Emergence of Perceptual Unity, Selfhood and Qualia in a Hierarchical Generative Model

Oscar Gilg

St Anne's College

University of Oxford

Supervised by Kobi Kremnitzer

Trinity 2023

Abstract

This dissertation explores the idea that consciousness can arise from a mathematical model. Borrowing from illusionism, we propose a scientific template for how to study consciousness. Instead of focusing on the nature of consciousness, we investigate how behaviours associated with consciousness might emerge from the idiosyncrasies of a mathematical model. The model we construct draws on the free energy principle and predictive coding. It suggests that the backbone of conscious experience can be construed in terms of prediction error minimization in a Bayesian hierarchical generative model of the world. By extending the model with extra machinery (temporal depth and counterfactual richness), we show that subjective properties of consciousness such as perceptual unity, selfhood and ineffability of qualia come to light.

Contents

Introduction

The problem of consciousness is one of the most profound and enigmatic questions in philosophy and science. Consciousness refers to the subjective qualities of experience, such as the way things look, sound, feel, and smell. Despite the tremendous advances made in our understanding of the brain, we still lack a comprehensive theory of how conscious experience arises from the activity of neurons. This problem is peculiar because it seems intuitively to be beyond the realm of science. Humans, regardless of their religious inclinations, seem to possess an indescribable hunch that the mind is separate from the material body. This hunch is not only ubiquitous across epochs and cultures, it occurs in young infants (Bloom 2004), before the prospect of influence by religion or philosophy.

It is often said that extraordinary problems require extraordinary solutions. Our approach towards consciousness relies as much on arguing that consciousness is not so extraordinary as it does on developing tools to explain it.

Frank Jackson, in his seminal article "Epiphenomenal Qualia" (Jackson 1982), presents a captivating thought experiment that has been the source of much debate in philosophy of mind. The experiment involves a scientist, Mary, who is confined to a room where she can only perceive the colours black and white. Despite this limitation, Mary studies colour vision and memorizes everything there is to know about it. Jackson then posits a hypothetical scenario in which Mary is released from the monochromatic room and exposed to the colourful world outside. The key question that arises is whether Mary, with her complete knowledge of colour vision, would learn anything new upon experiencing colours for the first time. Would Mary know exactly what to expect when the first sight of red hits her?

Most people reading this for the first time think that Mary *does* learn something when she sees colours. This, however, contradicts materialism: the view that nothing exists except matter (i.e. the default assumption of science). Predictably, materialists have contested this thought experiment. They often (justly) claim that, given the very strong assumptions, Mary would in fact *not* learn anything new (Dennett 1991). In this dissertation I will focus on a different problem altogether: why it is that we readily believe this argument in the first place?

The problem with *Mary's room* is that it places all the weight of resolving the paradox on an explanation of consciousness. This is perhaps why some theories of consciousness resort to provocative propositions or even overhauls of science. My

claim is that the mystery lies as much in the way we intuit our own consciousness as in any grand theory.

The present work proposes a rigorous, scientific approach to studying consciousness, referred to as the *illusionist strategy*, presented in Chapter 1. In Chapter 2, a plausible mathematical framework of cognition is constructed, drawing on the concepts of the free energy principle and predictive coding. In the final chapter, Chapter 3, we turn to showing that certain subjective properties of consciousness arise from the model. My main contribution lies in the methodology of the illusionist strategy, and the novel mathematical intricacy of the examples I present. It is noteworthy that the claim advanced in this dissertation does not rely on the model from chapter 2 being an accurate depiction of the brain. Rather, the main thrust of the argument is that, given a credible model of the brain, the subjective properties of consciousness may arise from it.

Chapter 1

A methodology for studying consciousness mathematically

1.1 The Hard problem and the Meta problem

In his seminal paper "Facing up to the problem of consciousness" (Chalmers 1995), David Chalmers hones in on the philosophical point which was being made by the *Mary's room* thought experiment. Chalmers proposes a distinction between what he calls the *easy* and *hard* problems of consciousness.

The easy problems of consciousness are the ones which can be readily tackled by cognitive science. They are concerned with functional properties of the human brain such as "the ability to discriminate, categorize, and react to environmental stimuli". The term easy here is meant to be understood in a metaphysical rather than a scientific sense. The idea is that, even if it takes centuries, it is easy to conceive of the success of the cognitive sciences in explaining these phenomena.

The hard problem of consciousness on the other hand, is the problem of understanding why the various activities of the brain are accompanied by experience. In Chalmers' words:

> even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience [...] there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?* (Chalmers 1995)

Put differently, solving the easy problem might not be enough to explain conscious experience. There may remain an "explanatory gap" (Levine 1999). Chalmers' argument is significant because, like Jackson's *Mary's room* argument, it would contradict materialism if true.

Twenty-three years after proposing the hard problem, David Chalmers laid out an updated framework for studying consciousness by publishing "The Meta-Problem of Consciousness" (Chalmers 2018). The first lines read:

> The meta-problem of consciousness is (to a first approximation) the problem of explaining why we think that there is a [hard] problem of consciousness.

The meta problem is a subset of the easy problem. However its judicious formulation provides a gateway into an alternative to the hard problem of consciousness: illusionism (Frankish 2016; Dewhurst and Dolega 2020).

1.2 Illusionism

Illusionism is an alternative to the hard problem. It is a subtle position which is often misunderstood. To explain it we begin with a quote from the American philosopher Joseph Levine (Levine 1999):

> The explanatory gap argument doesn't demonstrate a gap in nature, but a gap in our understanding of nature.

The "explanatory gap" Levine is referring to is between our subjective experience and objective scientific explanations. This quote emphasizes that consciousness is a subjective artefact. Illusionism goes further: it says that consciousness is *only* a subjective artefact.

Perhaps the easiest way to view illusionism is as the position that equates the hard problem and the meta problem of consciousness. It proposes that in order to explain consciousness, it is sufficient to explain why we think we are conscious, and why we believe consciousness to be the way we believe it to be.

How could consciousness be an illusion when it is the only thing one can be certain of? This is a common objection which misunderstands the use of the word "illusion" in illusionism. Illusionism (or at least our flavour of it) does not deny the existence of subjective experience. It simply denies that experience has any intrinsic phenomenal qualities, also known as qualia.

Consider the case of white light, which was previously regarded as an intrinsic property of nature until Newton's discovery that it is, in fact, composed of seven distinct colours. Clearly the primary focus of explanation should not be on the nature of "whiteness" itself, but rather on how the visual system processes light of

different wavelengths when it hits our retina. My claim is that consciousness is an "illusion" in a similar vein to white light. To assert that white light is illusory is not a denial of its existence, but rather an acknowledgment that our experience of it is misleading in terms of its objective existence in the world.

Regardless of whether one believes in it, illusionism also has the advantage of providing a (if not *the*) principled, scientific approach to consciousness. It is arguably the only approach which 1) recognises the ineffability and unity of subjective experience and 2) does not require a major revision of science to explain it. In fact illusionism can be viewed as a parsimonious explanation: if evolution needed humans to be imbued with a strong sense of experience, then *actually* fabricating phenomenal experience would be "mother nature doing things the hard way" (Frankish 2016).

Illusionism disambiguates the problem of consciousness and allows us to focus on a well-defined scientific question: why do we believe consciousness to be the way we believe it to be (i.e. the meta problem)? This is the problem we will attempt to tackle in this dissertation by proposing ways in which reports about consciousness could arise from a plausible model of the brain.

1.3 Heterophenomenology

Taking the illusionist position allows us to brush aside much of the philosophical tangles relating to consciousness. The question now is: how do we actually study consciousness in the real world?. Building on illusionism, Daniel Dennett proposes what he calls "heterophenomenology" as a scientific blueprint for the study of consciousness (Dennett 1991).

Heterophenomenology arises from an obvious problem when studying consciousness in humans: verbal reports about what we are conscious of are often wrong (e.g. optical illusions). On the other hand subjective reports clearly contain useful information. This is where heterophenomenology steps in. Instead of considering subjective reports as ultimate sources of truth, heterophenomenology treats them as fictions which carry valuable information[1]. This methodology is best understood through a simple example:

Agent A:	"I see a car."
Phenomenologist:	"Why does A see a car?"
Heterophenomenologist:	"Why does A say it sees a car?"

[1]There is a potent analogy to be made with the way anthropologists delve deep into various people's deities without casting any judgement on their existence.

At first sight heterophenomenology can seem too weak to make any headway into understanding consciousness. However it is important to bear in mind that the class of all subjective reports is a very rich class. A subjective report can encapsulate feelings, qualitative aspects of experience (qualia), counterfactual scenarios and deep introspection. The following report highlights the richness of heterophenomenology:

> As a red car drove past, I was struck by its vivid redness. Watching it speed down the road, I imagined myself behind the wheel, feeling the wind in my hair. At that point in time I felt that there was so much more to me than a simple machine.

Attempting to explain this report directly in terms of neuroscience is futile. Instead I propose to decompose such reports into a list of subjective properties of consciousness. This leads to a novel three part methodology to study consciousness: the *illusionist strategy*.

1.4 The Illusionist Strategy

The illusionist strategy is a novel three-part scheme that I propose as a way to study consciousness scientifically. The three steps are as follows:

1. Enumerate the subjective properties of consciousness.
2. Describe (to a sufficient extent) how the brain functions.
3. Explain why, given that the brain functions the way it does, the subjective properties of consciousness arise.

The emphasis here is on the word "subjective". The properties listed describe how consciousness seems to us, in line with the heterophenomenological approach. To explain such subjective properties of consciousness, it is therefore sufficient to explain why we perceive them, and why we perceive them the way we do.

In the present work, we will construct a plausible model of brain function. However the crux of our argument lies not in the accuracy of this model, but in showing that subjective properties of consciousness emerge from it.

1.5 Proposed subjective properties of consciousness

Here I put forward my best attempt at enumerating subjective properties of consciousness, heavily inspired by Jakob Hohwy (Hohwy 2022) but in a heterophenomenologist spirit. This list of properties may not be exhaustive or disjoint, but the properties enumerated are undeniably integral to the issue at hand.

