between action and sensory input, even if some actions remain potential only. It has been argued that the phenomenal richness

of our experience of any particular gestalt is bound up in the counterfactual encoding of sensorimotor contingencies [38][37]; what it is like for me to perceive a die as square, for example, is a function of how I would expect my die-caused sensory input to change were I to pick it up and turn it round. This reinforces links between the FEP and older ideas of embodiment and sensorimotor contingency theory (SMC) [11].

To probe some of the claims the FEP, in this project we have built a Free Energy minimising system that simulates the visual sampling of an image to disambiguate between competing prior expectations of what the system might be looking at. Our aims were twofold: (A) to explore in depth how exactly a Free Energy minimising system would act to disambiguate between competing prior expectations, (B) to demonstrate whether the simplifications required to make a computationally tractable model rendered the resultant system empirically worthless. As a starting point, we have relied upon Friston et al *Perceptions as Hypotheses: Saccades as Experiments* [22]. Where our model differs in implementation, we have made this clear, and our use of the model to investigate (A) and (B) we believe to be unique.

We investigate (A) with respect to the simple case of Kanizsa's Triangle (Fig. 3.1), demonstrating both a system which acts to confirm and a system which acts to disconfirm the current hypothesis. There is a noticeable difference in which parts of the image are foveated, which suggests the potential for empirical follow-up. Our approach to (B) is to compare data from the model to a well known result in bistable perception in humans; that the rates of alternation between percepts follow a gamma distribution [8]. We initially attempt this with respect to the classic Rubin's Vase illusion (Fig. 3.5), and then on an image constructed to be ambiguous from the point of view of our system, given the simplifications we have made.

Chapter 1 outlines the necessary theoretical background to the project. It includes a derivation of the main mathematics of the FEP from first principles, and a short discussion which attempts to tie the mathematical results into a larger conceptual framework. Chapter 2 outlines the specific implementation of our system, and discusses what extra assumptions and simplifications were made. Chapter 3 details the experiments run, and presents the results which are most pertinent to the investigation. For both (A) and (B) it includes a succinct discussion of the implications and limitations of the investigation, along with recommendations for improvements.

Chapter 1

Background: Free Energy

1.1 Introduction

The FEP [20][19][17] is a model of brain dynamics which considers the brain to be fundamentally Bayesian, and purports to tie previous work on approximate Bayesian inference into a larger framework which encompasses both action and perception [23]. In doing so, it attempts to provide a unifying explanation for psychological phenomena as diverse as attention and learning. This chapter outlines the mathematical underpinnings of the minimisation which is at the core of the approach, and follows with a brief, and hopefully somewhat illuminating discussion.

1.2 Perception as Inference

The idea that perception can be thought of as inference on the causes of sensory data can be considered to originate with Helmholtz [26]. For an organism to successfully interact with the world, it must have a model of the world which is at least good enough to capture the statistics of those world states W which directly affect the organism. However, the states of the world are hidden from the organism, which only has access to its sensory data S.

This means that a complex organism's brain must maintain a **generative** model; i.e. a model of how supposed world states generate sensory data, which can then be inverted to infer the state of the world ω ($\omega \in W$) given some sensory data s ($s \in S$). This inversion can then be used to update the generative model such that the posteriors of earlier experiences form the priors of later ones. In an ideal world, this inference would be performed according to Bayes' Rule (1.1),

$$p(\omega|s) = \frac{p(s|\omega)p(\omega)}{p(s)} \tag{1.1}$$

where the posterior probability of some world state given some sense data is computed from the likelihood of that sense data given that world state and the

prior probabilities of both world state and sense data.

It should be noted that this proposed inference is statistical; the brain is representing the world and the potential inputs from the world as probability densities over possible states. This arises naturally from the fact that the brain is selecting from multiple competing hypotheses presented by an inverse model which is itself only as good an approximate as previous sensory experience allows it to be. The generative model can therefore be thought of as being represented by the brain in the generative density

$$p(\omega, s) \tag{1.2}$$

i.e. the joint density of world states and sensory states. Interestingly, one of the assumptions Friston makes in deriving a method of approximate inference is that organisms treat the world itself as essentially deterministic; it is organisms' subjective representations of world states which are treated as probabilistic. We will return to this in a moment.

The problem with ideal Bayesian inference, under Friston's account [20], is that the prior $p(s)$ is difficult, if not generally impossible, to compute. $p(s)$ can be extracted from the generative distribution by marginalisation:

$$p(s) = \int p(\omega, s) d\omega \tag{1.3}$$

or equivalently,

$$p(s) = \int p(s|\omega) p(\omega) d\omega \tag{1.4}$$

which involves integrating over all possible world states. This integral is likely to be analytically intractable due to the complexity of the world. In addition, it is also empirically intractable simply because an organism cannot access hidden world states W.

1.3 Introducing Free Energy

This means that we need a tractable approximation to ideal Bayesian inference. There are several ways to do this, but the method Friston extends is the Variational Bayes approach developed by Hinton and colleagues in the 1990s for several machine learning tasks [14][27].

Let us assume that the brain also encodes a second density, $q(\omega)$, which is its best current approximation to the true posterior density $p(\omega|s)$ from (1.1). Let us call this density the **recognition density**, and assume the brain acts to minimise some measure of the difference between it and the posterior. Finally, let us assume that this measure is the Kullback-Leibler divergence [12] between the two distributions, namely (1.5).

$$D_{KL} = \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega|s)} \tag{1.5}$$

By noting that

$$p(\omega|s) = \frac{p(\omega, s)}{p(s)} \tag{1.6}$$

and using the facts that $\ln(AB) = \ln(A) + \ln(B)$, and $\int d\omega q(\omega) = 1$, we can rewrite (1.5) as

$$D_{KL} = \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega, s)} + \ln p(s) \tag{1.7}$$

where the first term is the Kullback-Leibler divergence between the recognition density and the generative density (1.2) and the second term is the log probability of the sensory data. As discussed, $p(s)$ is intractable, which means the second term is impossible for the organism to minimise. However, the divergence term can be minimised. It is this information-theoretic quantity with which we shall concern ourselves for the rest of this report. We will call it 'Free Energy', or F.

$$F = \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega, s)} \tag{1.8}$$

Free Energy has two important properties which should be derived straight away. Firstly, it is a tight upper bound on surprisal $-\ln p(s)$, the measure of how unexpected some particular sense data is. Secondly, it can be written as the 'Energy' (an information-theoretic object to be defined below (1.11)) of the generative density averaged over the recognition density minus the (Shannon [39]) entropy of the recognition density. This second property underpins an intuitive parallel with thermodynamics [7], and provides us with a heuristic to understand Free Energy.

As with our transformation from (1.5) to (1.7), we can rearrange (1.8) as follows:

$$\begin{aligned} F &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega, s)} \\ &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega|s)p(s)} \\ &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega|s)} - \ln p(s) \\ F &\geq -\ln p(s) \end{aligned} \tag{1.9}$$

The final step holds because the first term in the third line is the Kullback-Leibler divergence between two probability densities, which is always positive [12]. This means that F is an upper bound on surprisal. It is a tight bound because when the divergence between $q(\omega)$ and $p(\omega|s)$ is zero, i.e. the brain's approximation to the posterior exactly matches the true posterior, then Free Energy will equal surprisal. As it is an upper bound on surprisal, minimising free energy will minimise surprisal.

Under Friston's account, minimising surprisal is something that organisms are bound to do by virtue of their existence [19]. This is because Friston assumes *a priori* that for an organism to remain a functioning entity it must minimise

the entropy of the environments it finds itself in. To do this, it must minimise the entropy of its sensory states, with respect to action. Fishes which find themselves in a (for them) surprising environment, such as on land, rapidly cease to be functioning fishes. Under ergodic assumptions, the long term average of surprisal converges to the entropy (i.e. the ensemble average of surprisal):

$$\begin{aligned} H(S) &= -\int p(s) \ln p(s) ds \\ &= \lim_{T \to \infty} \int_0^T -\ln p(s) dt \end{aligned} \tag{1.10}$$

This means that acting over time to minimise surprisal (or, as is actually the case, some approximating upper bound to surprisal) minimises the entropy of an organism's external states. Thus according to Friston, minimising Free Energy not only allows an organism to more tightly model its external environment, but also fulfils an intrinsic requirement of survival.

The second property of Free Energy, which brings out parallels to thermodynamics, can be shown firstly by defining a second information-theoretic quantity 'Energy', as

$$E(\omega, s) = -\ln p(\omega, s) \tag{1.11}$$

and then rearranging (1.8) and substituting.

$$\begin{aligned} F &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega, s)} \\ &= \int d\omega q(\omega) \ln q(\omega) - \int d\omega q(\omega) \ln p(\omega, s) \\ &= \int d\omega q(\omega) \ln q(\omega) + \int d\omega q(\omega)(-\ln p(\omega, s)) \\ &= \int d\omega q(\omega) \ln q(\omega) + \int d\omega q(\omega) E(\omega, s) \\ &= -ENTROPY + AVERAGE\ ENERGY \end{aligned} \tag{1.12}$$

It is this rearrangement that provided Hinton and colleagues with the motivation to call the quantity we are minimising 'Free Energy', and $-\ln p(\omega, s)$ 'Energy' [14][31]. This is because in physics, the thermodynamic potential Helmholtz Free Energy, A, of a thermodynamically closed system is defined as

$$A = U - TS \tag{1.13}$$

where U is the internal energy of the system, T is its absolute temperature, and S is its (thermodynamic) entropy [7]. The parallel with the final line of (1.12) is clear if we set $T = 1$.

In physics, Helmholtz Free Energy is a measure of the energy of a system available to do 'useful' work. We can think of our information-theoretic parallel in a similar way, if we think of 'useful' in terms of modelling the environment. As will become clear later, minimising free energy under a few straightforward

assumptions equates to minimising the difference between our model predictions and the evidence. This means that when Free Energy reaches a minimum, we have effectively maximised the log model evidence, and the brain's model is as good as it can be. In the ideal case that Free Energy reaches 0, the brain's model would be exact; although the simplifying assumptions we make about the form the brain's model can take renders this impossible in realistic environments.

Some caution should be urged when thinking of Free Energy in this way. This is an intuitively appealing parallel; it is not clear that there is a direct physical interpretation of the information-theoretic quantities so defined. Rather, it is a useful heuristic which helps us to think about what minimising Free Energy is doing, with respect to the brain's efforts at approximate Bayesian inference.

1.4 How the Brain Encodes World States W

An important issue is how the brain encodes the hidden states of the world as part of the recognition density $q(\omega)$ and the generative density $p(\omega, s)$, given that all the brain has access to are sensory data S. We shall consider the issue with respect to the recognition density $q(\omega)$, and derive an expression for F.

Friston's first major assumption is that the recognition density can be factorised into several approximately independent sub-densities[10]. This is because world states are assumed to be factorisable into multiple subsets $\omega_i, i = 1...N$, each corresponding to a distinct time-scale $\tau_1 < \tau_2 < ... < \tau_N$.

$$q(\omega) = \prod_{i=1}^{N} q_i(\omega_i) \tag{1.14}$$

In his complete model, Friston proposes three distinct timescales that, from shortest to longest, represent neuronal activity, synaptic efficacy/plasticity, and synaptic gain respectively.

In the full model, $N > 1$, the distribution of a particular sub-density i can be shown to depend upon its 'partially averaged energy' $\mathcal{E}_i(\omega_i, s)$[18], which is the Energy, $-\ln p(\omega, s)$, with the contributions from all other sub-densities averaged out. This means

$$q_i^*(\omega_i) = \frac{e^{-\mathcal{E}_i(\omega_i, s)}}{\int d\omega_i e^{-\mathcal{E}_i(\omega_i, s)}} \tag{1.15}$$

where the * indicates this is the optimal recognition density at this timescale, selected from an ensemble of recognition densities. This is reminiscent of the Boltzmann distribution in statistical physics [7]. However, in our model of saccadic behaviour, we will be concerning ourselves with the specific case where $N = 1$, i.e. the case of a single timescale, which we will assume corresponds to neuronal activity. This means that for us, partially averaged energy just is energy, and the optimal recognition density is directed to the posterior when the Free Energy is minimised.

Our second major assumption is that the brain models the world as having a Gaussian form:

$$\begin{aligned} q_i(\omega_i) &\equiv G_i(\omega_i; \mu_i, \sigma_i) \\ &= \frac{1}{\sqrt{2\pi\sigma_i}} e^{\frac{-(\omega_i - \mu_i)^2}{2\sigma_i}} \end{aligned} \tag{1.16}$$

where μ_i, the mean, and σ_i, the variance, are the sufficient statistics of the distribution. This is the Laplace approximation, and can be justified by both appeal to the maximum entropy principle [20] - that a Gaussian distribution has the maximum entropy of all form that can be specified by two moments - and the central limit theorem, which states that a statistical system with large degrees of freedom (such as the world) would admit a Gaussian distribution [12].

By modelling the states of the world as Normally distributed, the brain only has to encode the sufficient statistics of those distributions. A further simplifying assumption (see below, p9) permits the brain to discard σ_i, meaning that the recognition density simply becomes a function of the expectations, μ_i, encoded by particular brain states. This allows the formulation of Free Energy in terms of brain states, rather than hidden world states. In this form, the brain can act to minimise Free Energy.

Recall that the free energy can be decomposed into an entropy term and an average energy term (1.12).

$$F = \int d\omega q(\omega) \ln q(\omega) + \int d\omega q(\omega) E(\omega, s) \tag{1.17}$$

We now assume that the Gaussian form of $q(\omega)$ is *sharply peaked*. In other words, that the distribution approaches a delta function. This means that we are assuming that the brain models the world via the recognition density as being *determined*, rather than probabilistic; that the variance of states in the world is minimal, at least with respect to their means [24]. This is not to be confused with the variances of the generative model, which encode how reliable the brain thinks its own expectations are. These variances are not necessarily minimal, and in Friston's model enact the role of attention[28]. More on this later (1.9).

This assumption allows us to perform a Taylor expansion of $E(\omega, s)$ in the second term around $\{\mu_i\} = \{\omega_i\}$:

$$\begin{aligned} \int d\omega q(\omega) E(\omega, s) &\approx \int d\omega q(\omega) \left\{ E(\mu, s) + \sum_i \left[\frac{\partial E}{\partial \omega_i}\right]_\mu \delta\omega_i + \frac{1}{2} \sum_{i,j} \left[\frac{\partial^2 E}{\partial \omega_i \partial \omega_j}\right] \delta\omega_i \delta\omega_j \right\} \\ &= E(\mu, s) + \frac{1}{2} \sum_i \left[\frac{\partial^2 E}{\partial^2 \omega_i}\right] \sigma_i \end{aligned} \tag{1.18}$$

Here the first order terms vanish, and only the diagonal terms of the second order terms are non-zero [10]. Friston introduces a second simplification by assuming that the variances are optimal [24], which means we can vary Free Energy with respect to σ_i. As it turns out, this allows us to write $\partial^2 E / \partial^2 \omega_i$

purely in terms of σ_i^*, which means that F, when (1.18) is substituted into (1.17),

$$F = \int d\omega q(\omega) \ln q(\omega) + E(\mu, s) + \frac{1}{2} \sum_i \left[\frac{\partial^2 E}{\partial^2 \omega_i} \right] \sigma_i \tag{1.19}$$

can be seen to be comprised of two terms which only depend on the variance, σ_i^*, and the energy term, $E(\mu, s)$. This is because the entropy of a Gaussian distribution (the first term) depends only on its variance [12]. Hence, given we have assumed the variances to be small and optimal, with respect to the minimisation we can discard the first and third terms of (1.19).

We now have a model of how Free Energy is encoded by the sufficient statistics, μ, of the Gaussian form assumed for the recognition density. Furthermore, we assume that μ are representations of the hidden world states that cause sensory data, and are reflected in brain states. The Energy term in this encoding is referred to as 'Variational Energy' [10], $\mathcal{E}$, and differs from the Energy E (1.11) by a constant in the case of multiple time scales. In our model, there is only a single time scale, but we will retain the nominal distinction to highlight that the variational energy is a function of μ, rather than of ω:

$$\mathcal{E}(\mu, s) = -\ln p(\mu, s) \tag{1.20}$$

where $p(\mu, s)$ is the Laplace-encoded generative density, i.e. the brain's model of the relationship between sense data s and its own states μ, which represent hidden states in the world ω. In this form, the generative density is now accessible to the brain, and can be used to minimise Free Energy.

1.5 Minimising Free Energy

Friston's working assumption is that for any particular brain state, μ_i, the brain updates μ_i by straightforward gradient descent with respect to the Free Energy surface. This is both computationally tractable and biologically plausible [20].

$$\begin{aligned} \dot{\mu}_i &= -\kappa_i \frac{\partial F}{\partial \mu_i} \\ &= -\kappa_i \frac{\partial \mathcal{E}}{\partial \mu_i} \end{aligned} \tag{1.21}$$

Where κ_i is a learning rate. Here the second line holds because as we have shown (Section 1.4), only the variational energy $\mathcal{E}$ depends upon μ, and so varying F with respect to any μ_i is equivalent to varying $\mathcal{E}$ with respect to that μ_i. Technically, when we consider dynamical models, there is an additional μ' term, but we will discuss this below (p13, (1.39)).

This means that to build a program that minimises free energy, we simply have to specify a generative process GP, which describes how sense data is actually generated by the world, and a generative model GM, which describes the brain's model of how sense data is generated with respect to its expectations.

This latter allows us (under specific assumptions) to specify $\mathcal{E}$ in terms of prediction errors, and thence compute $\dot{\mu}_i$. Integrating the system numerically drives it towards a local Free Energy minimum, and drives the recognition density to a close approximation of the true posterior.

To illustrate the encoding of the generative model GM, let us first consider the simplest case.

$$\begin{aligned} s &= g(\mu) + v \\ \mu &= \bar{\mu} + z \end{aligned} \tag{1.22}$$

Here, some sense data s is modelled by some function of the brain state μ, plus some noise, whilst μ itself varies around an expected value of $\bar{\mu}$. For simplicity, let us assume s and μ are scalar values. We assume that v and z are normally distributed, which allows us to specify the form of the distributions when we run the generative processes around the means $g(\mu)$ and $\bar{\mu}$:

$$\begin{aligned} p(s|\mu) &= \frac{1}{\sqrt{2\pi\Omega_v}} exp\Big(\frac{-(s-g(\mu))^2}{2\Omega_v}\Big) \\ p(\mu) &= \frac{1}{\sqrt{2\pi\Omega_z}} exp\Big(\frac{-(\mu-\bar{\mu})^2}{2\Omega_z}\Big) \end{aligned} \tag{1.23}$$

Recall (1.20) the definition of the variational energy, $\mathcal{E}(\mu, s) = -\ln p(\mu, s)$. As $p(\mu, s)$ can be written as $p(s|\mu)p(\mu)$, (1.23) can be substituted into (1.20). With a simple manipulation this yields, up to a constant,

$$\mathcal{E}(\mu, s) = \frac{1}{2\Omega_v}\varepsilon_v^2 + \frac{1}{2\Omega_z}\varepsilon_z^2 + \frac{1}{2}\ln(\Omega_v\Omega_z) \tag{1.24}$$

where

$$\varepsilon_v = s - g(\mu) \tag{1.25}$$

$$\varepsilon_z = \mu - \bar{\mu} \tag{1.26}$$

Minimising $\mathcal{E}$ with respect to μ, we can see that it is effectively the error terms, ε_v and ε_z, that are driving the minimisation. This makes sense; the errors can be thought of as indicative of how close the system's current model is to the sense data it is attempting to model. Due to how we have defined variational energy we can see that under Gaussian assumptions, the prediction errors are literally the surprisal, i.e. the negative log probability, associated with that particular sensory input under our current generative model.

These errors are weighted by the inverse variances - the precisions - of the model. This means that the minimisation will be primarily driven by those parts of the model with the largest precisions; i.e. those errors in which the system (or the brain) has most confidence. For Friston, this precision weighting can be thought of as attention [22]; it picks out those parts of the sensorium with respect to which the system is trying to minimise Free Energy, either by perception or action, at any point in time.

There are two main improvements to the simple model which will now be discussed. The first of these is extending the model to the dynamic case, i.e. where the brain or system models the evolution of its own states over time. Consider the system,

$$\begin{aligned} s &= g(\mu) + v & \mu &= \bar{\mu} + z \\ \frac{ds}{dt} &= \frac{\partial g}{\partial \mu}\frac{d\mu}{dt} + \dot{v} & \frac{d\mu}{dt} &= f(\mu) + x \end{aligned} \tag{1.27}$$

where the first order time derivatives of s and μ are also considered. In general, Friston considers all temporal orders [17], so that s and μ are replaced by $\tilde{s}$ and $\tilde{\mu}$ where

$$\begin{aligned} \tilde{s} &= \left(s, \frac{ds}{dt}, \frac{d^2 s}{dt^2}, ..., \frac{d^n s}{dt^n}\right) \\ &= \left(s_{[0]}, s_{[1]}, s_{[2]}, ..., s_{[n]}\right) \end{aligned} \tag{1.28}$$

and

$$\begin{aligned} \tilde{\mu} &= \left(\mu, \frac{d\mu}{dt}, \frac{d^2 \mu}{dt^2}, ..., \frac{d^n \mu}{dt^n}\right) \\ &= \left(\mu_{[0]}, \mu_{[1]}, \mu_{[2]}, ..., \mu_{[n]}\right) \end{aligned} \tag{1.29}$$

where $n = \infty$. In this case, the system outlined in (1.27) would have infinitely many equations, two at each dynamical order, which would provide a generalized map in the case of $\tilde{s}$, and generalized equations of motion in the case of $\dot{\tilde{\mu}}$, where $\mu_{[0]}$ is ignored because we assume that the precision of z is very low; we assume that the system has no expectations about its own state at any particular point in time, only about the way that state will evolve over time. We also assume that only linear derivative terms are collected at any dynamical level; Friston refers to this as the 'local linearity assumption' [25].

For our purposes, however, the system in (1.27) is adequate - the infinite dimensional model with full generalized co-ordinates $\tilde{s}$ and $\tilde{\mu}$ would be impossible to code without some truncation of higher order dynamics. We take $\tilde{s} = \{\dot{s}, s\}$ and $\tilde{\mu} = \{\dot{\mu}, \mu\}$, and as with simple system (1.22) rewrite the generative distribution of which the variational energy is a function as follows:

$$\begin{aligned} p(\tilde{s}, \tilde{\mu}) &= p(\tilde{s}|\tilde{\mu})p(\tilde{\mu}) \\ &= p(\{\dot{s}, s\}|\{\dot{\mu}, \mu\})p(\{\dot{\mu}, \mu\}) \\ &= p(\dot{s}|\dot{\mu})p(s|\mu)p(\{\dot{\mu}, \mu\}) \\ &= p(\dot{s}|\dot{\mu})p(s|\mu)p(\dot{\mu}|\mu)p(\mu) \end{aligned} \tag{1.30}$$

The third line holds because we are assuming that dynamical levels are themselves independent; this is a necessary consequence of the assumption of local linearity. This means that we take it to be the case that the sensory data at a particular dynamical order n only interacts with, and is represented by, brain states of the same dynamical order, $\mu_{[n]}$. Finally, as mentioned, we ignore $\mu_{[0]}$ and thus discard $p(\mu)$. This gives us

$$\mathcal{E}(\tilde{\mu}, \tilde{s}) = -\ln\left[p(\dot{s}|\dot{\mu})p(s|\mu)p(\dot{\mu}|\mu)\right] \tag{1.31}$$

As with the simple system, we assume the noise terms $\dot{v}, v, z$ are normally distributed, and run the generative processes in (1.27) around their respective means to derive expressions for the contributing probabilities in terms of the errors and their variances.

$$\begin{aligned} p(\dot{s}|\dot{\mu}) &= \frac{1}{\sqrt{2\pi\Omega_{\dot{v}}}} exp\Big(\frac{-(\dot{s}-\dot{g}(\dot{\mu}))^2}{2\Omega_{\dot{v}}}\Big) \\ p(s|\mu) &= \frac{1}{\sqrt{2\pi\Omega_{v}}} exp\Big(\frac{-(s-g(\mu))^2}{2\Omega_{v}}\Big) \\ p(\dot{\mu}|\mu) &= \frac{1}{\sqrt{2\pi\Omega_{x}}} exp\Big(\frac{-(\dot{\mu}-f(\mu))^2}{2\Omega_{x}}\Big) \end{aligned} \tag{1.32}$$

Here,

$$\dot{g}(\dot{\mu}) \equiv \frac{\partial g}{\partial \mu}\frac{d\mu}{dt} \tag{1.33}$$

and for higher dynamical orders n (with which we are not concerning ourselves),

$$g_{[n]} \equiv \frac{\partial g}{\partial \mu}\frac{d^n\mu}{dt^n} \tag{1.34}$$

Substituting into (1.31) results in the following expression for the variational energy,

$$\mathcal{E}(\mu, s) = \frac{1}{2\Omega_{\dot{v}}}\varepsilon_{\dot{v}}^2 + \frac{1}{2\Omega_{v}}\varepsilon_{v}^2 + \frac{1}{2\Omega_{x}}\varepsilon_{x}^2 + \frac{1}{2}\ln(\Omega_{\dot{v}}\Omega_{v}\Omega_{x}) \tag{1.35}$$

where

$$\varepsilon_{\dot{v}} = \dot{s} - \dot{g}(\dot{\mu}) \tag{1.36}$$

$$\varepsilon_{v} = s - g(\mu) \tag{1.37}$$

$$\varepsilon_{x} = \dot{\mu} - f(\mu) \tag{1.38}$$

which can then be minimised with respect to μ and $\dot{\mu}$. This removes the final term of (1.35), again yielding an expression in terms of the errors weighted by their respective precisions.

There is one final nuance in the dynamical minimisation. Note that the straightforward minimisation by gradient descent described in (1.21) is not guaranteed to reach a minimum, as the $\dot{\mu}_{[n]}$ term ($\mu_{[n+1]}$) is already part of the generalised co-ordinates defined by $\tilde{\mu}$, as is $\mu_{[n]}$. To get around this, we introduce a second term, $\mu'_{[n]}$, which can be thought of as the brain's representation of the next dynamical order up[24]. It is a time derivative not necessarily equal to $\dot{\mu}_{[n]}$ Thus,

$$\dot{\mu}_{[n]} = \mu'_{[n]} - \kappa_{[n]}\frac{\partial F}{\partial \mu_{[n]}} \tag{1.39}$$

is guaranteed to reach a local Free Energy minimum. If we think about it, this makes sense; $\partial F/\partial\mu_{[n]} = 0$ when $\dot{\mu}_{[n]} = \mu'_{[n]}$. This states that when the

brain's model of the evolution of the state of some dynamical level is equal to the actual evolution of the state of that dynamical level, then the Free Energy is minimised, with respect to that dynamical level. As we have discussed (Sec (1.3)), Free Energy is a proxy to our measure of the difference between our brain's model of the hidden states of the environment which cause sense data and the true posterior (i.e. true distribution) of those hidden states. Thus, if our model is exact, with respect to some dynamical level, Free Energy should necessarily be minimised with respect to that level, which the $\mu'_{[n]}$ term enforces.

The second main improvement to the simplistic model is to make it hierarchical. As will become clear in Chapter 2, our model of saccades approximates the action of minimising Free Energy across this hierarchy, rather than being explicitly hierarchical - we concern ourselves only with the lowest level of the hierarchy. However, we will provide a brief overview of the ideas behind the hierarchical model, as this is important to an understanding of the full dynamical hierarchical GM, and its claims to model the brain.