Conscious perception and Qualia

Any account of consciousness must grant a central role to conscious perception: the awareness of sensory information that is processed by the brain, such as sights, sounds, smells etc. It involves the integration of sensory information with emotions, thoughts, feelings and past experiences. Note that conscious perception is also furnished with thoughts whose objects are thoughts themselves. We call this meta-cognition (i.e. cognition about cognition).

Conscious perceptions (and experiences at large) seem to have subjective and intrinsic qualities which we call qualia. They refer to the particular way we experience things which is unique to individuals and cannot be fully captured or communicated to others, we say that qualia are ineffable. From a heterophenomenologist point of view, we are not interested in qualia themselves but rather reports about them. Out of these reports, claims of ineffability are the most difficult and most crucial to explain. Later we will attempt to give a formal definition of ineffability and study how it might arise from our mechanistic model.

Perceptual Unity

Perception is unified into a single coherent scene. Our brain has the ability to perceive the world as a single and stable entity, despite the different and often conflicting signals that come from our senses. This phenomenon can be observed through reports of unity, and also through the suppression of sensory signals which do not fit into the unified scene.

United sense of self

Consciousness is anchored in a fist-person perspective, leading to an experience of selfhood. This is, like qualia, difficult for individuals to communicate, yet it is a phenomenon which we report and experience richly. Selfhood is influenced by an

individual's perception of their body and mental states (introspection), as well as their interactions with others and their social environment.

When explaining selfhood we must also account for the instances where selfhood is lost: so-called "selfless" experiences (Letheby and Gerrans 2017) for instance under a state of psychosis or under the influence of psychedelic drugs. We must also consider situations where selfhood is maintained but misattributed such as in the rubber hand illusion (Botvinick and Cohen 1998; Seth 2021).

Conscious flow

Conscious experience seems to flow continuously in what we often capture by the phrase "stream of consciousness". Interestingly reported sequences of events need not follow what subjects actually perceive. One prominent example of this is the Cutaneous Rabbit illusion (see figure 1.1 and Dennett 1991). In fact there is a rich body of work on the temporal structure of conscious experience dating back to William James (James 1890; Andersen and Grush 2009).

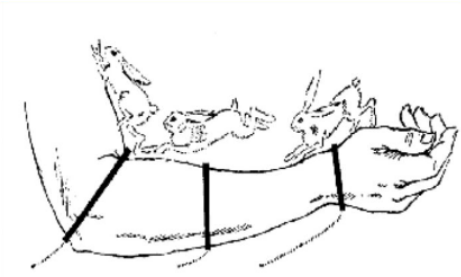

Figure 1.1: *The Cutaneous Rabbit Experiment*
Subjects are tapped 4 times at each of three regularly spaced locations on the arm but surprisingly report feeling a sequence of taps continuously moving up (like a rabbit). The thought-provoking element of this experiment is that subjects *retrospectively modify* their perception of the initial taps when they report. Indeed tapping just 4 times on the wrist results in an accurate report of static taps. However adding the subsequent taps at further locations leads to subjects reporting that *even the first taps* were already in motion!

Attention

Last but not least we must consider the relationship between attention and consciousness. Attention and consciousness are not the same thing: states of unconscious attention and, slightly more controversially, inattentive consciousness exist (Koch and Tsuchiya 2007). However it is also clear that they are closely linked: devoting one's

attention to a task has a strong bearing on what is (and isn't) conscious (Simons and Chabris 1999). This phenomenon is often referred to as inattentional blindness.

Other heuristics

Before diving into our proposed model, we make clear some heuristics which we will use throughout our argument.

- Firstly consciousness is the result of evolution through natural selection. When modelling the brain we must keep in mind that its ultimate objective is to ensure survival.
- Secondly consciousness is embodied. It is centred around an awareness of one's own body and is deeply tied to its state (hunger, pain etc.).
- Finally the brain is skull-bound. It is just an adaptive system receiving inputs. Even signals from its own body contain noise. This does not contradict the embodied aspect of consciousness. In fact it motivates it, because it mandates the brain to infer the state of its own body.

Chapter 2

A plausible model of cognition based on the free energy principle

The free energy principle (FEP) is a mathematical framework proposed by Karl Friston which describes the dynamics of physical systems (Friston 2009; Friston 2010). It is a theory of every "thing" (Friston 2019) in the sense that it provides a unified framework to study a range of systems (from a drop of oil to an ecosystem) and a range of functions (from perception to learning).

This treatment considers the FEP solely insofar as it applies to the brain. As a "unified theory of cognition", the FEP follows in the footsteps of Helmholtz and his proposal that perceptual inferences are extracted from sensory data through probabilistic modelling (Helmholtz et al. 1925). Although a first-principles account, the FEP has been extensively validated through experiment (Friston 2009) and is one of the most studied theories of the mind.

2.1 Free energy

The basic idea behind the free energy principle is that the driving force behind all biological organisms is the minimisation of a quantity known as free energy. Free energy can be thought of as a measure of an information-theoretic quantity called surprisal.

Definition 2.1.1 (Surprisal)**.** Given a random variable X with probability mass function $p_X(x)$ the surprisal of outcome x is

$$h_X(x) = -\log(p_X(x))$$

2.1.1 Why minimise surprisal?

Staying alive requires agents to maintain certain variables within bounds to maintain their internal balance (known as homeostasis). To do this, organisms use adaptive feedback loops, for example eating when hungry. Minimising surprisal can therefore be seen as preserving "expected" states which keep essential variables within the bounds of biological viability. However as the distribution of these states is unknown to agents, they are unable to directly measure surprisal and hence turn to a tractable quantity known as free energy (Christopher L. Buckley et al. 2017).

2.1.2 Deriving free energy from a variational Bayes scheme

To develop the theory rigorously, we need to define some variables which represent the world. Let $x \in \mathcal{X}$ characterise the (hidden) states of the environment, and $o \in \mathcal{O}$ denote sensory stimuli which arise as a result of these hidden states.

We make the assumption that complex states of the environment are not directly accessible to agents through sensory stimuli, but instead that the brain uses a process of Bayesian inference to reconstruct the world given sensory input. For example perception of "a bear standing in front of me" is the result of processing a set of stimuli (light hitting the retina) using prior knowledge.

We assume the brain encodes prior beliefs about the joint distribution between hidden states and sensory inputs in a **generative model**: $p(x, o) = p(x)p(o|x)$ which can be factorised into prior and likelihood. Given an observation $o = o_1$, we can calculate the posterior belief using Bayes' theorem:

$$p(x|o_1) = \frac{p(o_1|x)p(x)}{p(o = o_1)} = \frac{p(o_1|x)p(x)}{\int p(o_1|x)p(x)dx} \tag{2.1}$$

As is often the case in Bayesian inference, the integral in the denominator is intractable. To circumvent this we introduce a **variational density** (also called the recognition density) $q(x)$. We are now in a position to define free energy.

Definition 2.1.2 (Free Energy)**.** Given a generative model $p(x, o)$ and an variational posterior $q(x)$, free energy is defined as:

$$\mathcal{F} \equiv D_{KL}[q(x)||p(x|o)] - \log p(o)$$

Where D_{KL} denotes the Kullback-Leibler divergence:

$$D_{KL}[a(x)||b(x)] = \int a(x) \log \frac{a(x)}{b(x)} dx$$

This definition may seem somewhat arbitrary. As it turns out free energy possesses very desirable properties namely:

1. It can tractably be computed using densities which the brain has access to.
2. The second term $\log p(o)$ is independent of the variational posterior q, hence by minimising $\mathcal{F}$ with respect to q we are ensuring q approximates the true posterior.
3. As KL-divergence is always non-negative, we get the inequality $\mathcal{F} \geq -\log p(o)$. Hence $\mathcal{F}$ is an upper bound on surprisal, which gets tighter and tighter as we minimise $\mathcal{F}$.

2.1.3 Unpacking free energy

The formula for free energy given above arguably doesn't do it full justice. Free energy can be expressed in different ways leading to a stronger intuition about its potency. We start by deriving a formulation in terms of entropy and energy:

$$\begin{aligned}
\mathcal{F} &= \mathbf{D}_{KL}[q(x)||p(x|o)] - \log p(o) \\
&= \int q(x) \log \frac{q(x)}{p(x|o)} dx - \log p(o) \\
&= \int q(x) \log \frac{q(x)}{p(x,o)} dx \\
&= \underbrace{\mathbb{E}_{q(x)}[\log q(x)]}_{Entropy} - \underbrace{\mathbb{E}_{q(x)}[\log p(o,x)]}_{Energy}
\end{aligned} \tag{2.2}$$

From this we can derive yet another expression which gives further intuition.

$$\begin{aligned}
\mathcal{F} &= \int q(x) \log \frac{q(x)}{p(x,o)} dx \\
&= \int q(x) \log \frac{q(x)}{p(x)} - q(x) \log p(o|x) dx \\
&= \underbrace{\mathbf{D}_{KL}[q(x)||p(x)]}_{Complexity} - \underbrace{\mathbb{E}_{q(x)}[\log p(o|x)]}_{Accuracy}
\end{aligned}$$

Here the trade-off between complexity and accuracy illustrates the idea that a good explanation of a cause is accurate (fits the data) and simple (does not overfit the data).

2.1.4 Parameterising the variational posterior q

The free energy principle is based on a variational Bayes scheme. In order to implement this scheme the brain must explicitly represent the variational posterior $q(x)$. For our model we will assume a very simple form for the variational posterior, namely a Dirac delta distribution will all its mass at one point. This will enable us to derive a simplified form for the free energy.

Let $\mu \in \mathbb{R}^d$ and δ_μ be a measure defined on subsets of $\mathbb{R}$ such that $\delta_\mu(A) = 1 \iff 0 \in A$. We can thus formally define the integral of a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with respect to the measure δ_μ:

$$\int_{-\infty}^{\infty} f(x)\delta_\mu(dx) = f(\mu)$$

We can now define our variational posterior as a Dirac delta distribution centered at $\mu \in \mathbb{R}^d$. Going back to the entropy/energy formula for free energy, as the Dirac delta distribution is a point mass distribution its entropy is 0 (this can be shown rigorously by looking at a sequence of approximations of the Dirac delta function). Hence only the energy part of the formula remains:

$$\begin{aligned}\mathcal{F} &= -\mathbb{E}_{q(x)}[\log p(o, x)] \\ &= -\int \log p(o, x)\delta_\mu(dx) \\ &= -\log p(o, \mu)\end{aligned} \tag{2.3}$$

Importantly it means that free energy can now be expressed as a function of the sufficient statistic of the variational posterior, namely μ.