The mathematical motivation behind the hierarchy is to reduce the contribution of the prior $p(\mu)$ to the generative distribution $p(s, \mu)$ to a minimum, which allows us to consider it simply as a noise term, and of minimal importance. Accounting for the priors of the generative model is one of the main practical problems faced by Variational Bayes approaches to approximate Bayesian inference (of which Free Energy minimisation is an instance).

Let us divide the brain state μ into multiple sub-states $\{\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, ..., \mu^{(M)}\}$, so that we can rewrite the generative distribution as

$$\begin{aligned} p(s, \mu) &\equiv p(s, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}, ..., \mu^{(M)}) \\ &= p(s|\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, ..., \mu^{(M)}) p(\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, ..., \mu^{(M)}) \\ &= p(s|\mu^{(1)}) p(\mu^{(1)}|\mu^{(2)}) \ \ ... \ \ p(\mu^{(M-1)}|\mu^{(M)}) p(\mu^{(M)}) \end{aligned} \tag{1.40}$$

The final line assumes that the sub-states are conditionally independent of one another. We can usefully think of this as representing a hierarchical system where each $\mu^{(i)}$ corresponds to the state of the ith level, and we have made a Markov[16] assumption so that the contributions to the probability of any level can be defined just in terms of the level above. This means the posterior probability of any particular level acts as the prior probability of the level below. To make this explicit, we have written the hierarchical system below:

$$\begin{aligned} s &= g^{(1)}(\mu^{(1)}) + v^{(0)} \\ \mu^{(1)} &= g^{(2)}(\mu^{(2)}) + v^{(1)} \\ &. \\ &. \\ \mu^{(M-1)} &= g^{(M)}(\mu^{(M)}) + v^{(M-1)} \\ \mu^{(M)} &= v^{(M)} \end{aligned} \tag{1.41}$$

Here we can see that the lowest level models the sense data (which can be thought of as $\mu^{(0)}$) directly, and each level above attempts to model the sufficient

statistics of the level below. At the top level $\mu^{(M)}$, now a minor contributor as the final prior of the system, is simply noise. As ever, we assume each noise term $v^{(i)}$ is Gaussian in form, and this means that we can write the generative density, and thus the variational energy, in terms of the errors between the expectations of the $(i+1)^{th}$ level, $g^{(i+1)}(\mu^{(i+1)})$, and the actual state of the ith level, $\mu^{(i)}$. In practice this means that errors rise up the hierarchy, and priors flow down.

To combine the dynamical and hierarchical models into a full generative model, Friston assumes that the hidden states of the world, W, can be divided into causal and dynamic states, $W = V \times X$ [20]. These are modelled by brain states μ_v and μ_x respectively. This separation allows us to think of causal states, μ_v, as those which contribute to the inter-layer dynamics. These are the states that the functions $g^{(i)}$ attempt to model, such that

$$\tilde{\mu}_v^{(i)} = \tilde{g}^{(i+1)}(\tilde{\mu}_v^{(i+1)}, \tilde{\mu}_x^{(i+1)}) + v^{(i)} \tag{1.42}$$

where the tildes indicate that there is an equivalent function at every dynamical level, which we have assumed to be independent (see (1.30)). Conversely, the dynamic states, μ_x, are those which contribute to intra-layer dynamics, i.e. the brain's attempt, by $f^{(i)}$, to model the generalized motion of its own dynamics at that hierarchical layer.

$$\dot{\tilde{\mu}}_x^{(i)} = \tilde{f}^{(i)}(\tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)}) + x^{(i)} \tag{1.43}$$

By assuming the $v^{(i)}$ and $x^{(i)}$ noise terms are normally distributed, this allows for a complete determination of the variational energy, which is then minimised with respect to each brain state by gradient descent. We can think of the variational energy as a surface above all brain states ($\mu_i \in \mathbb{R}$), where the brain attempts to find a local minimum by updating brain states accordingly. This corresponds to perception; action requires the provision of a forward model, which we discuss briefly below.

1.6 Action as a Result of Minimisation

Thus far, we have described a hierarchical system in which sensory data is an uncontrollable input to the bottom layer. By changing brain states to minimise free energy and thus select the brain state which corresponds to a model of a possible world state for which the log sensory evidence is then maximised, the system perceives its environment. However, this does not by itself guarantee the minimisation of the surprisal (and thus by ergodic assumptions, the entropy). All that the organism can do by perception is tighten the Free Energy bound on the surprisal associated with a particular sensory state. To truly minimise the entropy of its sensory input, an organism must interact with its environment to change that sensory data directly (in the simplest case, it should be able to move). For Friston, this action can also be derived from the Free Energy minimisation [22].

To make sense of this claim in the context of our discussion, recall (1.9) our demonstration that the Free Energy, F, the divergence between the recognition distribution $q(\omega)$ and the generative distribution $p(\omega, s)$, can be written as the divergence between $q(\omega)$ and the true posterior $p(\omega|s)$ plus the surprisal $-\ln p(s)$. When we are perceiving, it is this divergence between $q(\omega)$ and $p(\omega|s)$ that is being minimised.

However, we can also rearrange F as follows:

$$\begin{aligned} F &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega, s)} \\ &= \int d\omega q(\omega) \ln \frac{q(\omega)}{p(\omega)} - \int d\omega q(\omega) \ln p(s|\omega) \end{aligned} \tag{1.44}$$

Friston characterises this as 'Complexity minus Accuracy' [23], but it is the 'Accuracy' term with which we are most concerned. This can be thought of as the conditional surprisal averaged across the recognition distribution $q(\omega)$, or equivalently, as the expected surprisal under some recognition distribution. Under Gaussian assumptions, as we have shown, this becomes simply the prediction error. If the agent can act to affect sensory data, then assuming it strives to minimise Free Energy, it will act to drive sensory states closer to its conditional expectations.

According to Friston, this minimisation occurs close to the boundary between the body and the world; when the priors which are flowing down the system take the form of expectations about *proprioceptive* sensations[23]. At this level, minimising Free Energy involves minimising proprioceptive prediction errors; whilst this can be done by altering the predictions, it can also be achieved by altering the proprioceptive sensations themselves. Friston hypothesises that this is effectively achieved by equipping the generalized predictive coding scheme with classic reflex arcs[22].

At first glance, this 'Active Inference' seems highly counter intuitive. Under this hypothesis there are no top-down motor commands. Rather, there are simply expectations about proprioceptive input which the body then acts to satisfy - we expect our hand to grasp a cup, and so it does. The motor commands are the proprioceptive errors; action occurs to minimise these errors. However, some thought will convince us that we have already implicitly accepted this if we have accepted the Free Energy account of perception; in the complementary problem of perceiving, the upflow of sensory data through our brain has no direct bearing our experience. All that flows up through the brain are the instructive error signals, which are decomposed across the hierarchy. What we *perceive* are those top down representations which best explain away our sensory input.

We take sense data to be a function of action, and thus assume we can minimise F with respect to action by gradient descent.

$$\begin{aligned} \dot{a} &= -\kappa_a \frac{\partial F}{\partial a} \\ &= -\kappa_a \frac{\partial F}{\partial s} \frac{\partial s}{\partial a} \end{aligned} \tag{1.45}$$

Where the second line holds by the chain rule because sensory states are a function of action, and $\partial F/\partial s$ can be thought of as the error at the lowest level of the hierarchy:

$$\varepsilon_v^{(1)} = (\mu^{(0)} - g^{(1)}(\mu^{(1)})) \tag{1.46}$$

The full FEP model, therefore, sees the brain-body system performing a dual minimisation of Free Energy with respect to action and perception:

$$\begin{aligned} \mu(t)^* &= \underset{\mu}{argmin}\ F(s(t), \mu(t)) \\ a(t)^* &= \underset{a}{argmin}\ F(s(t), \mu(t)) \end{aligned} \tag{1.47}$$

where * indicates the optimal brain state, in the case of μ, and optimal action, in the case of a.

This is compelling. By proposing an overall directive - that an organism minimise Free Energy - we have shown that two of the most fundamental contributors to intelligence both arise from the same process. Action and perception appear dissimilar because the process of minimising Free Energy is being embodied in different physical contexts. The structure of the body and the central nervous system mean that proprioceptive prediction errors can be minimised through an active forward model. Deeper within the brain, prediction errors are minimised by altering brain states, which corresponds to selecting those prior expectations which best explain away the errors rising through the hierarchy.

1.7 Decomposing Error Signals

It is perhaps helpful to step back at this point, and consider the Free Energy Principle from a more abstract standpoint.

Firstly, it should be noted that at any hierarchical level i the levels above are all combining to perform inference on the level $(i-1)$; as far as they are concerned, the $(i-1)^{th}$ level could be the sensory input from the world. This is a result of the Markov assumptions we have made in creating the hierarchy - in practice it means that we can think of each set of upper levels $\{i...M\}$ as embodied within a 'world' only accessible by the statistics of the state of the $(i-1)^{th}$ level. When $(i-1) = 0$, then this 'world' is the real world: the environment external to the system.

We can think of the hierarchy as *decomposing* the sensory input across its levels; in a way analogous to Fourier decomposition's division of a signal into the contributions from its various frequencies [9]. As we are minimising the negative log of a product of probability distributions we have assumed to be Gaussian, this decomposition is also subtractive. Instead of frequencies, the hierarchy decomposes the incoming signal into a set of statistical structures in the log domain, each of which corresponds to the predictions of a particular level, $g^{(i)}$. These structures can further possess their own dynamics, which the brain models by $f^{(i)}$. Each level subtracts away that part of the error signal

which it correctly models, and the remaining error is passed to the higher levels. We could say the brain decomposes a signal into a 'concept' domain.

This us a better insight into what it means for us to treat the top prior of the model, $p(\mu^{(M)})$, as Gaussian noise. By not modelling the error remaining at the top of the hierarchy (see (1.41)), we are effectively ignoring any structures remaining in the signal after it has passed through the brain. This remnant can be thought of as the **unexplained surprisal** in sensory states.

In reality, the brain is always going to be unable to fully model the sensorium; the world is far more complex than our brain's capacity to model it. If the brain does indeed enact some form of Variational Bayes, the maximum possible complexity of its model of world states can be seen to be a function of the depth of the hierarchy, and the structure of the generative functions f and g at each level. We could think of this in evolutionary terms; if an organism's model cannot capture the full complexity of the incoming sense data, then it must at least capture those structures in the sense data that correspond to states of the world that will directly affect the organism's survival. Dismissing the remaining error as Gaussian noise is not to say, therefore, that this remaining sensory input lacks structure, or fails to describe certain parts of the world. Rather, the organism is assuming that this input is not relevant with respect to the current demands of survival, and so can be treated as noise. Whilst surprising, it is not dangerously so.

This final noise term also acts as a buffer against over-fitting. The constraints of a finite hierarchy with respect to the complexity of possible models enact a real-world structural example of the classic bias-variance trade-off (see, e.g. [36] p705). By not perfectly accounting for sensory input, the brain avoids a high variance regime which would minimise surprisal in the moment, but might well fail to predict a future event which could be catastrophically surprising. Evolving some level of complexity allows the brain to avoid a high bias regime, in which it is incapable of building models complex enough to be useful. In short, we can see the structure of the brain as a sophisticated balancing act evolved under the straightforward constraint of minimising average surprisal over time, which under ergodic assumptions is equivalent to minimising the entropy of sensory data (1.10). We are effectively adding a regularization term to our model. Our first expectation is about the model itself; the topmost layer predicts that the layers below will explain away just those parts of the world which are necessary to survive in the present, and that what remains can be treated as Gaussian noise.

1.8 The Dark Room Problem

As we have discussed, for any particular level i, the $(i-1)^{(th)}$ level could be the sensory input from the world. We can turn this on its head; as far as the $i = 1$ level is concerned, the actual sensory input s could just be the error signal from lower levels of a larger hierarchy. In other words, we can think of sensory input itself as an *error* signal. Rather than being an instructive input, sensory data is

then the error created by our interactions with the world. For Friston, this leads to the conclusion that we tend to act to *confirm*, rather than *disconfirm*, our expectations of the world. For Friston's opponents, this is more problematic. It highlights the fact that minimal surprisal - i.e. the minimisation of Free Energy - could be most readily achieved by sitting in a room with the lights off and staring at the wall [35].

This 'dark room' problem is not the focus of this dissertation. For the interested reader we recommend Andy Clark's review paper, and the commentaries as a good place to start [13]. Here, we will briefly highlight what can be seen as the main rebuttal; that the brain has evolved to minimise long term surprisal, to maximise its survivability. This means that the nature of the world has trained us to expect things like light, and movement. This is because the reduction of surprisal gained by switching the lights off, or staying in dark places, has been outweighed in the long term by the enormous (and quite possibly terminal) increase in surprisal when a predator sneaks up on us [20].

In short, we minimise long-term surprisal with respect to our embodiment in a particular environment. We are able to do so accurately because we have been structured to do so by the pressures of natural selection. Even if we chose to sit in a darkened room for a prolonged period of time, the sub-environment of our body would eventually begin to make interoceptive demands on the higher levels of the nervous system (we might begin feeling hungry), which would percolate upwards as errors to be minimised by priors that would flow down the hierarchy and manifest as action at the lowest levels. We would get up and find food. This might require us to turn the light on.

The idea that we act to confirm our expectations follows, for Friston, directly from the idea that we select the action which most minimises Free Energy [20]. In the dual minimisation discussed at (1.47), by minimising Free Energy we act to minimise the surprisal between our model and the sensory input. Assuming that the system has some forward model of how actions bring about sensory input, and some current state of expectations about the world, then we act to bring our sensory experience in line with those expectations, rather than acting to disambiguate or check for evidence against a current hypothesis about the world. However, as we will explore, modelling expected future (i.e. counterfactual) surprisal can lead to disambiguatory actions.

1.9 Attention and Consciousness

Finally, thinking of perception as inference allows us to sketch a tentative hypothesis concerning what exactly comprises conscious experience. Under a predictive coding scheme, the inference on the causes of sensory data at any particular point in time is essentially the set of current brain states $\{\mu^{(1)}...\mu^{(M)}\}$ that are best able to explain the sensory data s. We can think of this total brain state, or some part of it, as giving rise to the conscious experience associated with a particular percept or gestalt. What it is like for me to *see* a red ball is what it is for me to be in the brain state that corresponds best to that par-

ticular inference ('Nagel's oft worn phrase': a mental state is conscious if there is *something it is like* for me to possess it [32]). If the 'Red Ball' prior results in the best minimisation of Free Energy throughout the hierarchy, given some sense data, then my brain will alter its state towards that which corresponds to that particular prior, and I will have a conscious visual experience of a red ball.

The FEP also provides a neat response to the Binding Problem [4]; why it is that the redness and roundness of the ball are phenomenally unified, even though they are processed in different parts of the brain, and at different timescales [40]. If we think of percepts as the result of our priors, then binding happens before sensation. We assume that there is a red ball in the world, and by virtue of doing so high-level expectations provide us with the unified phenomenological experience of a red ball. These priors then flow down the system, in the form of predictions about expected sensory input. If these predictions fail to be satisfied, the upflow of errors may modify the higher priors, and the character of our experience will change with our brain state. Rather than having to provide an account of some unifying mechanism that knits disparate sensory information together as it spreads through the brain, we have instead an account whereby objects are unified by virtue of us seeing them *as* objects.

The Free Energy framework also claims to explain attention; that sense that there are certain features of our experience at any particular moment of which we are more fully conscious than the rest. For example, when we attend to a particular aspect of a visual scene, such as a single tree in a forest, the tree dominates our conscious experience (common consensus, e.g. [15], 29/30, is that we are still conscious of the rest of the forest).

For Friston, attention can be thought of as a result of the precision weighting of error signals [20]. Weighting the error by our confidence in a particular part of the generative model drives the minimisation more with respect to that particular error. This has two effects. Firstly, we will be driven to act to minimise the largest error. This supports the view of attention as selection for action - those features in the world to which we are attending are those we are most likely to interact with. Under the FEP account, this is because attention is just precision weighting of the errors associated with the brain states which model those features. Secondly, heavily weighting an error from a particular level in the hierarchy means that higher levels in the hierarchy will select their states, $\mu^{(i)}$, to best explain away that error. This means that a greater proportion of total brain states $\{\mu^{(1)}...\mu^{(M)}\}$ will be representing that particular part of the world in which we have high confidence. If our phenomenal content can be described in terms of our particular brain state, then it follows that if a large proportion of our brain state at time t is representing a particular part of the world, our perception of that part of world will correspondingly dominate our conscious experience at time t. The FEP therefore can claim to provide a grounded account of both the role attention plays in conscious experience, and the role attention plays in the selection of features for subsequent action.

Chapter 2

System: Implementation

2.1 Introduction

As our starting point, we will be using the model outlined in Karl Friston's 'Perceptions as Hypotheses' [22]. Our model deviates from Friston's in that we streamline much of the gradient descent process, and our saccades are modelled as instantaneous. It is unclear whether they are or not in the original paper.

In addition, Friston simply evaluates the salience at the solutions to the generative model, such that $\mu_{x,q}(t+\tau) = \mu_{x,q}(t)$. Our system minimises Free Energy counterfactually to better simulate that disambiguating priors relies on future expectations about not just the behaviour of those priors, but also of the model itself. By computing the expected value of $\mu_{x,q}(t+\tau)$ separately for each fictive point of foveation (see Section 2.4), we assume that the hierarchy that we are approximating has expectations about its future states.

An additional distinction is that the original paper gives no indication as to how errors between 256 element arrays are computed (each visual input comprises 256 input channels). We took a mean square error approach which, as we will discuss, raised several interesting issues with respect to preserving certain structural aspects of sensory input.

As discussed in the introduction, we are intersted both in disambiguation and empirical viability.. The first is investigated by equipping the system with two priors, in this case a rectangle and a triangle, and then presenting an image which would be expected to provoke a preference between the two (see Section 3.2/3.3). The image presented was Kanizsa's Triangle, as in humans this provokes a strong visual experience of a triangle. The second scenario presented an ambiguous image to the system, to create a bistable percept (see Section 3.4/3.5).

2.2 The System

To build a program that minimises Free Energy we need to specify a generative process GP, which describes how sense data is actually generated by the world, and a generative model GM, which describes the brain's model of how sense data is generated with respect to its expectations.
Let us define our generative process as follows:

$$\begin{aligned} s_p &= x_p + w_p \\ s_q &= g(I, x_p) + w_q \\ \dot{x}_p &= a - Cx_p + w_{\dot{p}} \end{aligned} \tag{2.1}$$

Here w_p,w_q and $w_{\dot{p}}$ represent noise terms (not to be confused with ω, which in Chapter 1 represented world states). x_p ($\in \mathbb{R}^2$) is the point of foveation in 2-D co-ordinates (for consistency with Matlab indexing we set [0,0] as the top left corner of the image). s_p ($\in \mathbb{R}^2$) is the system's proprioceptive input: the system's sense of where it is currently foveating. s_q ($\in \mathbb{R}^{256}$) is the visual or perceptual input to the system, computed as a function of the image I and the true point of foveation.

$g(I, x_p)$ returns a 16x16 array of real number values in the range [0,1]. Each value corresponds to the output of a visual channel which outputs the maximum of a DoG (difference of Gaussian) filter across 1/256th of 1/36th of the image. This means the visual field itself covers 1/36th of the whole image. Performing a convolution with a DoG filter is equivalent to running a band-pass filter over the image, which permits only a certain range of spatial frequencies through.

$$DoG(\sigma_e, \sigma_i) = \Big[\frac{1}{\sqrt{2\pi\sigma_e}}\Big] e^{(-\frac{x^2}{2\sigma_e^2})} - \Big[\frac{1}{\sqrt{2\pi\sigma_i}}\Big] e^{(-\frac{x^2}{2\sigma_i^2})} \tag{2.2}$$

The net result is to edge-detect, by virtue of detecting sharp local second order changes in intensity, a task performed by ganglion cells in the retina in humans, and analogous to error detection with respect to an expectation of local smoothness [30]. With Friston, we have taken the ratio $(\sigma_i/\sigma_e) = 4$. A more common ratio is ~ 1.6 [33], but no difference was detected in our preliminary results, and for ease of debugging, a whole number ratio of four pixels to one pixel was retained. This preprocessing is simply intended to select only those parts of the image most likely to be informationally dense.

a ($\in \mathbb{R}^2$) is the action term; it denotes a vector of distance per time step, so that when the system is integrated, the new point of foveation can be computed by adding the vector a to the vector x_p (plus some noise). C is a small constant, and governs the decay of the point of foveation back to the centre of the image if $a = 0$.

This division of sensory input into proprioceptive and visual signals results in a generative model which can be thought of as two distinct processes, linked only by the fact that the perceptual input is a function of the current point of

foveation. We define the GM as follows:

$$
\begin{aligned}
&\textbf{Proprioception:} \\
&\qquad s_p = \mu_p + w_{vp} \\
&\qquad \mu'_p = \frac{1}{4}(\mu_u - \mu_p) + w_{xp} \\
&\textbf{Perception:} \\
&\qquad s_q = \sum_j exp(\mu_{qj}) g(I_j, \mu_p) + w_{vq} \\
&\qquad \mu'_q = 1 - \sum_j exp(\mu_{qj}) + w_{xq}
\end{aligned}
\tag{2.3}
$$

Here, the top two equations are the system's model of where it is looking, in terms of input $s_p(\in \mathbb{R}^2)$ and proprioceptive state $\mu_p(\in \mathbb{R}^2)$. w_{xp} and w_{vp} are noise terms, where the v subscript denotes the noise on a model of a causal state, and x the noise on the model of a dynamic state (see sect). $\mu_u(\in \mathbb{R}^2)$ is the brain's representation of some hidden control state $\mathbf{u}$; this involves the further subdivision of world states W such that $W = V \times X \times U$. μ_u can be thought of as the brain's expectation about those hidden states of the world U ($\mathbf{u} \in U$) which can be affected by action to change sensory input [22]. For our purposes, it is sufficient to note that μ_u acts as an attractor of the proprioceptive system; the time derivative of μ_p is zero when $\mu_u = \mu_p$.

The bottom two equations are the system's model of what it is seeing. w_{xq} and w_{vq} are noise terms. The sensory input $s_q(\in \mathbb{R}^{256})$ is modelled by the system as a weighted sum of prior expectations. These prior expectations are computed by providing the system with a series of images which constitute its world knowledge; priors on what it thinks the world might consist of. The system then calculates for each prior what the expected input, $g(I_j, \mu_p)$, would be assuming the same point of foveation μ_p for both prior image I_j and world image I.

The weights, $exp(\mu_{qj})$, can be thought of as the system's prior estimate of the probability that the jth expected image is actually the image being observed; note that the dynamics of the system are modelled such that $\mu'_q = 0$ when $\sum exp(\mu_{qj}) = 1$, where μ'_q is the time derivative of μ_q according to the model, which is distinct from the time derivative of μ_q derived from the gradient descent (see Chapter 1). This means that the second part of the perceptual system enacts a dynamic renormalisation that enforces that the sum of the weights is approximately equal to 1. This is effectively a softmax probability [6], as the approximate probability of any particular prior j can be computed as

$$
p(j) = \frac{exp(\mu_j)}{\sum_j exp(\mu_j)} \tag{2.4}
$$

Note the structural parallels with the distribution of a particular sub-density of the recognition density $q(\omega)$ under the assumption that it can be factorized by

timescale, as discussed in (1.15):

$$q_i^*(\omega_i) = \frac{e^{-\mathcal{E}_i(\omega_i,s)}}{\int d\omega_i e^{-\mathcal{E}_i(\omega_i,s)}}$$

Our priors are hierarchically flat; we are approximating the full activity of a deep learning hierarchy with several images I_j and an associated state μ_{qj}. From the parallel above we can postulate that the μ_{qj} term can be thought of as representing the negative partially averaged energy of the full system state that would, in the full hierarchy, be associated with that particular percept. We are not here factorising by timescale, but applying similar independence assumptions to particular sets of possible world states. This means we can treat each μ_{qj} variable as orthogonal to the rest.

It is not clear if this is valid; whilst it makes the mathematics tractable, it removes possible interdependencies which might contribute to disambiguation; for example, the prior expectation that one object cannot occupy the same space as another will be lacking from a model which assumes their statistical independence. To compensate, our counterfactual computations assume that the world is actually the current most likely prior; we enforce an expectation that the world can only be in one state.

2.3 Minimising the Free Energy of the System

As we have shown in Chapter 1, to minimise Free Energy, we minimise the variational energy, $\mathcal{E}$, which is the negative log of the joint distribution of system or brain states μ and sensory states s:

$$\mathcal{E}(\mu, s) = -\ln p(\mu, s) \tag{2.5}$$

The simplest way to approach this is to think of $\mathcal{E}$ as a surface above μ and s. To find a numerical approximation to a minimum we compute the gradient of the surface $\nabla\mathcal{E}$ at some initial point, and then take a small step in the direction of the negative gradient. We then recompute the gradient at this new point, and repeat. Assuming we choose our step size appropriately, and our function is smooth, this will bring us close to a local minimum (for an excellent overview of gradient descent, see [5], 263). For ease of computation, we will treat each variable independently, and compute the gradient with respect to that variable. Friston argues that the brain implements some variant of this gradient descent approach [20].

Firstly, we will rewrite the generative distribution $p(\mu, s)$ in a more accessible way. We assume that the two sensory states, s_q and s_p are independent, and we will define $\bar{\mu}_p = \{\mu'_p, \mu_p\}$ and $\bar{\mu}_q = \{\mu'_q, \mu_q\}$. This means, by Bayes' Rule, we can write

$$\begin{aligned} p(\bar{\mu}_{p/q}) &= p(\mu'_{p/q}, \mu_{p/q}) \\ &= p(\mu'_{p/q}|\mu_{p/q})p(\mu_{p/q}) \end{aligned} \tag{2.6}$$

Rewriting $p(\mu, s)$, we get

$$\begin{aligned} p(\mu, s) &= p(s, \mu) \\ &= p(s_p, s_q, \bar{\mu}_p, \bar{\mu}_q) \end{aligned}$$

Which, under Bayes' Rule

$$= p(s_p, s_q|\bar{\mu}_p, \bar{\mu}_q)p(\bar{\mu}_p, \bar{\mu}_q)$$

As we have assumed the independence of sensory states,

$$= p(s_p|\bar{\mu}_p, \bar{\mu}_q)p(s_q|\bar{\mu}_p, \bar{\mu}_q)p(\bar{\mu}_p, \bar{\mu}_q)$$

We then note from the GM that s_p is conditionally independent of $\bar{\mu}_q$

$$= p(s_p|\bar{\mu}_p)p(s_q|\bar{\mu}_p, \bar{\mu}_q)p(\bar{\mu}_p, \bar{\mu}_q)$$

Finally, we assume $\bar{\mu}_p$ and $\bar{\mu}_q$ are independent, and expand them as described in (3.7):

$$\begin{aligned} &= p(s_p|\bar{\mu}_p)p(s_q|\bar{\mu}_p, \bar{\mu}_q)p(\bar{\mu}_p)p(\bar{\mu}_q) \\ &= p(s_p|\mu'_p, \mu_p)p(s_q|\mu'_p, \mu_p, \mu'_q, \mu_q)p(\mu'_p|\mu_p)p(\mu_p)p(\mu'_q|\mu_q)p(\mu_q) \\ &= p(s_p|\mu_p)p(s_q|\mu_p, \mu_q)p(\mu'_p|\mu_p)p(\mu_p)p(\mu'_q|\mu_q)p(\mu_q) \\ &= p(s_p|\mu_p)p(s_q|\mu_p, \mu_q)p(\mu'_p|\mu_p)p(\mu'_q|\mu_q) \end{aligned} \tag{2.7}$$

Here we have noted from the generative model that s_p and s_q are conditionally independent of states of a higher dynamical order (μ'_p,μ'_q), and we have discarded the priors for the reasons discussed in the simple generative model examined in Chapter 1; that the system does not have prior expectations about its state at any particular point in time, only about how that state might evolve.