This simplification rests on a strong assumption: that $q(x)$ is a Dirac delta distribution. One considerable advantage of this is that it removes any integral or expectation from the expression, allowing us to consider the generative model p directly applied to the variational posterior's estimate μ, which parameterises beliefs.

2.2 Predictive Coding

The free energy principle provides an imperative which guides the brain. However the theory does not say anything about the form of the generative model p and the variational posterior q. These densities are where all the richness and complexity of human behaviour is encoded. Hence we need to make some assumptions about what they capture and what form they take.

In doing this we will make our way towards a variant of a prominent theory of brain function: predictive coding (Rao and Ballard 1999; Friston 2003; Friston 2005). Predictive coding (PC) can be derived from the free energy principle (Friston 2010) with some additional assumptions which we make clear in this section. Importantly, PC provides a stronger and more realistic account of cerebral operations. I posit that it may serve as a viable foundation to study consciousness (see chapter 3).

2.2.1 A hierarchical model

The expressiveness of models with a single level of latent variables is limited. The success of deep learning has demonstrated the capacity for hierarchical (i.e. deep) models to learn complex and abstract representations, the likes of which are necessary for humans to navigate the world. The free energy principle framework can be extended to multiple layers by decomposing the generative model $p(x)$ over a sequence of latent variables as follows:

$$p(x^0, ..., x^{(L)}) = p(x^{(L)}) \prod_{l=0}^{L-1} p(x^{(l)}|x^{(l+1)})$$

where we set the first variable to be the sensory input: $x^{(0)} = o$.

This hierarchy allows us to capture the richness of human cognition in a succinct way. We know from deep learning that, even if each component (namely $p(x_l|x_{l+1})$) remains simple, the combined model can be trained to learn highly complex functions. In fact the predictive coding framework is being used to train artificial neural networks in what is shaping to be an exciting new research area (Millidge et al. 2022).

2.2.2 Modelling a dynamic world

Until now we have assumed that the environment is static. Of course in reality the brain receives a constant flow of sensory data, and has to act in real time.

To deal with temporal data, we represent the world in terms of *generalized coordinates of motion*, an approach introduced in K. J. Friston, Stephan, et al. 2010. This means that, on top of representing hidden states and observations, we include all their time derivatives in our model. Thus given a parameterisation μ, we denote its generalized coordinate representation by the vector $\tilde{\mu} = [\mu, \mu_{[1]}, \mu_{[2]}, ...]$. We also define the operator $\mathcal{D}$ which maps each coordinate to its time derivative (B. Millidge and C. L. Buckley 2022) such that $\mathcal{D}[\mu, \mu_{[1]}, \mu_{[2]}, ...] = [\mu_{[1]}, \mu_{[2]}, \mu_{[3]}, ...]$.

We are now in a position to produce a dynamical generative model. We model sensory input o as a function of hidden states x and some Gaussian noise w_o. We

use a Langevin-type equation for the dynamics of hidden states themselves. These two equations can be differentiated. Here we make what is called the *local linearity* assumption for x and o so that we only end up with linear terms in the derivative computations:

$$\begin{aligned} o &= f(x) + \omega_o & x' &= g(x) + \omega_x \\ o' &= f(x)x' + \omega_o' & x'' &= g(x)x' + \omega_x' \\ o'' &= f(x)x'' + \omega_o'' & x''' &= g(x)x'' + \omega_x'' \\ &\dots & &\dots \end{aligned}$$

Or expressed more succinctly with our notation:

$$\tilde{o} = \tilde{f}(\tilde{x}) + \tilde{\omega}_o \qquad \mathcal{D}\tilde{x} = \tilde{g}(\tilde{x}) + \tilde{\omega}_x \tag{2.4}$$

where $\tilde{f}$ and $\tilde{g}$ are functions over generalized coordinates.

We make another assumption, namely that the noise is independent between coordinates. Whilst a strong assumption, this enables us to decompose the prior over hidden states as

$$p(\tilde{x}) = p(x, x_{[1]}, x_{[2]}, \dots) = p(x) \prod_{n=0}^{\infty} p(x_{[n+1]}|x_{[n]})$$

We can also write the likelihood of sensory data conditional on hidden states:

$$\begin{aligned} p(\tilde{o}|\tilde{x}) &= p(o, o_{[1]}, \dots | x, x_{[1]}, \dots) \\ &= \prod_{n=0}^{\infty} p(o_{[n]}|x_{[n]}) \end{aligned}$$

Hence we can rewrite the whole generative model as

$$p(\tilde{x}, \tilde{o}) = p(x) \prod_{n=0}^{\infty} p(o_{[n]}|x_{[n]}) p(x_{[n+1]}|x_{[n]})$$

2.2.3 Combining the hierarchical and dynamical models

We have made two assumptions, that the brain is a hierarchical system and that it models its dynamic environment by using generalized coordinates. By combining these two insights we arrive at the following decomposed form for the generative

model:

$$\begin{aligned} p(\tilde{x}, \tilde{o}) &= p(\tilde{o}, \tilde{x}^{(1)}, \tilde{x}^{(2)}, ...\tilde{x}^{(L)}) \\ &= p(\tilde{x}^{(L)}) \prod_{l=0}^{L-1} p(\tilde{x}^{(l)} | \tilde{x}^{(l+1)}) \\ &= p(x^{(L)}) \prod_{n=0}^{\infty} p(x^{(L)}_{[n+1]} | x^{(L)}_{[n]}) \prod_{l=0}^{L-1} \prod_{n=0}^{\infty} p(x^{(l)}_{[n]} | x^{(l+1)}_{[n]}) \end{aligned} \tag{2.5}$$

where in the final line we assume independence among hierarchical levels at different dynamical orders.

2.2.4 The generative model

So far we have refrained from assigning any particular distribution to the generative model $p(x, o)$. Here we will commit to a particular distribution. Note that the claim is not that this is exactly how the brain operates. The aim is simply to produce a template upon which to study consciousness.

From 2.5, the only densities we need to define are that of the final layer prior $p(\tilde{x}^{(L)})$ and the intra-layer likelihood $p(x^{(l)}_{[n]} | x^{(l+1)}_{[n]})$.

For the intra-layer likelihood, we stick to convention (Christopher L. Buckley et al. 2017) in defining it as a Gaussian with mean parameterised by a function $f_{l,n}$ and covariance matrix $\mathbf{\Sigma}_{l,n}$:

$$p(x^{(l)}_{[n]} | x^{(l+1)}_{[n]}) = \frac{\exp\Big(-\frac{1}{2}(x^{(l)}_{[n]} - f_{l,n}(x^{(l+1)}_{[n]}))^T \mathbf{\Sigma}^{-1}_{l,n}(x^{(l)}_{[n]} - f_{l,n}(x^{(l+1)}_{[n]}))\Big)}{(2\pi)^{d/2} |\mathbf{\Sigma}_{l,n}|^{1/2}} \tag{2.6}$$

where d denotes the dimension of the vector representation of x and $|\mathbf{\Sigma}|$ denotes the determinant of the matrix $\mathbf{\Sigma}$.

For the final layer prior we have already decomposed it into a product:

$$p(\tilde{x}^{(L)}) = p(x^{(L)}) \prod_{n=0}^{\infty} p(x^{(L)}_{[n+1]} | x^{(L)}_{[n]})$$

In line with Christopher L. Buckley et al. 2017, we assume that the noise $\tilde{\omega}_x$ in 2.4 is independent in each generalized coordinate and define the distribution of order $n+1$ conditional on order n as a Gaussian with mean parameterised by a function g_n and covariance matrix $\mathbf{\Theta}_n$:

$$p(x^{(L)}_{[n+1]} | x^{(L)}_{[n]}) = \frac{\exp\Big(-\frac{1}{2}(x^{(L)}_{[n]} - g_n(x^{(L)}_{[n+1]}))^T \mathbf{\Theta}^{-1}_{n}(x^{(L)}_{[n]} - g_n(x^{(L)}_{[n+1]}))\Big)}{(2\pi)^{d/2} |\mathbf{\Theta}_{n}|^{1/2}} \tag{2.7}$$

For the prior on the 0^{th} coordinate of motion (i.e. the state itself), we break with common practice and argue that a Gaussian is not suitable. Instead we use a multivariate Laplace distribution with independent components[1]. The contents of the final layer $x^{(L)}$ can be seen as encoding hypotheses about the world, because they condition all lower-level representations in a top-down way (more on this in 2.5). In this sense our model approximates Bayesian inference with a hierarchy of beliefs. For this reason I argue that each component of the topmost layer should be sparse, because only a small subset of all possible states of the world are plausible at any given time.

Imposing a Laplace distribution over the final layer can be viewed as a form of regularization. Indeed it has been shown that only a small proportion of neurons are active at any given time (Vinje and Gallant 2000). For these reasons I propose the following prior:

$$p(x^{(L)}) = \prod_{i=1}^{d} \frac{1}{2b_i} \exp\left(-\frac{(x_i^{(L)})^T x_i^{(L)}}{b_i}\right) \tag{2.8}$$

where $x_i^{(L)}$ represents the i-th component of the vector $x^{(L)}$.

The idea is that by optimizing the parameters $\{b_i\}_{0\leq i\leq d}$, it is possible to learn the fatness of the tails and the tightness of the peak around 0 for each component. The agent can thus learn to encourage sparsity in a flexible way.

2.3 Minimising free energy

By making assumptions, we have been able to decompose and simplify the generative model. Now we are in a position to write down an explicit formula for the free energy, and investigate how it can be minimised.