Let us now assume that all four noise terms, w_{xp}, w_{vp}, w_{xq} and w_{vq}, are Normally distributed, and run the four processes described by the generative model around their respective means. This yields the following expressions for the various conditional probabilities:

$$\begin{aligned} p(s_p|\mu_p) &= \frac{1}{\sqrt{2\pi\Omega_{vp}}} exp\Big(\frac{-(s_p - \mu_p)^2}{2\Omega_{vp}}\Big) \\ p(s_q|\mu_p, \mu_q) &= \frac{1}{\sqrt{2\pi\Omega_{vq}}} exp\left(\frac{-(s_q - \sum_j exp(\mu_{qj})g(I_j, \mu_p))^2}{2\Omega_{vq}}\right) \\ p(\mu'_p|\mu_p) &= \frac{1}{\sqrt{2\pi\Omega_{xp}}} exp\Big(\frac{-(\mu'_p - \frac{1}{4}(\mu_u - \mu_p))^2}{2\Omega_{xp}}\Big) \\ p(\mu'_q|\mu_q) &= \frac{1}{\sqrt{2\pi\Omega_{xq}}} exp\Big(\frac{-(\mu'_q - (1 - \sum_j exp(\mu_{qj})))^2}{2\Omega_{xq}}\Big) \end{aligned} \tag{2.8}$$

Substituting (2.9) into (2.8) and subsequently (2.8) into (2.6) yields the following expression (up to a constant) for the variational energy, where we have discarded the $1/\sqrt{2\pi\Omega}$ terms of the Gaussians as they do not affect the minimisation with respect to μ and s:

$$\mathcal{E}(\mu, s) = \frac{1}{2\Omega_{vp}}\varepsilon_{vp}^2 + \frac{1}{2\Omega_{vq}}\varepsilon_{vq}^2 + \frac{1}{2\Omega_{xp}}\varepsilon_{xp}^2 + \frac{1}{2\Omega_{xq}}\varepsilon_{xq}^2 \tag{2.9}$$

where

$$\begin{aligned}
\varepsilon_{vp} &= (s_p - \mu_p) \\
\varepsilon_{vq} &= (s_q - \sum_j exp(\mu_{qj}) g(I_j, \mu_p)) \\
\varepsilon_{xp} &= (\mu'_p - \frac{1}{4}(\mu_u - \mu_p)) \\
\varepsilon_{xq} &= (\mu'_q - (1 - \sum_j exp(\mu_{qj})))
\end{aligned}$$

Note that the negative of the variational energy cancels with the negative of the Gaussian distribution.

We will now make two simplifying assumptions. These are motivated because our interest is in how minimising Free Energy drives disambiguatory sampling of a visual image; we can approximate certain parts of the system as *fast*, which means we do not need to integrate them directly. This saves conceptual and computational space. The first assumption is that the saccades consist of a fast stage (action) and a slow stage (perception). This means that we can approximate the fast stage as instantaneous, and teleport the point of foveation to μ_u (plus some noise) between each saccade. In effect, rather than providing a forward model where we assume $\dot{x}_p = 0$ and integrate the proprioceptive system, we can ignore the vp and xp terms entirely, and compute μ_u directly from the salience map, as discussed in Section 2.4.

Secondly, we assume that the dynamic renormalisation is fast. This allows us to discard the xq term from the variational energy, and approximate it by enforcing $\sum_j exp(\mu_{qj}) = 1$ at the end of every saccade. Removing these dynamics allows us to set $\mu'_q = 0$, which means that we can perform gradient descent without requiring the alternate derivative term to guarantee that the process reaches a minimum (see Chapter 1).

This reduces us to the following system:

$$\begin{aligned}
\dot{\mu}_{q1} &= -\kappa_1 \frac{\partial \mathcal{E}}{\partial \mu_{q1}} \\
\dot{\mu}_{q2} &= -\kappa_2 \frac{\partial \mathcal{E}}{\partial \mu_{q2}} \\
&. \\
&.
\end{aligned} \tag{2.10}$$

Where μ_{qi} is the state associated with the ith prior, and κ_i is the learning rate of the gradient descent, which was the same across all μ_{qi}. Note we do not need to perform the minimisation with respect to μ_p, due to the assumptions detailed above.

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial \mu_{qi}} &= \frac{\partial}{\partial \mu_{qi}}\Big(\frac{1}{2\Omega_{vq}}\varepsilon_{vq}^2\Big) \\ &= \frac{1}{\Omega_{vq}}\varepsilon_{vq}\frac{\partial \varepsilon_{vq}}{\partial \mu_{qi}}\end{aligned} \tag{2.11}$$

There is an important nuance in the coding of the derivative. To avoid computing the minimisation with respect to every pixel, we reduced the perceptual error, $\varepsilon_{vq} = (s_q - \sum_j exp(\mu_{qj})g(I_j, \mu_p))$, to a single number by calculating the mean square error (MSE) across all 256 individual channels. The ε_{vq} term in (3.12) is therefore straightforward to compute. However, the derivative of ε_{vq} cannot be reduced, as is generally the case (and as Friston describes in the original paper) to $\partial g^{(i)}/\partial \mu_{qi}$, where here $g^{(i)} = \sum_j exp(\mu_{qj})g(I_j, \mu_p)$. This is a result of the fact that MSE maps from $\mathbb{R}^{256}$ to $\mathbb{R}^1$; by doing so we lose structural information that plays a role in determining the rate of change of the error. To illustrate this, consider the simple two prior case below, when we derive $\partial \mathcal{E}/\partial \mu_{q1}$:

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial \mu_{q1}} &= \frac{\partial}{\partial \mu_{q1}}\Big(s_q - \sum_j exp(\mu_{qj})g(I_j, \mu_p)\Big) \\ &= \frac{\partial}{\partial \mu_{q1}}\Big(s_q - exp(\mu_{q1})g(I_1, \mu_p) - exp(\mu_{q2})g(I_2, \mu_p)\Big) \\ &= \frac{\partial}{\partial \mu_{q1}}\Big(s_q - exp(\mu_{q1})g(I_1, \mu_p)\Big) \\ &= \frac{\partial}{\partial \mu_{q1}}\frac{1}{N}\sum_1^N \Big(s_{qi} - exp(\mu_{q1})g_i(I_1, \mu_p)\Big)^2\end{aligned}$$

where $N = 256$ and s_{qi} and g_i denote the respective real valued inputs from the ith channels, in the range [0,1].

$$= -exp(\mu_{q1})\frac{2}{N}\sum_1^N \Big(s_{qi} - exp(\mu_{q1})g_i(I_1, \mu_p)\Big) \tag{2.12}$$

Here we are averaging across all 256 channels. The problem with this is that for every channel where $g_i \simeq 0$, the error for that channel will not change if μ_{q1} changes. By averaging all of the channels, however, we are implicitly assuming that it will. Thus (2.13) is not the true expression for $\partial \mathcal{E}/\partial \mu_{q1}$. To approximate the true expression, we only count the error across those channels where the value of g_i is appreciably bigger than zero. We still divide by 256, to encourage maximal information gain from each foveation.

The above constitutes the *perception* step of the dual minimisation discussed in Chapter 1. We integrated the system for 16 time steps with a learning rate

κ of 0.3 and an initial μ_p of [1500,1200], the approximate centre of the image. We then calculated the next location to saccade by constructing a Salience map. This location, μ_u, was then set (with the addition of a small amount of noise) equal to μ_p, and to x_p. The visual input s_q was recalculated from the generative process, and the next saccade begun. Constructing the Salience map and recomputing μ_u can be thought of as the *action* step of the dual minimisation, and will be discussed below.

2.4 Hidden Controls and Salience

Friston argues that we can approximate the action of a full hierarchy by modelling the system as minimising the entropy expected across world states at some future time $(t + \tau)$ [22]. The system selects the hidden control state, $\mathbf{u}$, represented by the system as μ_u accordingly:

$$\mu_u^* = \underset{\mu_u}{argmin}\ \mathbf{H}\big(q(\omega|\mu_{x,q}(t+\tau), \mu_u)\big) \tag{2.13}$$

Where * denotes the optimal μ_u. Other than stating that the 'the fictive prediction errors at each location were evaluated at their solution under the generative model, namely: $\mu_{x,q}(t+\tau) = \mu_{x,q}(t)$', the original paper does not explain how this is calculated. We have therefore taken our own approach.

Recall (Section 1.3) the central motivation behind Free Energy minimisation; we are minimising a measure of the distance between two probability distributions, the recognition density and the generative density. If we assume that the perception step has driven Free Energy to be a tight bound on surprise, it follows that the generative density $p(\omega, s)$ can be used as a good approximation to the recognition density, $q(\omega)$. This means that we can use the states μ_{qi}, which are the sufficient statistics of the generative density, as a way to calculate an approximation to the entropy of the recognition density.

In addition, we assume that the only possible world states (as far as the system is concerned) correspond to the priors, so that they represent the complete knowledge the system has of the world (we can assume that in a full biological system the subdivision of the world into just these prior expected states would have been learned). This means that $q(\omega)$ is a discrete probability distribution across the priors. As we have enforced $\sum_j exp(\mu_{qj}) = 1$, we take $exp(\mu_{qj})$ to be the softmax probability of the jth prior. Thus we can estimate the entropy of $q(\omega)$ as follows:

$$\mathbf{H}(q(\omega)) = -exp(\mu q1)\mu_{q1} - exp(\mu_{q2})\mu_{q2}\ ... - exp(\mu_{qM})\mu_{qM} \tag{2.14}$$

Where M is the number of priors.

To compute the entropy of the recognition density conditioned on some future states, $q(\omega|\mu_{x,q}(t+\tau), \mu_u)$, as with Friston, we constructed a map of salience $\mathbf{S}$ for a 32x32 grid of possible μ_u, i.e. possible points of foveation, across the image. Salience is the negative of the entropy, $\mathbf{H}$. For each point k we performed

a *fictive* integration for τ of 8 time steps, with $s_q = g(I_B, \mu_{uk})$, where I_B is the prior with the highest value μ_{qB}, i.e. the system's current best guess at the state of the world. We then used the states of the system after the fictive integration, $\{\mu_{q1}(t+\tau)...\mu_{qM}(t+\tau)\}$ to estimate the expected salience at that μ_{uk} at time $(t+\tau)$.

To drive the actual saccades, as with Friston we introduced an inhibition of return, which depressed the salience S at any particular point i at saccade k according to

$$\begin{aligned} S_{i,k} &= S_{i,k} - (S_{i,k} \times R_{i,k-1}) \\ R_{i,k-1} &= \rho(S_{i,k-1}) + \frac{1}{2} R_{i,k-2} \end{aligned} \tag{2.15}$$

where $\rho(S_{i,k})$ is a gaussian function, with a standard deviation 1/16 the size of the image, of the distance of i from the target of the kth saccade. Finally, μ_u^* was then taken to be the highest point on the salience map (as $argmax\ \mathbf{S} = argmin\ \mathbf{H}$), and the next point of foveation.

Assuming that the agent acts to minimise the entropy of world states comes directly, for Friston, as a necessary consequence of their existence [17]. In addition, if we recall that under ergodic assumptions the entropy of word states is the same as the long term average of surprise (1.10), and we note that by minimising Free Energy (as $-\ln p(\mu, s)$) we are at each time step minimising an upper bound on the surprise of sensory states, we can see that foveating the point of greatest expected future salience is equivalent to the minimisation of Free Energy. Note that by itself acting to foveate the most salient features of an image is equivalent to infomax, a pre-existing model of efficient action which Friston has shown to be a specific case of the FEP, given a sparse prior [20]. For us, the main difference is that our salience is counterfactual, derived by minimising Free Energy with respect to fictive sensory inputs.

By integrating into the future with respect to an expected (but fictional) sensory input, we are assuming that our priors model *counterfactual* aspects of the world. Thus, for the system, where it will look next is in part a function of what its current state leads it to believe about what the world will be like at some future time $(t+\tau)$. We could think of this counterfactual information in the real hierarchy as being encoded in the dynamical equations of motion, $f^{(i)}$, at each hierarchical layer. The main point is that the information gathered by foveating a local part of the image drives an expectation about, i.e. selects states μ_{qi} that correspond to, a particular global image. If we take our conscious experience to be a function of our current brain states, then it is hoped that this salience map captures something of the gestalt phenomenology that we experience, top down, as incoming sensory data from a small part of the world drives brain states to a Free Energy minimum.

Chapter 3

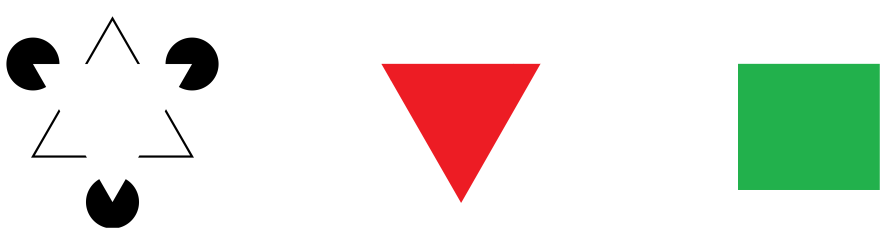
System: Evaluation and Discussion

3.1 Introduction

Here we present our investigations into (A) how exactly a Free Energy minimising system would act to disambiguate between competing prior expectations, and (B) whether the simplifications required to make a computationally tractable model rendered the resultant system empirically worthless. We follow each with a discussion of the implications and limitations of the investigation, along with suggestions for future investigation.