Going back to our approximation of free energy in 2.3 in terms of beliefs and sensory input:

$$\mathcal{F} = -\log p(\mu, o)$$

Substituting our expression for the generative model into this we get:

$$\begin{aligned} -\mathcal{F} &= \log\left[p(\mu^{(L)}) \prod_{n=0}^{\infty} p(\mu_{[n+1]}^{(L)} | \mu_{[n]}^{(L)}) \prod_{l=0}^{L-1} \prod_{n=0}^{\infty} p(\mu_{[n]}^{(l)} | \mu_{[n]}^{(l+1)})\right] \\ &= \log p(\mu^{(L)}) + \sum_{n=0}^{\infty} \log p(\mu_{[n+1]}^{(L)} | \mu_{[n]}^{(L)}) + \sum_{l=0}^{L-1} \sum_{n=0}^{\infty} \log p(\mu_{[n]}^{(l)} | \mu_{[n]}^{(l+1)}) \end{aligned} \tag{2.9}$$

[1]Note that using a prior with independent components does not necessarily mean that the components of the posterior will also be independent.

Now substituting the distributions defined in 2.8, 2.7 and 2.6 we finally get an explicit formula for free energy in terms of internal parameters:

$$
\begin{aligned}
-\mathcal{F} = &\sum_{i=1}^{d} \frac{(\mu_i^{(L)})^T \mu_i^{(L)}}{b_i} + \log(2b_i) \\
&+ \sum_{n=0}^{\infty} \frac{1}{2}(\mu_{[n]}^{(L)} - g_n(\mu_{[n+1]}^{(L)}))^T \mathbf{\Theta}_n^{-1}(\mu_{[n]}^{(L)} - g_n(\mu_{[n+1]}^{(L)})) + \log\Big((2\pi)^{d/2}|\mathbf{\Theta}_n|^{1/2}\Big) \\
&+ \sum_{l=0}^{L-1}\sum_{n=0}^{\infty} \frac{1}{2}(\mu_{[n]}^{(l)} - f_{l,n}(\mu_{[n]}^{(l+1)}))^T \mathbf{\Sigma}_{l,n}^{-1}(\mu_{[n]}^{(l)} - f_{l,n}(\mu_{[n]}^{(l+1)})) + \log\Big((2\pi)^{d/2}|\mathbf{\Sigma}_{l,n}|^{1/2}\Big)
\end{aligned}
$$

According to the free energy principle, this is the quantity which the brain minimises by modifying its internal variables.

2.3.1 How does the brain minimise free energy

The internal variables which the brain can tweak can be separated into two broad categories according to the time-scales of their updates. Firstly the beliefs about hidden states:

$$\{\tilde{\mu}^{(l)}\}_{1\leq l\leq L}$$

Recall that these variables originally come from the variational posterior $q(x;\mu) = \delta_\mu(x)$ which we defined as a Dirac delta function. These beliefs about hidden states are updated on a fast time-scale and we will argue that they can be identified with perception. Secondly, the model contains other variables which we call parameters (see 2.3.1). These parameters are updated on slower time-scales. They enable the brain to learn accurate representations of the world.

Parameter	Definition
$\{b_i\}_{0\leq i\leq d}$	Scales of components in the topmost layer prior.
$\{g_n\}_{n\in\mathbb{N}}$	Parameterisation of the final layer density mean.
$\{\mathbf{\Theta}_n\}_{n\in\mathbb{N}}$	Final layer density covariance.
$\{f_{l,n}\}_{0\leq l\leq L, n\in\mathbb{N}}$	Parameterisation of the layer-to-layer density.
$\{\mathbf{\Sigma}_{l,n}\}_{0\leq l\leq L, n\in\mathbb{N}}$	Parameterisation of the layer-to-layer density covariance.

Table 2.1: List of parameters.

Under the free energy principle, it is suggested that the brain updates its variables and parameters through an implementation of *gradient descent* (Christopher L.

Buckley et al. 2017; Karl Friston and S. Kiebel 2009; Bogacz 2017). For example for beliefs we get the dynamics

$$\frac{d\tilde{\mu}}{dt} = -\frac{\partial \mathcal{F}}{\partial \tilde{\mu}}$$

where analogous equations exist for all other parameters.

Here we derive explicit formulas for the gradient of free energy with respect to beliefs $\mu^{(l)}_{[n]}$ at different layers and different dynamical orders. To my knowledge this is novel in the sense that this scheme uses a hierarchical model, generalized coordinates and does not assume independence between the dimensions of the vector representations. Our calculations use, amongst others, the rules stated clearly in 2.2.

For clarity we separate different cases based on the parameters l and n, starting with the topmost layer states $\mu^{(L)} = \mu^{(L)}_{[0]}$ (i.e. when $l = L$ and $n = 0$).

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \mu^{(L)}} &= 2\left[\frac{\mu^{(L)}_1}{b_1}, \frac{\mu^{(L)}_2}{b_2}, ..., \frac{\mu^{(L)}_d}{b_d}\right]^T \\ &+ \mathbf{\Theta}_0^{-1}(\mu^{(L)} - g_0(\mu^{(L)}_{[1]})) \\ &+ \left(\frac{\partial f_{L-1,0}}{\partial \mu^{(L)}}(\mu^{(L)})\right)^T \mathbf{\Sigma}^{-1}_{L-1,0}(\mu^{(L-1)} - f_{L-1,0}(\mu^{(L)}))\end{aligned}$$

Where the first line is the derivative of the component-wise Laplace prior.

Next we look at $\mu^{(L)}_{[n]}$ where $n > 0$, i.e. the topmost layer but this time over dynamical orders greater than 1.

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \mu^{(L)}_{[n]}} &= \mathbf{\Theta}_n^{-1}(\mu^{(L)}_{[n]} - g_n(\mu^{(L)}_{[n+1]})) \\ &+ \left(\frac{\partial g_{n-1}}{\partial \mu^{(L)}_{[n]}}(\mu^{(L)}_{[n]})\right)^T \mathbf{\Theta}^{-1}_{n-1}(\mu^{(L)}_{[n-1]} - g_{n-1}(\mu^{(L)}_{[n]})) \\ &+ \left(\frac{\partial f_{L-1,n}}{\partial \mu^{(L)}_{[n]}}(\mu^{(L)}_{[n]})\right)^T \mathbf{\Sigma}^{-1}_{L-1,n}(\mu^{(L-1)}_{[n]} - f_{L-1,n}(\mu^{(L)}_{[n]}))\end{aligned}$$

Variables	**Rule**
$\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{A}^T = \mathbf{A}$.	$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$
$z = f(\mathbf{y})$ where $f : \mathbb{R}^d \mapsto \mathbb{R}$, and $\mathbf{y} = g(\mathbf{x})$ where $g : \mathbb{R}^d \mapsto \mathbb{R}^d$.	$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T \frac{\partial z}{\partial \mathbf{y}}$

Table 2.2: Useful rules for differentiation in multiple variables.

Finally we derive the gradient of free energy with respect to beliefs from any other layer ($1 \leq l \leq L-1$), over generalized coordinates of motion.

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \mu_{[n]}^{(l)}} &= \boldsymbol{\Sigma}_{l,n}^{-1}(\mu_{[n]}^{(l)} - f_{l,n}(\mu_{[n]}^{(l+1)})) \\ &+ \Big(\frac{\partial f_{l-1,n}}{\partial \mu_{[n]}^{(l)}}(\mu_{[n]}^{(l)})\Big)^T \boldsymbol{\Sigma}_{l-1,n}^{-1}(\mu_{[n]}^{(l-1)} - f_{l-1,n}(\mu_{[n]}^{(l)}))\end{aligned}$$

These equations show that the gradients all depend only on local information (i.e. from adjacent values of l or n). This means that we could implement the model with simple components each of which performs local computations. Indeed such implementations for similar models have shown promising results in machine learning tasks (Millidge et al. 2022). Moreover, local computation is also compatible with the way the brain functions.

2.4 Active inference

In actual fact, there are two ways for the brain to minimise free energy. The first, which we have already surveyed, is to update beliefs about the world to explain lower-level prediction errors (perception). The second is to modify the sensory samples to make them more like expectations, in other words: action. Active inference refers to a framework which views perception and action as complementary processes that minimise free energy.

It may appear that modifying sensory input to match expectations is akin to engaging in self-fulfilling prophecies. In some sense this is the case. However note that the world at large does not always cooperate with our expectations (J. Hohwy 2013) leading to high prediction errors and the need to revise hypotheses.

Action does not appear directly in the free energy formulation. Instead it is considered through its effect on sensory data. We rewrite our model as $\tilde{o} = \tilde{f}(\tilde{x}, a) + \omega_o$ where a is a vector representing an action. This relationship allows us to write down the gradient descent scheme for action:

$$\frac{da}{dt} = -\frac{d\tilde{o}}{da}\frac{\partial \mathcal{F}}{\partial \tilde{o}}$$

While action influences sensory input, it is also the result of the brain's processing. For this reason we write them as functions of internal beliefs: $a = \mathcal{A}(\{\tilde{\mu}^{(l)}\}_{1 \leq l \leq L})$. By letting actions depend on the whole hierarchy of beliefs, we can account for low-level

actions (e.g. reflexes) as well as high-level cognition (e.g. speech, reasoning etc.). It is noteworthy that verbal reports, crucial to the heterophenomenological approach, depend on the hypotheses represented in the topmost layer beliefs. We will revisit this formulation in 3.2. An illustration of the active inference loop is given in figure 2.1.

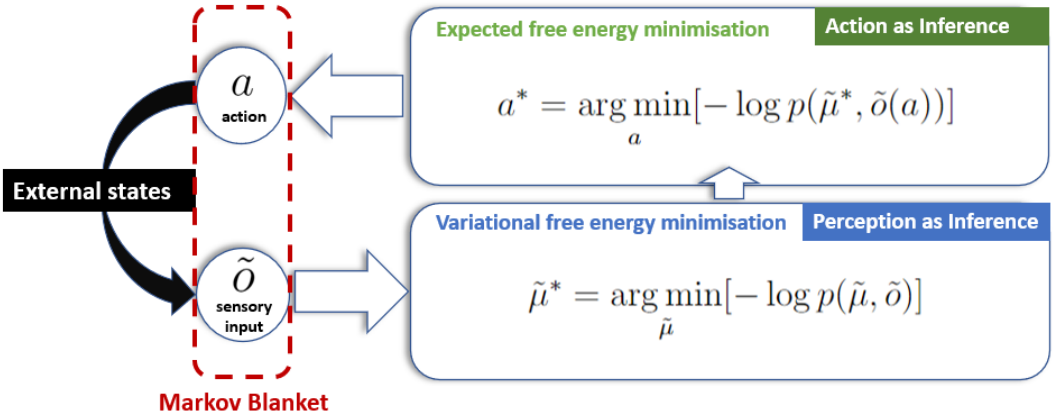

Figure 2.1: Diagram illustrating the active inference loop. By minimising free energy, the agent optimises beliefs and subsequently selects an action. Note that the agent's interaction with the environment is done through what is known in the FEP literature as a *Markov blanket* (M. J. D. Ramstead et al. 2023). A Markov blanket for agent A is a subset of states conditional upon which A's states are independent of external states.

2.5 The model in action

In this section I review the model devised and run through an example of how it might operate in the real world. Figure 2.2 gives an overview of the dynamics of the model.

Note that there are two methods of message-passing within the hierarchy: top-down predictions and bottom-up errors. Hence each layer in our model can be seen as making predictions about the previous layer, and then receiving back the error of that prediction. Perception happens by optimizing beliefs $\mu^{(l)}$ in parallel across the hierarchy.

To gain an understanding of the capabilities of our model, it is useful to explore a specific example. Let us consider an agent, whom we will refer to as agent A, walking down the street and catching sight of what appears to be a tall man wearing

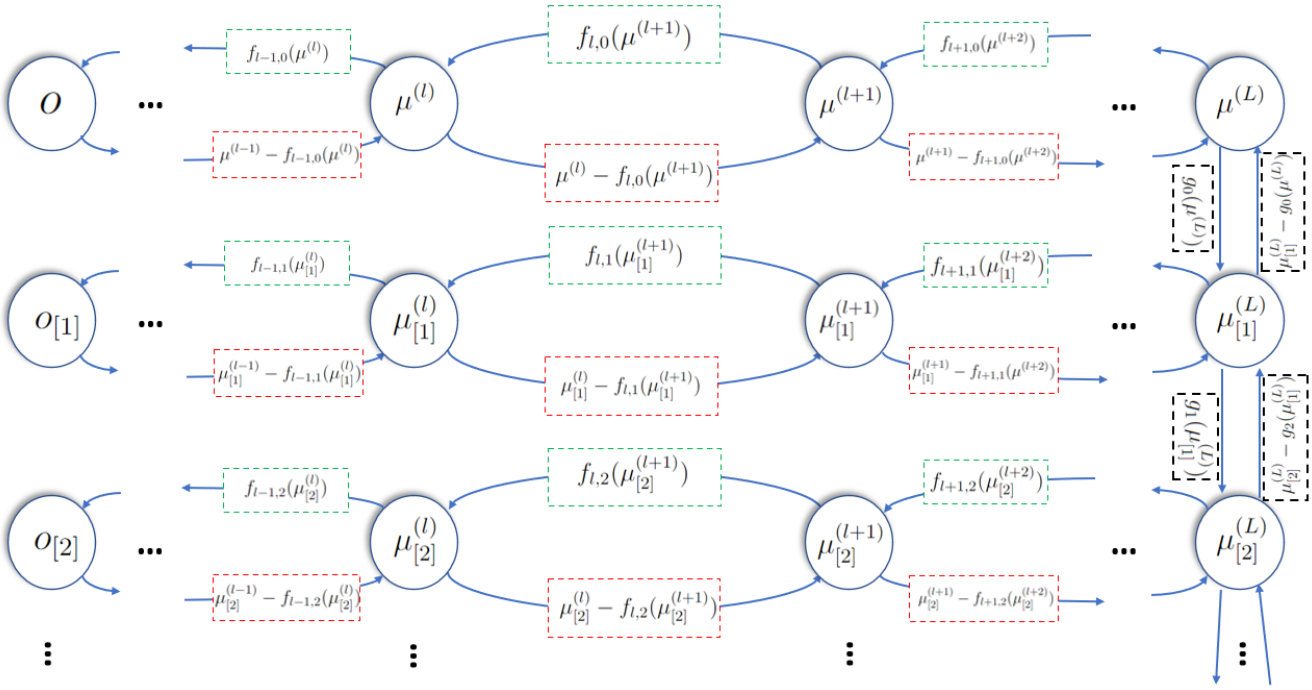

Figure 2.2: Diagram of interaction between components of the hierarchical model. Top-down predictions appear in green boxes, bottom-up error propagation in red. Note that, contrary to all other levels of beliefs $\{\tilde{\mu}^{(l)}\}_{1\leq l\leq L}$, the 0^{th} layer $\tilde{o}$ is not updated through gradient descent but rather is fixed to sensory input.

a dark coat in the distance. Our model of agent A's brain includes a final layer $\mu^{(L)}$, which represents a hypothesis H_{man} that "there is a tall man wearing a dark coat in the distance". If agent A were to be probed, this is what it would report. This hypothesis in turn acts top-down on the whole hierarchy through its predictions on the previous layer. It could for example mandate lower-level representations to infer different aspects of the man, his coat and perhaps even his behaviour (e.g. the man is hunched forward therefore he might be looking at his phone).

Now suppose that what agent A thinks is a man is actually not a man but a dark bear standing in the middle of the street. Agent A's top-down hypothesis H_{man} initially prescribes lower-level representations consistent with the sight of a human being. However as agent A gets closer to the bear, its beliefs $\{\tilde{\mu}^{(l)}\}_{0\leq l\leq L}$ become increasingly out of tune with predictions received from higher levels. In precise terms, the prediction errors $\epsilon_{l,n} = \mu^{(l-1)}_{[n]} - f_{l-1,n}(\mu^{(l)}_{[n]})$ start to increase. Once agent A is close enough to the bear, the bottom-up process of error propagation sounds the alarm: the hypothesis H_{man} has to be revised in the topmost layer. All of a sudden agent

A seamlessly shifts to a new hypothesis H_{bear} which in turn acts on lower-levels, and agent A starts to subjectively experience the animal's "bear-ness".

Here arises a crucial question. In this scenario, A's reports would undoubtedly express surprise, a range of emotions, and most significantly, accounts of the qualia associated with seeing the bear. If probed, Agent A would most likely describe a sense of "ineffability" of its private experiences throughout this short anecdote. The question is, how does our model explain such reports of ineffable subjective experience?

The next chapter is devoted to answering this and other related questions, both in the context of this anecdote and beyond.

Chapter 3

Emergence of the subjective properties of consciousness: perceptual unity, selfhood and the ineffability of qualia

The final step of the illusionist strategy is to take our model of cognition, and show that subjective properties of consciousness arise from it. The present work focuses on three of the five properties listed, namely **perceptual unity** in 3.1, **selfhood** in 3.2 and **qualia** in 3.3. Keeping with the heterophenomenologist approach, we investigate these through verbal reports, which are viewed as a type of action within active inference.

3.1 Perceptual Unity

Perceptual unity refers to the phenomenon of experiencing the world as a single, coherent entity. Humans report a single stream of consciousness and often neglect sensory input which doesn't fit into a unified scene.

3.1.1 A single hypothesis

The core aim of the brain is to control the body in order to stay alive. Crucially the brain controls only *one* body, and that body is heavily constrained by the laws of physics. For this reason the brain must act in a unified way. This echoes the claim that the brain approximates Bayesian inference, which is unified by definition. In the words of Hohwy (J. Hohwy 2013):

> Put simply, we can only do one thing at a time and this thing is prescribed by our singular hypothesis about what we are doing. This necessarily entails predictions that are internally unified throughout the hierarchy - in a global sense.

This yields a handy explanation of the phenomenon of binocular rivalry (Blake and Logothetis 2002). Despite alternating, perceptual unity is maintained at any given point in time by a single hypothesis acting top-down on the perceptual hierarchy.

3.1.2 An example of perceptual binding

How do reports of unity arise in our model? Suppose that given sensory input $\tilde{o}_{scene}$, our brain minimises free energy and produces the action a_{report} of reporting seeing an integrated scene containing a red square and a green ball.

Presumably the processing of the colours red and green are done in different parts of the brain, as are the processing of squares and balls. Why then do we experience a red square and a green ball rather than say a red ball and a green square or four totally separate entities? This is known as the binding problem (J. Hohwy 2013) and is a subset of the perceptual unity problem.

Here is a sketch of how things might play out in our model: Within the visual scene represented by $\tilde{o}_{scene}$, the first few levels $1, ..., l_1$ of the hierarchical model extract attributes red, green, square, ball as well as some information about the way these attributes spatially co-occur (think of the way a Convolutional Neural Network extracts low-level features and recombines them). Further levels $l_1 + 1, ..., l_2$ extract more abstract features so that $\tilde{\mu}^{(l_2)}$ represents the bindings "red square" and "green ball" as latent variables. As we move further up the hierarchy, more and more attributes are bound together. For example the sound of bouncing could be bound to the ball.

It should not be expected that this behaviour arise from *any* implementation of our model. The binding of the ball with the colour green originates from evolutionary heuristics. I also postulate that these unity heuristics are in practice enforced because top-down predictions $\tilde{f}_l(\tilde{\mu}^{(l+1)})$ contain a bias towards binding. This bias is a result of the sparsity of the topmost layer enforced by its prior distribution $\tilde{\mu}_i^{(L)} \sim \text{Laplace}(0, b_i)$ (regularisation can lead to bias).

To be concise, the brain enforces perceptual unity because it needs to act in a unified way. In our model perceptual unity is enforced through hierarchical binding of attributes into higher-level latent variables.

3.2 Sense of self

Humans experience a strong sense of self, as observed by reports of feelings of mineness, agency, and generally attributing special status to one's body and feelings (J. Limanowski and Blankenburg 2013).

3.2.1 Selfhood as a hypothesis

In our model, we view selfhood as nothing more than a hypothesis. This hypothesis emerges as the most accurate and parsimonious explanation of patterns of causation by inferring that the system is itself a cause of its sensory input (Limanowski and Friston 2020; Hohwy and Michael 2017; Deane 2021).