3.2 Disambiguation of Priors

We ran the system with an input image of Kanizsa's Triangle, and two priors: one rectangle, and one triangle (Fig. 3.1).

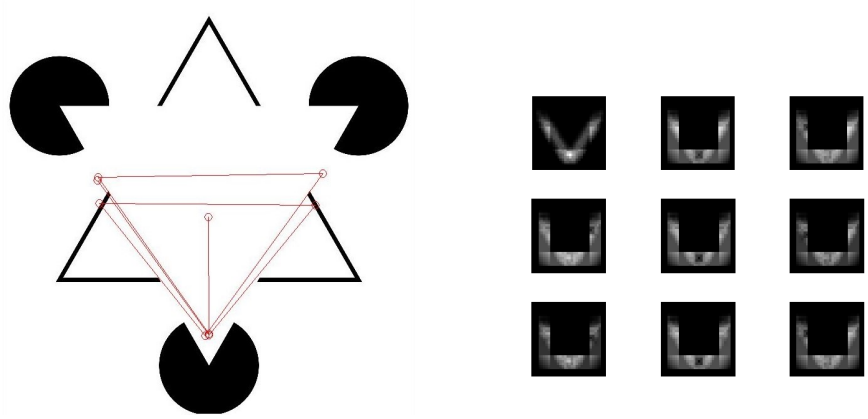


(a) Kanizsa's Triangle (b) Triangle Prior (c) Rectangle Prior

Figure 3.1: Image and priors for investigation of disambiguation.

Examination of the salience maps driving each saccade shows the system to be acting to test negative hypotheses: once the system has decided it is looking

at a triangle, the salience map (Fig. 3.2(b)) restricts itself to those parts of the image where a square would be if it were wrong.



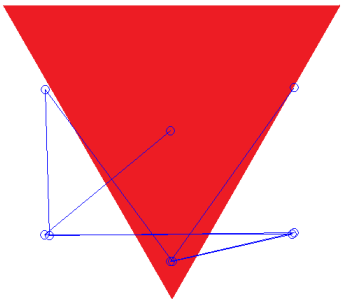
(a) Saccades (b) Driving Salience Map

Figure 3.2: Nine-saccade examination of Kanizsa's Triangle, and associated salience map. Saccades 1-9 are driven by map top-left to bottom-right.

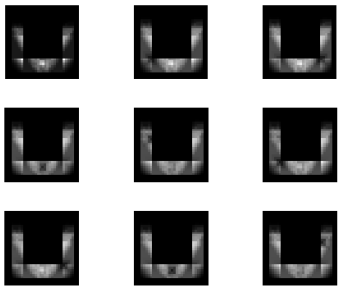
This was theorised to be due to a larger contribution to $\dot{\mu}_{qj}$ from negative results than from positive (see (2.13)), combined with the renormalisation $\sum_j \mu_{qj} = 1$ at each stage (see discussion at the end of p27). This was due to a combination of factors which meant that for those channels where the prior and the image agreed, they generally agreed fully. This resulted in a zero or near zero error value, and thus no contribution to $\partial\mathcal{E}/\partial\mu_{qj}$. Hence, the locations on the salience map that drove μ_{q1} and μ_{q2} apart most rapidly (and thus maximise salience by creating the most peaked distribution $q(\omega)$) were those where there should be the most *absent* information about the other prior - the system expects to gain the most information by confirming its negative expectations.

To test this, we examined the situation where the system was given the triangle prior as its input image. Because the priors are two simple shapes, the resultant salience maps are easy to distinguish. Figure 3.3 shows the disconfirmatory case. As we can see, the system became steadily more convinced that it was looking at a triangle, but its pattern of sampling was a combination of disconfirmatory and disambiguatory; the latter because it was somewhat interested in those areas near the intersections of the edges of the priors. This is because there was some positive contribution from those places where the prior and the image did not agree fully, so $\dot{\mu}_{q1}$ would be slightly positive, which drove μ_{q1} slightly further from μ_{q2} than a large negative $\dot{\mu}_{q2}$ by itself (assuming μ_{q1} and μ_{q2} correspond to triangle and rectangle priors respectively).

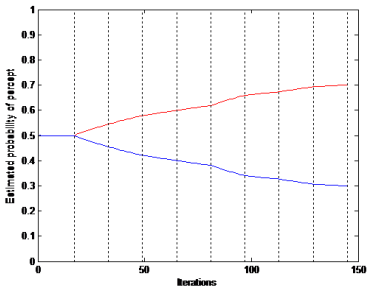
Secondly, we set the system up so that the weighting contribution $exp(\mu_{qj})$ (see 2.13) was such that unless the system was fully convinced that it was looking at the jth prior, there would still be a large enough positive contribution to the respective $\dot{\mu}_{qj}$ to offset negative contributions from the competing prior. We



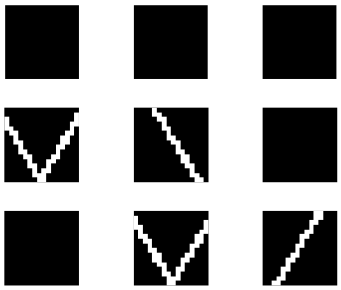
(a) Saccades



(b) Driving Salience Map



(c) Derived probability of world states (R=Triangle, B=Rectangle)

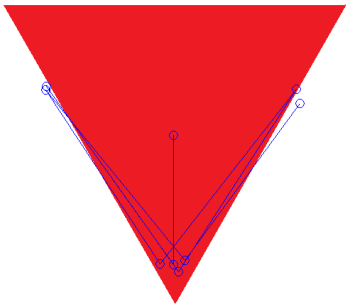


(d) s_q: sensory input

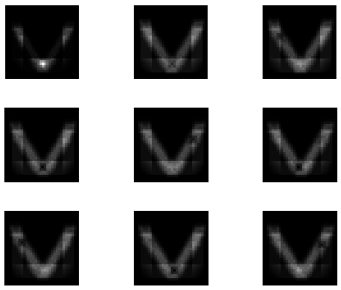
Figure 3.3: Nine-saccade when disconfirming system presented with a triangle.

can see the results for a nine-saccade run in Figure 3.4. As with Figure 3.3(c), the system becomes steadily more convinced that it is looking at a triangle (at about the same rate). However, we can see that it is now primarily acting to disambiguate; it foveates the bottom corner of the triangle and the two points of intersection between the sides of the priors. We can hypothesise that the bottom corner is the most appealing initial saccade due to the fact that it contains two intersections, and therefore the most disambiguatory information.

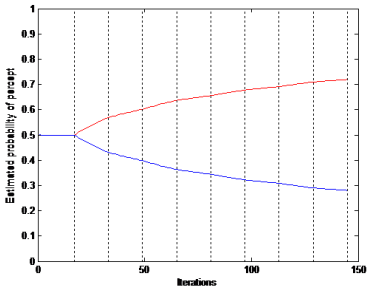
Interestingly, disambiguation still appears to rate much more highly than confirmation. Whilst we can see in Figure 3.4(b) that the system is willing to consider sampling to confirm its current hypothesis, i.e. to sample the sides of the triangle, it still expects to receive more useful information at the points of intersection. We could expect to see confirmatory saccades only if the inhibition of return were to last long enough (and be large enough) that the three main peaks on the salience map, once foveated, remained suppressed long enough for



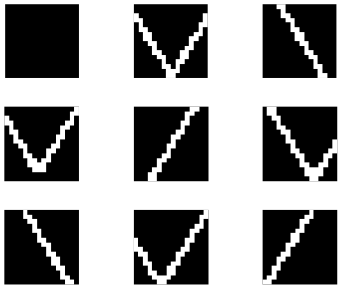
(a) Saccades



(b) Driving Salience Map



(c) Derived probability of world states (R=Triangle, B=Rectangle)



(d) s_q: sensory input

Figure 3.4: Nine-saccade when disambiguating system presented with a triangle.

other sampling to promise more information.

Finally, it should be noted that in the salience maps, the top edge is never expected to provide salient information. This is a consequence of the fact that the top edges of the two priors coincide, and so no useful information will be forthcoming.

3.3 Disambiguation Discussion

These results have interesting implications. If we are correct to approximate the action which best minimises Free Energy as that which maximises salience, and our other simplifications are valid, then we can make a hypothesis about the likely points of foveation if we prime a subject with simple priors. We can hypothesise that the first few targets of saccades will be those locations which most disambiguate. However, according to our model, if the subject

becomes convinced that what they are seeing is indeed a particular prior (in our terms, if for a particular j, $\mu_{qj} \approx 1$), then the positive contribution to that $\dot{\mu}_{qj}$ will be close to zero, and they would be expected to sample to disconfirm the alternative.

There are two large caveats to this hypothesis. Firstly, it will only hold if our GM, (2.3) is indeed a reasonable approximation to the workings of a full predictive hierarchy. This seems unlikely; having a single state approximate some set of various brain states that contribute to a particular gestalt experience is useful for our purposes of drawing graphs, but almost certainly fails to capture exactly what is going on in the brain. What, for example, is the neural correlate to this sort of disambiguation? With respect to a constructed PP hierarchy, what would be the multilayer correlate to the single state μ_{qj}?

Secondly, disconfirmation may be a result of having only two priors. Consider that we calculate salience as the weighted sum of the negative surprise of each prior (2.15). With two priors, a sharply peaked distribution can be achieved by either depressing the probability of the prior we consider to be wrong, or increasing the probablity of that we take to be correct. This corresponds to disconfirming or confirming. However, with *multiple* priors, given we may only sample one part of the image at a time, it is almost always going to be more efficient to sample to confirm, because increasing the probability of our current expectation will depress the probabilities of all of the other priors (due to normalisation). Acting to disconfirm will only be effective in the case where most alternatives coincide at a specific point; so that foveating that point will simultaneously depress the posterior probabilities of multiple competing inferences.

In short, we think that it is necessary to model both the case of multiple priors, and for those priors to be the product of a full predictive hierarchy, before results can be produced which would have a good claim to be human-like.

3.4 Distribution of rate of percept alternation

To examine whether our non-hierarchical, two-prior model was empirically useful, we set it up to compare to a well known result in cognitive psychology; that in the temporal dynamics of bistable perception, the durations of percepts follow a gamma distribution [8]. This phenomenon is usually grounded in neural mechanics, but as Friston explicitly draws close parallels between his account of the FEP and the neuronal activity of the brain [20], we do not consider it unreasonable to see whether this characteristic behaviour is simulated by a Free Energy minimising model.

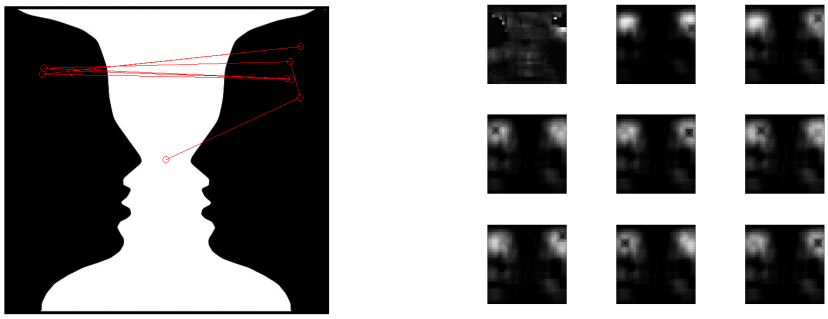
Initially we attempted to simulate the bistable image using the classic Rubin's Vase illusion (Fig. 3.5). However, due to our priors being hierarchically flat, there was no way to provide the system with a sense of which prior was which, as the driving ambiguity of the illusion is the fact that a single line can be seen as contributing to the evidence for two distinct gestalts. Recall however that the counterfactual salience maps in the Kanizsa case assigned no value to

(a) Rubin's Vase (b) Faces Prior (c) Vase Prior

Figure 3.5: Bistable: Rubin's Vase and Priors.

the top edge, which was shared by the two priors (Fig 3.3(b),3.4(b)). Similarly here, the system presupposes no information will be gleaned from the shared line, and so will not be driven to it.

We attempted to distinguish the priors by adding more information to the 'faces' in the one case and the 'vase' in the other (in the form of lines), however the system simply foveated the area of most information (see Fig. 3.6), and failed to settle into a properly bistable regime. We therefore moved on from Rubin's Vase.

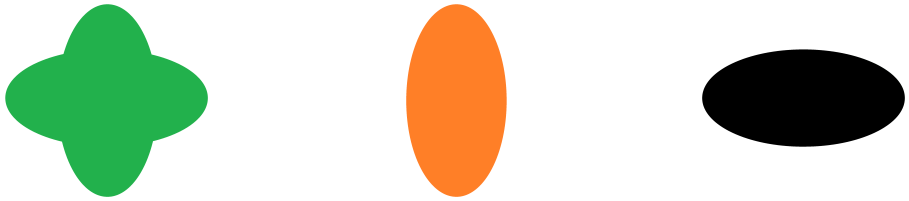


(a) Saccades (b) Driving Salience Map

Figure 3.6: Nine-saccade examination of Rubin's Vase, and associated salience map. Saccades 1-9 are driven by map top-left to bottom-right.

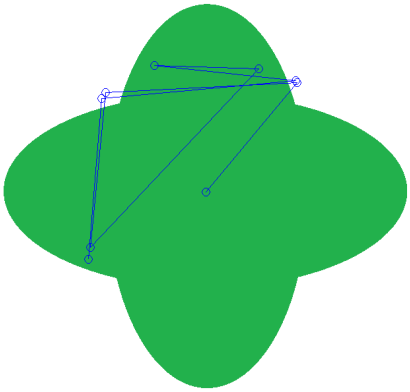
To take into account the limitations of our hierarchically flat priors, we instead constructed an image and prior pair so that the image would provide roughly equal evidence for both priors, such that the interaction of the saccades with the minimisation should result in the system alternating its inference on the causes of its sensory input (see Fig. 3.7). We ran the system with this input image and these priors for an initial nine saccades, to confirm that the resultant system did indeed alternate between percepts (Fig. 3.8).