Why would our model come up with such a hypothesis? Recall that actions are generated through what is effectively an inverse model of how they affect sensory input.

$$a^* = \arg\min_a [-\log p(\tilde{\mu}, \tilde{o}(a))]$$

It follows that the model *has* to account for its own actions, because these actions themselves act on sensory input. The whole system can be seen to model its own actions, with each layer responsible for modelling the previous one. As a consequence the brain comes to grant privileged status to artefacts which it deems to have control over (limbs, muscles, thoughts etc.).

Is modelling of one's actions enough to give rise to selfhood? No. We argue that a certain kind of hierarchical modelling is necessary, namely one that is temporally deep and entertains counterfactual scenarios.

When generating new actions, it has been argued that the human brain entertains multiple counterfactual scenarios into the future (Seth 2021, Hohwy 2022). It follows that the brain must model these prospective contexts[1]. Not only that, as an inferred proximal cause of its sensory input, the system must also model *itself* into the future[2]. The skull-bound brain effectively treats its future states as part of the hidden states of the world. The agent's beliefs $\{\tilde{\mu}^{(l)}\}_{1 \leq l \leq L}$ effectively become *beliefs about beliefs*. To formalise this, we extend our hierarchical model with a scheme which allows explicit modelling of the future.

[1] Here we borrow from cybernetics where the Good Regulator Theorem (Conant and Ashby 1970) states that any good regulator of a system must be a good model of that system.

[2] We assume that the brain only models these recursive dynamics down to a certain depth. This avoids an infinite regress where to model its actions the system must model itself and to model itself it must model its actions.

3.2.2 An extended model with temporal depth and counterfactual richness

Here we extend our model with a scheme which endows it with deep temporal structure, and the ability to consider counterfactual scenarios.

We start by postulating that each layer of beliefs $\tilde{\mu}^{(l)}$ can be partitioned into three vectors $(\tilde{s}^{(l)}, \tilde{\alpha}^{(l)}, \tilde{\phi}^{(l)})$ where $\tilde{s}$ represents hidden states, $\tilde{\alpha}$ represents active states and $\tilde{\phi}$ represents observation states (K. J. Friston, Rosch, et al. 2018). This partition is defined precisely in terms of how the dependencies can be decomposed (see 3.1). To account for modelling deep into the future, we shift the focus of active inference to choosing policies $\pi \in \prod$ which are none other than sequences of actions $(a_1, a_2, ...a_T)$. Active states are precisely the beliefs upon which policy selection depends, allowing us to rewrite 2.4:

$$\pi = \mathcal{A}(\{\tilde{\alpha}^{(l)}\}_{1 \leq l \leq L})$$

This grants the possibility of both actions from low-levels of the hierarchy (e.g. reflexes) to higher-level planning (e.g. "I infer that I am an agent that likes chocolate, therefore I will seek chocolate").

We can also be more specific about the inter-layer interactions for active states,

$$p(\tilde{\alpha}^{(l)}|\tilde{\mu}^{(l+1)}) = p(\tilde{\alpha}^{(l)}|\tilde{s}^{(l+1)}) \tag{3.1}$$

hidden states,

$$p(\tilde{s}^{(l)}|\tilde{\mu}^{(l+1)}) = p(\tilde{s}^{(l)}|\tilde{\alpha}^{(l)}, \tilde{s}^{(l+1)}) \tag{3.2}$$

and observation states.

$$p(\tilde{\phi}^{(l)}|\tilde{\mu}^{(l+1)}) = p(\tilde{\phi}^{(l)}|\tilde{s}^{(l)}) p(\tilde{s}^{(l)}|\tilde{\mu}^{(l+1)}) \tag{3.3}$$

To get temporal depth, we decompose beliefs $\tilde{s}$, $\tilde{\phi}$ and $\tilde{\alpha}$ into different time-steps $(t, t+1, ..., T)$. We can further decompose the dynamics from 3.2 into

$$p(\tilde{s}^{(l)}|\tilde{\alpha}^{(l)}, \tilde{s}^{(l+1)}) = p(\tilde{s}_1^{(l)}|\tilde{s}^{(l+1)}) \prod_t^T p(\tilde{s}_{t+1}^{(l)}|\tilde{s}_t^{(l)}, \alpha_t^{(l)}) \tag{3.4}$$

those from 3.1,

$$p(\tilde{\alpha}^{(l)}|\tilde{s}^{(l+1)}) = p(\tilde{\alpha}_1^{(l)}|\tilde{s}_1^{(l+1)}) \prod_t^T p(\tilde{\alpha}_{t+1}^{(l)}|\tilde{\alpha}_t^{(l)}, \tilde{s}_t^{(l+1)}) \tag{3.5}$$

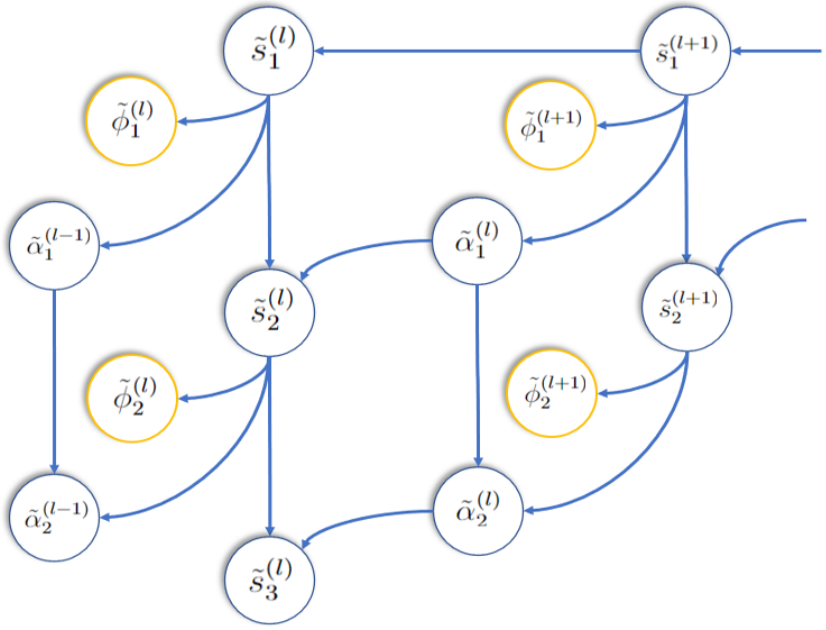

Figure 3.1: Bayesian graph depicting conditional dependencies of hidden states, active states and observation states through (counterfactual) time, between two layers of the hierarchy.

as well as the dynamics from 3.3:

$$p(\tilde{\phi}^{(l)}|\tilde{s}^{(l)}) = \prod_t^T p(\tilde{\phi}_t^{(l)}|\tilde{s}_t^{(l)}) \tag{3.6}$$

Hence the model has a way to model future hidden states and observations. We make a couple of points to emphasize the richness of this model. Firstly note that the distribution $p(\tilde{\phi}^{(l)}|\tilde{s}^{(l)})$ can be multimodal. This enables elaborate hierarchical modelling of counterfactual scenarios. Secondly, within these counterfactual scenarios, the system is mandated to model *itself* as a proximal cause of these counterfactual observations.

We also show that, under this scheme, our system can be viewed as minimising expected free energy. Recall that the expression for free energy is:

$$\mathcal{F} = \mathbb{E}_{q(\tilde{x}|\tilde{o})}[\log q(\tilde{x}|\tilde{o}) - \log p(\tilde{o}, \tilde{x})]^3$$

[3]Here we have written $q(\tilde{x}|\tilde{o})$ instead of $q(\tilde{x})$ previously. Even though the variational posterior does not depend on $\tilde{o}$ directly, it results from a minimisation which depends on $\tilde{o}$ so we make the dependence explicit.

In order to sample counterfactual observations under a policy the agent must represent a distribution $q(\tilde{o}|\pi)$. This is a distribution over counterfactual observations $o \in \mathcal{O}_{\text{counterfactual}}$. We assume that this set of events is represented by the observation states $\tilde{\phi}^{(l)}$.

We can now take the expectation of $\mathcal{F}$ with respect to this counterfactual distribution:

$$\mathbb{E}_{q(\tilde{o}|\pi)}[\mathcal{F}] = \mathbb{E}_{q(\tilde{o}|\pi)}[\mathbb{E}_{q(\tilde{x}|\tilde{o})}[\log q(\tilde{x}|\tilde{o})] - \log p(\tilde{o}, \tilde{x})]] \tag{3.7}$$

$$= \mathbb{E}_{q(\tilde{o},\tilde{x}|\pi)}[\log q(\tilde{x}|\tilde{o})] - \log p(\tilde{o}, \tilde{x})] \tag{3.8}$$

This quantity is called expected free energy and is generally used as the objective function under which actions are chosen (Friston 2019; Karl Friston, Fitzgerald, et al. 2017).

In this treatment we refrain from attaching any specific distributions. This is partly for conciseness but more importantly because this "extension" of our model should perhaps be viewed as an idealised example of how the brain might implement temporal depth and counterfactual richness. It is likely that there are many other ways of doing so. This goes back to what the goal of this dissertation is: not to perfectly model the brain, but to show how a plausible model of the brain would exhibit consciousness-reporting behaviour.

3.2.3 Emergence of the self

The argument is as follows. The brain infers itself as a proximal cause for much of its sensory input. With this hypothesis, it mandates its entire hierarchy of beliefs to model the system *itself*. This process of modelling involves temporal depth and considers counterfactual scenarios. The system thus has to gauge itself throughout this rich set of counterfactual scenarios into the future. On top of that some of the latent variables which are inferred (such as emotions, thoughts and feelings) are not projected onto any physical medium. This leads to reports of a vivid and integrated sense of self coupled with a strong "sense of being".

3.2.4 An example of inferring selfhood

Now that we have proposed an explanation for *why* selfhood arises, we can look at an example of *how* it might arise, specifically in a context where selfhood is "tricked": the rubber hand experiment (Botvinick and Cohen 1998).

Consider agent A as the subject of the rubber hand experiment. Agent A begins the experiment with a hierarchical set of beliefs $\{\tilde{\mu}^{(l)}\}_{1\leq l\leq L}$. Within these beliefs, the final layer can be seen as encoding a hypothesis $H_{rubber\ hand}$ which acts top-down to direct perception of the rubber hand as an exogenous object. Next, the experimenter produces concurrent stimuli on both the rubber hand and the participant's real hand. Even though A's prior beliefs captured by $H_{rubber\ hand}$ are strong, the temporal overlap between the seen and felt touch is so unlikely under the current hypothesis that the bottom-up errors overthrow $H_{rubber\ hand}$ in favour of a new hypothesis $H_{my\ hand}$.