Having built a bistable system, we ran it for 200 saccades, and produced a

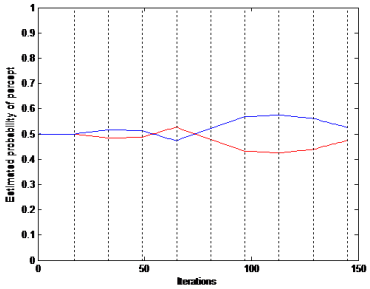


(a) Ambiguous Image (b) First Prior (c) Second Prior

Figure 3.7: Bistable: Hierarchically Flat Ambiguity



(a) Saccades



(b) Derived probability of world states (R=Triangle, B=Rectangle)

Figure 3.8: Nine-saccade for hierarchically flat bistable system

histogram of the duration of percepts. By dividing through by the area of each bar, we obtained an approximation to the pdf of the rate of alternation (Fig. 3.9). As can be seen, this is not a gamma distribution.

3.5 Bistability discussion

The attempt to use Rubin's vase to simulate a case of bistable perception offered an interesting insight into bistable perception under the FEP account. It highlights that these phenomena probably require the presence of a hierarchy, as a single level of ambiguity can possess no salience; the system will sample anywhere but the ambiguous region. The 'gestalt flip' that characterises our phenomenal experience when we are presented with such images is likely due to the fact that ambiguities between larger structures only emerge at high levels of abstraction in the hierarchy. The Rubin's vase figure is not locally ambiguous - a line is just a line - it is only with respect to high level priors of things like faces

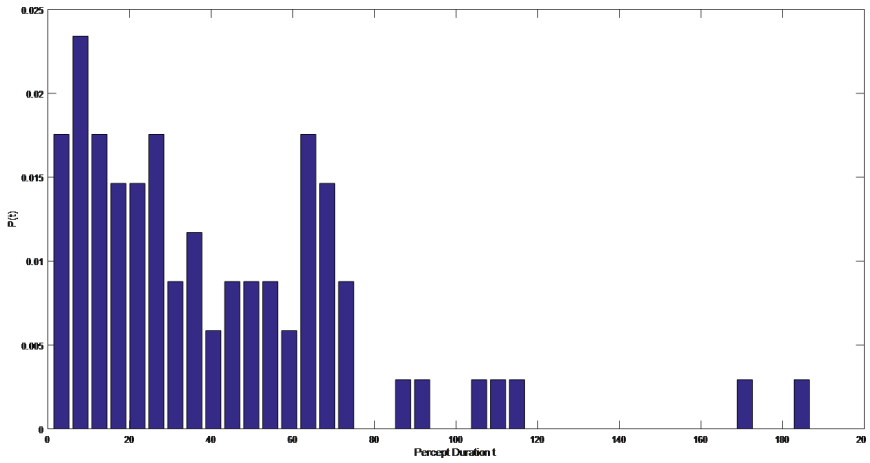


Figure 3.9: Distribution of Percept Duration after 200 saccade run

that the ambiguity emerges. Indeed, it can also be seen as evidence of action to confirm; because as we have shown, neither disconfirmation nor disambiguation would involve the foveation of the ambiguous region.

Thus it is the interplay of expectations across the hierarchy that must give rise to the perceptual bistability experienced by humans. Sensory signals prompt a particular high-level prior, which causes proprioceptive expectations to flow down and drive action. This action gathers evidence that promotes a competing expectation, whilst also confirming the original expectation. Further action to confirm the original may result in no more evidence (if we look for facial features where there are none, for example), and so the second prior then supplants the first.

If this account is correct, then it is unsurprising that a simulation of Free Energy minimisation without a full hierarchy failed to produce the expected gamma distribution. As with our investigations into disambiguating priors, we are limited by the same two oversimplifications.

Conclusion

Our system does not understand structure, and our system thinks that the world is effectively binary. Compensating for these limitations would provide potentially interesting results. Building a hierarchy would allow the system to simulate both an understanding of structure and the dynamical delays we hypothesise are the result of the interactions between high level priors and low level action. Adding more priors should, as discussed in Section 3.3, move the action of the system from disconfirmation/disambiguation to confirmation/disambiguation.

On a more positive note, our modelling has provided fertile ground for investigation into the way in which the FEP could drive the disambiguation of priors. We have shown that counterfactual PP is computationally tractable, and that the results provide some insight into how minimising Free Energy drives action. In addition, we have made minor hypotheses about a suitably primed subject's likely targets of foveation when presented with Kanizsa's triangle.

As discussed, the FEP's attraction is in the claims it makes to unite phenomena as wide ranging as consciousness and action [17][13]. It is the belief of the author that understanding intelligence requires understanding the informational constraints placed upon an organism as a driving force behind its evolution. Thinking of organisms as experimenters, constantly sampling and updating, poised between exploitation and exploration [21], between overfitting and underfitting, is a very appealing way to approach this. We hope that our deconstruction of the workings of Free Energy minimisation at the level of active inference has provided an interesting and stimulating read.

Bibliography

[1] Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries and Karl J. Friston, *Canonical Microcircuits for Predictive Coding*, Neuron 76(4) 695-711, November 21, 2012. doi: 10.1016/j.neuron.2012.10.038

[2] W Ross Ashby, *Design for a brain*, London, UK: Chapman and Hall, 1952.

[3] W Ross Ashby, *An introduction to cybernetics*, London, UK: Chapman and Hall, 1956.

[4] Tim Bayne and David Chalmers, *What is the unity of consciousness?*, in *The Unity of Consciousness, Binding, Integration, Dissociation*, ed Cleeremans, Oxford, UK: Oxford University Press, 2003.

[5] Christopher Bishop *Neural Networks for Pattern Recognition*, Oxford, UK: Oxford University Press, 1995.

[6] Christopher Bishop, *Pattern Recognition and Machine Learning*, New York, USA: Springer-Verlag, 2006.

[7] Stephen J. Blundell and Katherine M. Blundell, *Concepts in Thermal Physics*, Oxford, UK: Oxford University Press, 2006.

[8] Jan W. Brascamp, Raymond van Ee, Wiebe R. Pestman and Albert V. van den Berg, *Distributions of alternation rates in various forms of bistable perception*, Journal of Vision 5: 287-298, 2005. doi: 10.1167/5.4.1

[9] William L. Briggs and Van Emden Henson, *The DFT: An Owners' Manual for the Discrete Fourier Transform*, Society for Industrial and Applied Mathematics, January 1987.

[10] Chris Buckley, Chang Sub Kim, Simon M. McGregor and Anil K. Seth, *The free energy principle in neuroscience: A technical evaluation*, in preparation.

[11] Thomas Buhrmann, Ezequiel Alejandro Di Paolo and Xabier Barandiaran, *A dynamical systems account of sensorimotor contingencies*, Frontiers in Psychology 285(4):1-19, May 2013. doi: 10.3389/fpsyg.2013.00285

[12] George Casella and Roger L. Berger, *Statistical Inference* Pacic Grove, US: Duxbury, 2002.

[13] Andy Clark, *Whatever next? Predictive brains, situated agents, and the future of cognitive science*, Behavioural and Brain Sciences 36(3): 2013. doi: 10.1017/S0140525X12000477

[14] Peter Dayan, Geoffrey E. Hinton and Radford M. Neal *The Helmholtz Machine* Neural Computation 7: 889-904, 1995.

[15] Barry Dainton, *Stream of Consciousness: Unity and Continuity in Conscious Experience*, London, UK: Routledge, 2000.

[16] Yadolah Dodge and David Cox, *The Oxford Dictionary of Statistical Terms*, Oxford, UK: Oxford University Press, September 7 2006.

[17] Karl J Friston. *A Theory of Cortical Responses*, Phil. Trans. R. Soc. B 360: 815836, 2005. doi: 10.1098/rstb.2005.162

[18] Karl Friston, *Variational ltering*, NeuroImage, 41:747766, 2008.

[19] Karl Friston. *The free-energy principle: a rough guide to the brain?*, Trends in Cognitive Sciences 13(7):293-301, 2009. doi: 10.1016/j.tics.2009.04.005

[20] Karl Friston, *The free-energy principle: a unified brain theory?*, Nature Reviews, Neuroscience, Volume 11:127-138, February 2010. doi: 10.1038/nrn2787

[21] Karl Friston, *Embodied inference: Or I think therefore I am, if I am what I think*, The implications of embodiment(Cognition and Communication),ed. W. Tschacher and C. Bergomi, pp. 89125. Imprint Academic, 2011.

[22] Karl Friston, Rick A Adams, Laurent Perrint and Michael Breakspear, *Perceptions as hypothesis: Saccades as Experiments*, Frontiers in Psychology 151(3):1-20, May 2012. doi: 10.3389/fpsyg.2012.00151

[23] Karl J. Friston, Jean Daunizeau, James Kilner and Stefan J. Kiebel, *Action and behavior: a free-energy formulation*, Biological Cybernetics, published online 11 February 2010. doi: 10.1007/s00422-010-0364-z

[24] Karl Friston, James Kilner, and Lee Harrison, *A free energy principle for the brain*, J Physiol Paris, 100:7087, 2006.

[25] Karl Friston, Jeremie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny, *Variational free energy and the Laplace approximation*, Neuroimage, 34:220234, 2007.

[26] Hermann von Helmholz, *Treatise on physiological optics*, Hamburg: Voss, 3rd edition, 1909.

[27] Geoffrey E. Hinton and Richard S. Zemel, *Autoencoders, Minimum Description Length and Helmholtz Free Energy*, Advances in Neural Information Processing Systems 6 (NIPS 1993), accessed via www.cs.toronto.edu/ fritz/absps/cvq.pdf

[28] Jakob Hohwy, *Attention and conscious perception in the hypothesis testing brain*, Frontiers in Psychology 96(3):1-14, April 2012. doi: 10.3389/fpsyg.2012.00096

[29] Jakob Hohwy, *The predictive mind*, Oxford, UK: Oxford University Press, 2013.

[30] Toshihiko Hosoya, Stephen A. Baccus and Markus Meister, *Dynamic predictive coding by the retina*, Nature 436(7):7177, 2005.

[31] Kevin G. Kirby, *A Tutorial on Helmholtz Machines*, June 2006, accessed via www.nku.edu/∼kirby/docs/HelmholtzTutorialKoeln.pdf

[32] Uriah Kriegel, *Consciousness as Intransitive Self-Consciousness: Two Views and an Argument*, Canadian Journal of Philosophy, 33(1): 9, March 2003.

[33] D. Marr and E. Hildreth, *Theory of Edge Detection*, Proc. R. Soc. Lond. B 207:187-217, 1980. doi: 10.1098/rspb.1980.0020

[34] Elizabeth Michael, Vincent de Gardelle, and Christopher Summerfield, *Priming by the variability of visual information* , PNAS January 2014. doi: 10.1073/pnas.1308674111

[35] David Mumford, *On the computational architecture of the neocortex. II. The role of cortico-cortical loops*, Biological Cybernetics 66(3):24151, 1992. doi: 10.1007/BF00198477

[36] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Upper Saddle River, USA: Prentice Hall, 3rd Edition, 2010.

[37] Anil K. Seth, *Interoceptive inference, emotion and the embodied self*, Trends in Cognitive Sciences xx:19, 2013. doi: 10.1016/j.tics.2013.09.007

[38] Anil K. Seth, *The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies*, In T. Metzinger and J. M. Windt (Eds). Open MIND: 35(T). Frankfurt am Main: MIND Group, 2015. doi: 10.15502/9783958570108

[39] Claude Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 27 (3): 379423, July-October 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x

[40] S Zeki, *The Disunity of Consciousness*, TRENDS in Cognitive Sciences, 7(5): 215, May 2003.