All of a sudden, this rubber hand is inferred to be a part of the body, and a cause of sensory input. Assuming (for sake of argument) that the illusion is perfect, counterfactual modelling of the future will all be done based on the $H_{my\ hand}$ hypothesis. This leads not only to vivid reactions when the hand is threatened (Botvinick and Cohen 1998), it also leads to reports of a strong sense of mineness.

We have explained why and how reports of selfhood might arise in our model, but so far only in a superficial way. We are yet to describe reports of the ineffability of selfhood, how it procures a higher sense of being. This requires an account of the ineffability of qualia in general.

3.3 Conscious perception and Qualia

3.3.1 Ineffability

The problem of conscious perception can be illustrated through the example of seeing the colour red.

1. On the one hand we can explain how perception of red arises: red is the result of certain wavelengths hitting our retina.

2. On the other hand, the experience seems to have certain intrinsic and indescribable qualities.

These intrinsic qualities are qualia, and we say they are ineffable to evoke our inability to communicate them. Ineffability can be an elusive concept. Here we explore how it could be made formal.

Consider two agents A and B. Here is a prospective definition of ineffability:

Definition 3.3.1 (First definition of ineffability)**.** We say that agent A's experience e is ineffable if there does not exist any formal language[4] in which A can communicate e to B to the extent that B does not learn anything new when it has experience e.

This means that whatever A says, however much it tries to describe its experience of redness, there will always be something left to grasp by agent B. This definition captures the essence of what Jackson was exploring with *Mary's room*. I argue that, like Jackson's thought experiment, this definition misfires. Instead I propose an illusionist definition:

Definition 3.3.2 (Illusionist definition of ineffability)**.** We say that agent A's experience e is ineffable if agent A *believes and claims* that there does not exist a formal language in which A can communicate e to B to the extent that B does not learn anything new when it has experience e.

Such a formal language may need to enable A to describe the entire workings of the brain. However there is no reason why this language shouldn't exist if one assumes materialism[5].

For the present work, our interest lies in explaining how this version of ineffability arises in our model.

3.3.2 Inaccessible content

Consider what happens under our model when an agent sees red and reports "My experience of red has an ineffable quality to it". There is nothing magical about our model. It is composed of simple components from top to bottom. So why does it lead to such a strange report?

If verbal reports had direct access to the whole processing of the model then they would easily convey all there is to convey. The issue is that they do not. As discussed in 2.4, speech depends largely on the abstract representations of high-level beliefs.

The brain has upwards of 100 billion neurons which each connect with up to 10,000 other neurons through synapses. It does not need to have a single location where all the information comes together (Dennett 1991). In other words, top-layer beliefs $\tilde{\mu}^{(L)}$ do not need to perfectly track low-level beliefs $\tilde{\mu}^{(l)}$[6]. In fact each component of layer l does not need to have access to the whole layer. This sparse coupling

[4]Formal languages are defined by means of an alphabet and some formation rules.

[5]And that formal languages are sufficiently powerful.

[6]Topmost layers could be biased towards more abstract representations, or they could update themselves at different rates (S. J. Kiebel, Daunizeau, and K. J. Friston 2008)

of interdependent components can be viewed as a set of nested Markov blankets (M. J. Ramstead et al. 2023). The brain is not optimised for accurate reports about conscious experiences, it is optimized for survival and efficiency.

3.3.3 What happens when we see red

So what is it exactly that our reports do not have access to when we see red? We begin with an idea which is already forged into our model: the different components of the system can be seen as building models of each other[7]. The mathematics of this is explicit in equations 3.1, 3.2 and 3.3.

Our next move borrows from the idea of "strange inversions" first proposed by Dennett (Dennett 2015; Andy Clark 2019). Consider the sweetness of chocolate. When prompted we might report that we like chocolate because it is sweet. We think that "sweetness" is an intrinsic property of chocolate. Dennett suggests an inverted story. What we project out onto the chocolate is our manifold of reactive dispositions to taste it, lick it, etc. In the words of Andy Clark (Andy Clark 2019): "it's not the sweetness which explains our response: it's our response [...] which explains the diagnosis of sweetness".

Hence the experience of sweetness is actually the result of modelling ourselves into the future, and projecting the ensuing inference onto an object in the real world, in this case chocolate. This self-modelling into the future will carry temporal depth and counterfactual richness as formally described in 3.2.2 (A. Clark, K. Friston, and Wilkinson 2019). Strange inversions exist for many other convictions: cuteness, funniness, sexiness and, we argue, redness. In fact wearing the colour red has been shown to enhance performance in sport (Hill and Barton 2005) which could lead to speculation as to what our reactive dispositions towards redness are.

Equipped with these insights, let us run through what might happen when our model sees red.

- First, sensory input o_{red} enters the system.
- The model then updates its beliefs. Layer $l+1$ comes to represent the manifold of reactive dispositions to the colour red as a latent variable through $s^{(l+1)}$.

[7]Here we again appeal to the Good Regulator Theorem (Conant and Ashby 1970) or ideally a parallel version of it: In any good regulator of a system which contains interacting components, the interacting components must model the system *and each other*.

- These beliefs act on the previous layer l. This layer then comes to represent the various dispositions and their outcomes explicitly, by modelling them as actions into the future. This is done through the distribution:

$$\begin{aligned} p(\tilde{s}^{(l)}|\tilde{s}^{(l+1)}) &= p(\tilde{s}^{(l)}|\tilde{\alpha}^{(l)}, \tilde{s}^{(l+1)})p(\tilde{\alpha}^{(l)}|\tilde{s}^{(l+1)}) \\ &= p(\tilde{s}_1^{(l)}|\tilde{s}^{(l+1)})p(\tilde{\alpha}_1^{(l)}|\tilde{s}_1^{(l+1)}) \prod_t^T p(\tilde{s}_{t+1}^{(l)}|\tilde{s}_t^{(l)}, \alpha_t^{(l)})p(\tilde{\alpha}_{t+1}^{(l)}|\tilde{\alpha}_t^{(l)}, \tilde{s}_t^{(l+1)}) \end{aligned}$$

  which accounts for actions into the (counterfactual) future. This can involve integrating information from past experiences or evolution.

- The topmost layer L exchanges information with these lower layers. However, post-update it ends up with a coarser, more abstract representation of beliefs.

- When the model then selects a policy $\pi_{report} = \mathcal{A}(\{\tilde{\phi}^{(l)}\}_{1 \leq l \leq L})$, this is heavily dependent on the topmost layer, because language is a high-level cognitive function.

- The report begins "I experience this redness as ..." On our account, the system cannot actually report on the mechanisms underlying its experience of red. What else would we expect to hear if not puzzlement?

To put it concisely, the ineffable nature of qualia comes from the fact that our higher-level beliefs, which are needed for speech, cannot fully grasp the mechanisms that drive them. Just glimpsing the colour red triggers a plethora of profound and intricate cognitive mechanisms, which then give rise to singular inherent qualities. Despite this the coarse top-level representations fall short in capturing the intricate processing that occurs.

Discussion

Conscious experience appears intuitively indescribable and puzzling to humans. Taking an illusionist point of view reframes the problem into one which is more amenable to scientific study. Following the illusionist strategy enabled us to reconsider consciousness in terms of reports, viewed as actions. With this in mind we devised a credible model of brain cognition, borrowing from the free energy principle and predictive coding. Our model included a hierarchy of components, each capable of modelling a dynamic world. This enabled it to approximate a Bayesian inference scheme for navigating a complex environment. We also provided an extension of the model which captured temporal depth and counterfactual richness explicitly. From this mathematical model, we showed that subjective properties of consciousness could arise. Perceptual unity is enforced through a dominant hypothesis in the topmost layer of the hierarchy. Selfhood is inferred as a latent variable which explains patterns of causation by including the system *itself* in models of the future. Finally we tackled qualia from an illusionist perspective, focusing on reports of ineffability. These reports have limited access to mechanisms underlying an experience. Hence they disregard large parts of what characterizes a conscious experience, most notably Dennett's "strange inversions". On this account it is unsurprising that reports on qualia include puzzlement and a sense of ineffability.

Although we argued in favour of illusionism in chapter 1, the present work can also yield tangible insights if the assumption is dropped. Rather than targeting consciousness, the illusionist strategy could be seen as a scientific template to study the meta problem of consciousness. In some sense the illusionist strategy can be disconnected from illusionism, as long as one is prepared to lower explanatory ambitions.

For illusionists, following this methodology for long enough would fully explain consciousness. The present work is far from that point and here we discuss its limitations.

From a mathematical modelling point of view, taking the variational posterior $q(x)$ to be a Dirac delta distribution limits the model's power. It restricts representations to only a single point in the space of beliefs. We also made strong assumptions regarding the generative model, most notably that noise across different dynamic orders was independent. Our discussion of the model did not investigate the role of precision optimisation which has been linked to attention (Feldman and Karl Friston 2010; Andy Clark 2013) and which plays a key role in the Bayesian brain literature (J. Hohwy 2013; Limanowski and Friston 2020). We also failed to tackle conscious

flow as one of the subjective properties of consciousness we initially listed. Indeed our model in 3.4 represented time explicitly, and hence would fail any attempt to explain the cutaneous rabbit effect (figure 1.1).

Above all the aim of this dissertation is to support the idea that behaviours associated with consciousness can arise from a mathematical model. This problem was tackled from two angles. First by contending that consciousness is not so mysterious after all, and can be captured through subjective properties. Second by building a model capable of capturing the depth and richness of human experience. While the explanatory gap persists, the present work strives to provide a compelling argument that it can in fact be bridged.

Appendix A: Mathematical glossary

Table 1: Mathematical Glossary: Free energy (2.1)

Symbol	Name	Description
x	Hidden states	These refer to states of the environment.
o	Sensory data	Input from environment.
$p(x, o)$	Generative model	Joint distribution over hidden states and sensory input.
$q(x)$	Variational posterior	Agent's representation of distribution over hidden states.
$\mathcal{F}$	Free energy	Functional in q which is minimised under the free energy principle.
μ	Beliefs about hidden states	Agent's representation of hidden states through Dirac delta distribution.