Appendix: Code

```
clear all
close all

%initialise image and priors - MUST BE SAME SIZE; of images provided, sets
%are
%{CrossOvals.png,HorzOval.png,VertOval.png},{Rubin.png,Vase2.png,faces.png},
%and {KaniszaBig.png,KaniszaBig2.png,TrianglePrior.png,RectanglePrior.png}.
%Note mispelling of Kanizsa was to differentiate with a different set of
%images.
F1 = fspecial('gaussian',5,1);%Difference of Gaussian filter.
F2 = fspecial('gaussian',5,4);
DoG = F1-F2;

preImage = imread('KaniszaBig.png');%Image
greyImage = rgb2gray(preImage);
ImageSmall = conv2(double(greyImage),DoG,'valid');
[Image,transIm] = margin(ImageSmall);

imsize = size(Image);

prePrior1 = imread('TrianglePrior.png');%First prior
greyPrior1 = rgb2gray(prePrior1);
Prior1Small = conv2(double(greyPrior1),DoG,'valid');
[Prior1,transP1] = margin(Prior1Small);

prePrior2 = imread('RectanglePrior.png');%Second prior
greyPrior2 = rgb2gray(prePrior2);
Prior2Small = conv2(double(greyPrior2),DoG,'valid');
[Prior2,transP2] = margin(Prior2Small);
```

```
%initialise system
numsaccades = 9;
saccadelength = 16;
tau = 8;%counterfactual integration length

x_p = [floor(imsize(1)/2) floor(imsize(2)/2)]';%true proprioceptive states

s_p = x_p + [floor(10*randn(1)) floor(10*randn(1))]';%perceptual states: assuming some noise
s_q = G(Image,x_p); %+ floor(randn(1)/30); - s_q noise terms currently suppressed due to chaotic behaviour of system.

mu_p = s_p;%expected proprioceptive states
mu_p_dash = zeros(2,1);

mu_q = [-0.6931 -0.6931]';%expected perceptual states; initialise at approx p(P1)=p(P2)=0.5
mu_q_dash = zeros(2,1);

precisionVP = 8;%precisions in F (how certain the system is of its models)
precisionXP = 8;
precisionVQ = 4;
precisionXQ = 4;

%Saccades
saccadestore = zeros(2,numsaccades);
probscore = zeros(2,numsaccades);
pointlist = [16 16]';
maplist = zeros(32,(32*numsaccades));
stateprogress = zeros(2,(numsaccades*saccadelength+1));
stateprogress(:,1) = exp(mu_q);
for i = 1:numsaccades
    saccadestore(:,i) = mu_p;
    %perception step: minimise FE to a tight bound
    [mu_q,mu_q_dash,states] = int(s_q,mu_q,mu_q_dash,mu_p,Prior1,Prior2,saccadelength);
    probscore(:,i) = exp(mu_q);
    %action step: act to minimise surprise of sensory input, given model
    [mu_u,pointlist,map] = computeU(mu_q,mu_q_dash,Prior1,Prior2,tau,transIm,pointlist);
    mu_p = mu_u +[floor(10*randn(1)) floor(10*randn(1))
```

```
        ]';%adding some noise to teleport.
    s_q = G(Image,mu_p);% + floor(randn(1)/30);
    %recording the salience maps
    for j = 1:32
        for k = 1:32
            maplist(j,(k+32*(i-1))) = map(j,k);
        end
    end
    %recording the states
    stateprogress(:, (((i-1)*saccadelength)+2:(i*saccadelength)+1)) = states;
    i %- uncomment this if doing a long run:  helps keep track of how it's
    %running. One saccade is approx 40seconds on 16GB RAM.
end

%Imaging. Note fig (2) and fig(4) only work with exactly 9 saccades; will
%require adjustment or commenting out if running for fewer or more. fig(5)
%is set up for long runs.
figure(1)
imshow(preImage);
coords = zeros(2,numsaccades);
for i=1:numsaccades
    for j=1:2
        coords(j,i) = saccadestore(j,i)-transIm(j,1);
    end
end
hold on
plot(coords(2,:),coords(1,:),'-bo','MarkerSize',10);

figure(2)
for k = 1:9
    im = zeros(32,32);
    for i =1:32
        for j=1:32
            im(i,j) = maplist(i,(j+32*(k-1)));
            subplot(3,3,k)
            imshow(im)
        end
    end
end

figure(3)
```

```
x = stateprogress(1,:);
y = stateprogress(2,:);
plot(x,'r')
hold on
plot(y,'b')
ylim([0 1])
for d= 1:numsaccades
    x = (saccadelength)*(d)+1;
    plot([x x],[0 1],':k')
end
xlabel('Iterations');
ylabel('Estimated probability of percept');

figure(4)
for k = 1:9
    image = G(Image,saccadestore(:,k));
    subplot(3,3,k)
    imshow(image)
end

%figure(5)
l = size(stateprogress);
sp = zeros(2,l(2));
sp(1,:) = stateprogress(1,:)+0.05;%adjusting as image biased slightly towards one percept
sp(2,:) = stateprogress(2,:)-0.05;
l = l(2);
PD = [];%don't know in advance number of switches, so cannot preallocate.
count = 0;
score = 0;
for z = 1:l
    top = sp(1,z);
    bottom = sp(2,z);
    if z==l
        if count ~= 0
            PD = [PD count];
        end
    elseif top == bottom
        if score == 1
            PD = [PD count];
            count = 0;
            score = 0;
        elseif score == -1
            PD = [PD count];
            count = 0;
```

```
            score = 0;
        end
    elseif top > bottom
        if score == 1
            count = count +1;
        elseif score == -1
            PD = [PD count];
            count = 1;
            score = 1;
        elseif score == 0
            score = 1;
            count = 1;
        end
    elseif bottom > top
        if score == 0
            score = -1;
            count = 1;
        elseif score == -1
            count = count + 1;
        elseif score == 1
            PD = [PD count];
            count = 1;
            score = -1;
        end
    end
end
%[f,x] = hist(PD,40);
%bar(x,f/trapz(x,f));
%xlabel('Percept Duration t');
%ylabel('P(t)');
```

```
function [ MarginImage, translation ] = margin( image )
%Margin builds a margin of blackspace around the original image, so that
%foveating near to the edge of an image does not throw index out of bounds
%errors.

imsize = size(image);
Vincrease = floor(imsize(1)*(3/2));%creating margins, to avoid indexing errors
Hincrease = floor(imsize(2)*(3/2));
imageL = zeros(Vincrease,Hincrease);
marginTop = floor(imsize(1)*(1/4));
marginLeft = floor(imsize(2)*(1/4));
translation = [marginTop marginLeft]';
```

```
for i = 1:imsize(1)
    for j = 1:imsize(2)
        imageL(marginTop+i,marginLeft+j) = image(i,j);
    end
end

MarginImage = imageL;
end
```

```
function [ output ] = G( image, location )
%G returns a 256 element vector where each element corresponds to a visual
%channel displaced across a grid 1/6 of the size of the original image (1/9 the
%size of the margined image) centred on the location.

PoF = location;%Point of foveation
imsize = size(image);
%imshow(image,'InitialMagnification',25);
verticalfov = floor(imsize(1)/9);
horizontalfov = floor(imsize(2)/9);
FoVdim = [verticalfov horizontalfov]';%Field of view dimension


%imshow(imageL,'InitialMagnification',17);
topleft = PoF - FoVdim./2;%Top left of field of view
visualInput = zeros(16);

minVI = Inf;
maxVI = -Inf;
for i = 1:16
    for j = 1:16
        %extract channel area (note already DoG filtered in MainRun).
        indexvert = ceil(topleft(1)+(i-1)*(verticalfov/16)):floor(topleft(1)+(i)*(verticalfov/16));
        indexhorz = ceil(topleft(2)+(j-1)*(horizontalfov/16)):floor(topleft(2)+(j)*(horizontalfov/16));
        channelij = image(indexvert,indexhorz);
        visualInput(i,j) = max(max(channelij));
        if visualInput(i,j) > maxVI
            maxVI = visualInput(i,j);
        elseif visualInput(i,j) < minVI
            minVI = visualInput(i,j);
```

```
        end
    end
end

%if (max(visualInput(:)) - min(visualInput(:)))>0.001% prevent minute rounding errors from producing vestigal effects.
    %visualInput = (visualInput-min(visualInput(:)))/(max(visualInput(:))-min(visualInput(:)));
%end
%output = visualInput;
output = im2bw(visualInput);%binary (channel is informative/isn't) can give better runs, in some circumstances.
%imshow(output);
end
```

```
function [mu_q_new,mu_q_dash_new,states] = int(s_q,mu_q,mu_q_dash,mu_p,P1,P2,iterations)
%Int integrates the perceptual system by one time-step, according to the
%minimisation of E (the variational energy) with respect to each mu state.
%Note int does not integrate the proprioceptive system, as we are
%approximating that integration as 'fast'. Dynamic integration (involving mu_qi_dash states) has been
%commented out for the actual run.

k = 0.3;%learning rate

deriv_mu_qi = @(e1,dG_d_mu_qi) 4*e1*(dG_d_mu_qi);%-4*e2*(exp(mu_qi))-(1/1024)*4*e2*mu_qi;
%deriv_mu_qi_dash = @(e2) 4*e2;

states = zeros(2,iterations);

for j = 1:iterations
    mu_q1 = mu_q(1);
    mu_q2 = mu_q(2);
    %mu_q1_dash = mu_q_dash(1);mu_q2_dash = mu_q_dash(2);%splitting the states to respective priors

    expectedQ = exp(mu_q1).*G(P1,mu_p)+exp(mu_q2).*G(P2,mu_p);
    error_mu_q = squarediff(s_q,expectedQ);%causal error
```

```
terms are the same for all prior states
    dG_d_mu_q1 = dG(s_q,mu_q1,mu_p,P1);%rate of change of error with respect to visual states mu_q
    dG_d_mu_q2 = dG(s_q,mu_q2,mu_p,P2);

    %error_mu_q1_dash = mu_q1_dash - 1 + exp(mu_q1)+exp(mu_q2)+(1/1024)*mu_q1;%first prior dynamic error
    %error_mu_q2_dash = mu_q2_dash - 1 + exp(mu_q1)+exp(mu_q2)+(1/1024)*mu_q2;%second prior dynamic error

    mu_q1 = mu_q1 + k*deriv_mu_qi(error_mu_q,dG_d_mu_q1);%+mu_q1_dash;%update by gradient descent
    mu_q2 = mu_q2 + k*deriv_mu_qi(error_mu_q,dG_d_mu_q2);%+mu_q2_dash;

    %mu_q1_dash = mu_q1_dash - k*deriv_mu_qi_dash(error_mu_q1_dash);%update dynamic model
    %mu_q2_dash = mu_q2_dash - k*deriv_mu_qi_dash(error_mu_q2_dash);

    normaliser = exp(mu_q1)+exp(mu_q2);%renormalise states (if dynamic renormalisation, no need to do this)
    ex1 = exp(mu_q1)/normaliser;
    ex2 = exp(mu_q2)/normaliser;
    mu_q1 = log(ex1);
    mu_q2 = log(ex2);

    states(1,j) = exp(mu_q1);%record causal states for plotting
    states(2,j) = exp(mu_q2);

    mu_q(1) = mu_q1;%recombine causal states
    mu_q(2) = mu_q2;

    %mu_q_dash(1) = mu_q1_dash;%recombine dynamic states
    %mu_q_dash(2) = mu_q2_dash;
end

mu_q_new = mu_q;
mu_q_dash_new = mu_q_dash;
end
```

```

function [U, pointlist2, map] = computeU( mu_q,mu_q_dash,Prior1,Prior2,tau,translation,pointlist)
```

```
%This generates a Salience map across the image, and returns the
%co-ordinates of the highest point on the map.

mu_q1 = mu_q(1);
mu_q2 = mu_q(2);
trans1 = translation(1);
trans2 = translation(2);
if mu_q1 >= mu_q2%picking what the system thinks is the real world
    fictiveimage = exp(mu_q1)*Prior1;
else
    fictiveimage = exp(mu_q2)*Prior2;
end
imsize = size(fictiveimage);
vertgap = imsize(1)/51;%1/34 of pre-margined image
horzgap = imsize(2)/51;

Umap = zeros(32,32,3);%First layer stores salience values, second two store co-ordinates.

for i=1:32
    for j=1:32
        mu_u = [trans1+i*vertgap trans2+j*horzgap]';%compute ficitve foveation
        Umap(i,j,2) = mu_u(1);
        Umap(i,j,3) = mu_u(2);
        s_q = G(fictiveimage,mu_u);%compute fictive perceptual input
        [ficmu_q,ficmu_q_dash] = int(s_q,mu_q,mu_q_dash,mu_u,Prior1,Prior2,tau);%counterfactual integration
        ficmu_q1 = ficmu_q(1);
        ficmu_q2 = ficmu_q(2);
        salience = (exp(ficmu_q1)*ficmu_q1 + exp(ficmu_q2)*ficmu_q2);
        Umap(i,j,1) = salience;%building the salience map.
    end
end

salmax = -Inf;%convert salience map to range 0-1, for ease of viewing.
salmin = Inf;
for i=1:32
    for j=1:32
```

```
        if Umap(i,j,1)>salmax
            salmax = Umap(i,j,1);
        end
        if Umap(i,j,1)<salmin
            salmin = Umap(i,j,1);
        end
    end
end

ump = Umap(:,:,1);
ump = (ump-salmin)/(salmax-salmin);
%imshow(ump);

%suppress map to simulate inhibition of return.
gauss = @(gap) 1/(2*sqrt(pi))*exp(-((gap)^2)/4);
s = size(pointlist);
histlen = s(2);
for i=1:32
    for j=1:32
        total = 0;
        for k = 1:histlen
            histi = pointlist(1,k);
            histj = pointlist(2,k);
            dev = max([abs(histi-i),abs(histj-j)]);
            if and(dev < 5,k==histlen)
                total = total + gauss(dev);
            elseif dev < 5
                total = total + (0.5^(histlen-k))*gauss(dev);
            end
        end
        pixel = ump(i,j);
        pixel = pixel*(1-(3)*total);
        ump(i,j) = pixel;
    end
end

bestij = zeros(2,1);%compute max U across Map.
maximum = -Inf;
for i=1:32
    for j=1:32
        if ump(i,j)>maximum
            maximum = ump(i,j);
            bestij(1) = i;
            bestij(2) = j;
        end
```

```
    end
end

I = bestij(1);
J = bestij(2);
IJ = [I J]';
pointlist2 = horzcat(pointlist,IJ);
U = [Umap(I,J,2) Umap(I,J,3)]';
map = ump;
end
```

```
function [ dGout ] = dG( senseq,mu_q,mu_p,Prior)
%dG returns a scalar corresponding to the nonlinear approximation to the
%rate of change of the error with respect to a particular prior image.

pview = G(Prior,mu_p);
count = 0;
derror = 0;
for i = 1:16
    for j = 1:16
        if pview(i,j) > 0 %ignoring those pixels for which an increase in mu_q will not contribute to an increase in dG/dmu_q
            derrorij = senseq(i,j) - exp(mu_q)*pview(i,j);
            derror = derror + derrorij;
            count = count + 1;
        end
    end
end

if derror == 0%avoiding infinite dG
    derror = derror;
else
    derror = derror/256;
end

%max(max(senseq));
%max(max(pview));
dGout = derror*exp(mu_q);
end
```

```
function [ difference ] = squarediff( array1,array2 )
%squarediff computes the least squares difference between two separate arrays,
```

```
%and returns a single scalar value. Separated from the rest of the run for
%convenience wrt trying different measures of distance.

S1 = size(array1);
S2 = size(array2);
sim = isequal(S1,S2);

if sim == 0%check they are the same size first
    error('Arrays submitted to squarediff are unequal');
else
    errors = array1-array2;
    diff = mean(mean(errors.^2));
end

difference = diff;
end
```