Table 2: Mathematical Glossary: Predictive Coding (2.2)

Symbol	Description
$\tilde{\mu}^{(l)}$	Beliefs at layer l in generalized coordinates of motion. Infinite vector containing successive time-derivatives $\tilde{\mu}^{(l)} = (\mu^{(l)}, \mu^{(l)}_{[1]}, \mu^{(l)}_{[2]}, ...)$.
$f_{l,n}(\mu^{(l+1)}_{[n]})$	Function of belief $\mu^{(l+1)}_{[n]}$ to represent the mean of the lower-level belief $\mu^{(l)}_{[n]}$ at order of motion n.
$\mathbf{\Sigma}_{l,n}$	Matrix representing the covariance of the density $p(\mu^{(l)}_{[n]}\|\mu^{(l+1)}_{[n]})$.
$g_n(\mu^{(L)}_{[n+1]})$	Parametrisation of mean of density $p(\mu^{(L)}_{[n+1]}\|\mu^{(L)}_{[n]})$ from one order of motion to the next in the final layer.
$\mathbf{\Theta}_n$	Matrix representing the covariance of the density $p(\mu^{(L)}_{[n+1]}\|\mu^{(L)}_{[n]})$.
b_i	Scales of component i in final layer Laplace prior.

Table 3: Mathematical Glossary: Extended model (3.2.2)

Symbol	Description
$\tilde{s}^{(l)}$	Beliefs about hidden states of the environment.
$\tilde{\alpha}^{(l)}$	Beliefs about active states of the environment.
$\tilde{\phi}^{(l)}$	Beliefs about observation states of the environment.
π	Policy, i.e. sequence of actions.

Bibliography

Andersen, Holly K. and Rick Grush (2009). "A Brief History of Time Consciousness: Historical Precursors to James and Husserl". In: *Journal of the History of Philosophy* 47.2, pp. 277–307. DOI: 10.1353/hph.0.0118.

B. Millidge, A. Seth and C. L. Buckley (2022). *Predictive Coding: a Theoretical and Experimental Review*. arXiv: 2107.12979 [cs.AI].

Blake, Randolph and Nikos Logothetis (Feb. 2002). "Visual competition". In: *Nature reviews. Neuroscience* 3, pp. 13–21. DOI: 10.1038/nrn701.

Bloom, Paul (Jan. 2004). "Descartes' Baby: How the Science of Child Development Explains What Makes Us Human". In.

Bogacz, Rafal (2017). "A tutorial on the free-energy framework for modelling perception and learning". In: *Journal of Mathematical Psychology* 76. Model-based Cognitive Neuroscience, pp. 198–211. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2015.11.003.

Botvinick, Matthew and Jonathan D. Cohen (1998). "Rubber hands 'feel' touch that eyes see". In: *Nature*. DOI: 10.1038/35784.

Buckley, Christopher L. et al. (2017). *The free energy principle for action and perception: A mathematical review*. arXiv: 1705.09156 [q-bio.NC].

Chalmers, David (1995). "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3, pp. 200–19.

— (2018). "The Meta-Problem of Consciousness". In: *Journal of Consciousness Studies* 25.9-10, pp. 6–61.

Clark, A., K. Friston, and S. Wilkinson (2019). "Bayesing Qualia: Consciousness as Inference, Not Raw Datum". In: *Journal of Consciousness Studies* 26.9-10, pp. 19–33.

Clark, Andy (2013). "The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science")". In: *Frontiers in Psychology* 4. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00270. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00270.

— (2019). "Consciousness as Generative Entanglement". In: *Journal of Philosophy* 116.12, pp. 645–662. DOI: 10.5840/jphil20191161241.

Conant, Roger C. and W. Ross Ashby (1970). "Every good regulator of a system must be a model of that system". In: *International Journal of Systems Science* 1.2, pp. 89–97. URL: https://doi.org/10.1080/00207727008920220.

Deane, George (Sept. 2021). "Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution". In: *Neuroscience of Consciousness* 2021.2. niab024. ISSN: 2057-2107. eprint: https://academic.oup.com/nc/article-pdf/2021/2/niab024/40179307/niab024.pdf. URL: https://doi.org/10.1093/nc/niab024.

Dennett, Daniel C. (1991). *Consciousness Explained*. Penguin Books.

— (2015). "Why and how does consciousness seem the way it seems?" eng. In: *Open MIND*. Frankfurt am Main: MIND Group, Kap. 10(T). DOI: http://doi.org/10.25358/openscience-139.

Dewhurst, J. and K. Dolega (2020). "Attending to the Illusion of Consciousness". In: *Journal of Consciousness Studies* 27.5-6, pp. 54–61.

Feldman, Harriet and Karl Friston (2010). "Attention, Uncertainty, and Free-Energy". In: *Frontiers in Human Neuroscience* 4. ISSN: 1662-5161. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2010.00215.

Frankish, Keith (2016). "Illusionism as a Theory of Consciousness". In: *Journal of Consciousness Studies* 23.11-12, pp. 11–39.

Friston (2003). "Learning and inference in the brain". In: *Neural Networks* 16.9. Neuroinformatics, pp. 1325–1352. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2003.06.005.

— (May 2005). "A Theory of Cortical Responses". In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360, pp. 815–36. DOI: 10.1098/rstb.2005.1622.

— (2009). "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7, pp. 293–301. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2009.04.005.

— (Feb. 2010). "Friston, K.J.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127-138". In: *Nature reviews. Neuroscience* 11, pp. 127–38. DOI: 10.1038/nrn2787.

— (2019). *A free energy principle for a particular physics*. arXiv: 1906.10184 [q-bio.NC].

Friston, Karl, Thomas Fitzgerald, et al. (Jan. 2017). "Active Inference: A Process Theory". In: *Neural Computation* 29.1, pp. 1–49. ISSN: 0899-7667. URL: https://doi.org/10.1162/NECO%5C_a%5C_00912.

Friston, Karl and Stefan Kiebel (2009). "Cortical circuits for perceptual inference". In: *Neural Networks* 22.8. Cortical Microcircuits, pp. 1093–1104. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2009.07.023.

Friston, Karl J., Richard Rosch, et al. (2018). "Deep temporal models and active inference". In: *Neuroscience Biobehavioral Reviews* 90, pp. 486–501. ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2018.04.004.

Friston, Karl J., Klaas Enno Stephan, et al. (2010). "Generalised Filtering". In.

Helmholtz, H. von et al. (1925). *Helmholtz's Treatise on Physiological Optics*. vol. 3.

Hill, Russell and Robert Barton (June 2005). "Psychology: Red enhances human performance in contests". In: *Nature* 435, p. 293. DOI: 10.1038/435293a.

Hohwy (2022). "Conscious Self-Evidencing". In: *Review of Philosophy and Psychology* 13.4, pp. 809–828. DOI: 10.1007/s13164-021-00578-x.

Hohwy, J. (2013). *The Predictive Mind*. OUP Oxford. ISBN: 9780199686735.

Hohwy and Michael (Apr. 2017). *Why should any body have a self?* DOI: 10.31234/osf.io/fm4cr. URL: psyarxiv.com/fm4cr.

Jackson, Frank (Apr. 1982). "Epiphenomenal Qualia". In: *The Philosophical Quarterly* 32.127, pp. 127–136. ISSN: 0031-8094. eprint: https://academic.oup.com/pq/article-pdf/32/127/127/4570853/pq32-0127.pdf. URL: https://doi.org/10.2307/2960077.

James, W. (1890). *The Principles of Psychology*. vol. 1.

Kiebel, Stefan J., Jean Daunizeau, and Karl J. Friston (2008). "A Hierarchy of Time-Scales and the Brain". In: *PLoS Computational Biology* 4.

Koch, Christof and Naotsugu Tsuchiya (2007). "Attention and consciousness: two distinct brain processes". In: *Trends in cognitive sciences* 11.1, pp. 16–22.

Letheby, Chris and Philip Gerrans (June 2017). "Self unbound: ego dissolution in psychedelic experience". In: *Neuroscience of Consciousness* 2017.1. nix016. ISSN: 2057-2107. eprint: https://academic.oup.com/nc/article-pdf/2017/1/nix016/25024251/nix016.pdf. URL: https://doi.org/10.1093/nc/nix016.

Levine, Joseph (1999). "Toward a science of consciousness III : the third Tucson discussions and debates". In.

Limanowski and Friston (2020). "Attenuating oneself: An active inference perspective on "selfless" experiences". In: *Philosophy and the Mind Sciences* 1.I, pp. 1–16.

Limanowski, Jakub and Felix Blankenburg (2013). "Minimal self-models and the free energy principle". In: *Frontiers in Human Neuroscience* 7. ISSN: 1662-5161. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00547.

Millidge, Beren et al. (2022). *Predictive Coding: Towards a Future of Deep Learning beyond Backpropagation?* arXiv: 2202.09467 [cs.NE].

Ramstead, Maxwell J et al. (Apr. 2023). *The inner screen model of consciousness: applying the free energy principle directly to the study of conscious experience*. DOI: 10.31234/osf.io/6afs3. URL: psyarxiv.com/6afs3.

Ramstead, Maxwell J. D. et al. (Apr. 2023). "On Bayesian mechanics: a physics of and by beliefs". In: *Interface Focus* 13.3. DOI: 10.1098/rsfs.2022.0029. URL: https://doi.org/10.1098%2Frsfs.2022.0029.

Rao, Rajesh and Dana Ballard (Feb. 1999). "Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects". In: *Nature neuroscience* 2, pp. 79–87. DOI: 10.1038/4580.

Seth, A. (2021). *Being You: A New Science of Consciousness*. Faber & Faber. ISBN: 9780571337705.

Simons, Daniel J and Christopher F Chabris (1999). "Gorillas in our midst: Sustained inattentional blindness for dynamic events". In: *perception* 28.9, pp. 1059–1074.

Vinje, William E. and Jack L. Gallant (2000). "Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision". In: *Science* 287.5456, pp. 1273–1276. DOI: 10.1126/science.287.5456.1273.