Probabilistic approaches to statistical and structure learning in human cognition

Francesco Silvestrin

School of Psychology
University of East Anglia

A thesis submitted in partial fulfilment of the requirements of the University of East Anglia for the degree of
Doctor of Philosophy

06/09/2022

Abstract

This thesis describes a series of empirical and simulation experiments, all in the broad area of probabilistic inference and learning. The first three experiments, described in Chapters 4 and 5, focus on a specific theoretical framework, predictive coding. We identified some critical issues (discussed in more detail in Chapter 1) and tackled them with a combination of techniques including pupillometry, electroencephalography (EEG) and computational modelling. In particular, we present an augmented version of classical predictive coding models incorporating dynamic precision estimation (Chapter 4) and show how human participants can successfully learn multimodal distributions, violating classical predictive coding (Chapter 5). In Chapter 6 we took a more theoretically agnostic approach (although we firmly remained within the Bayesian brain framework) to study structure learning. If in Chapter 5 we verified that humans could learn multimodal distributions, in Chapter 6 we asked how they do it without having any knowledge about the structure of the probabilistic model generating their observations. We also introduce a working memory component, with our simulations showing how revisiting past stimuli can benefit structure learning.

Overall, we contribute to the Bayesian brain framework both with empirical findings coming from both simulations and lab experiments. We augment current computational models increasing their flexibility, and thus their scope to be used in more diverse experimental contexts. Finally, we make a contribution to the field of computational rationality, discussing the trade off between working memory load and learning performance.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others. The research presented in Chapter 4 has been previously presented as a conference paper and in a poster format (Silvestrin et al., 2019) and has been published in a peer reviewed scientific journal (Silvestrin et al., 2021).

Francesco Silvestrin
06/09/2022

Table of contents

List of figures

List of tables

Notes on mathematical notation

The mathematical notation was made to be coherent across the Result Chapters (Chapters 4, 5 and 6). This made it impossible to keep the same notation in the Introduction ones (Chapters 1, 2, and 3) due to the limited number of possible latin and greek letters to be used to denote variables. This thesis has thus two separate notation groups: one for the Result Chapters and one for the Introduction. The reader should consider this when going through this thesis.

Chapter 1

Predictive coding: a mathematical introduction

1.1 The Bayesian brain

Humans find themselves immersed in a complex and multifaceted world, from which they only receive sparse and noisy sensory signals. The brain has the daunting task of making sense of this world, navigating it, and acting upon it. It must therefore identify the external and internal causes of the noisy input it receives, and, based on that, select the optimal course of action. Determining the (hidden) causes of incoming signal is often called the *inverse problem*, which, for most of the real-world scenarios the brain encounters daily, has no single solution. In fact, incoming sensory signals might have more than one (and possibly infinite) possible combinations of objects and situations causing them. The sound of steps in the night might be caused by a burglar, but also by your partner going to the bathroom. The brain must solve this ambiguity, as different interpretations of the sensory data have different optimal behavioural responses (in our example, calling the police or going back to sleep). According to probabilistic accounts of brain function, known collectively as the *Bayesian brain* hypothesis, (Knill and Pouget, 2004), the brain deals with this uncertainty by combining sensory input coming from different sources with prior knowledge using the Bayes rule. The brain is thus conceptualised as an inference machine which, given multidimensional sensory observations $\mathbf{o}$ with a set of N possible hidden causes $\mathbf{z}$ (also called *hidden states* or *latent variables*), infers the probability of a certain cause z_n generating the observations as given by

$$p(z_n \mid \mathbf{o}) = \frac{p(\mathbf{o} \mid z_n)p(z_n)}{\sum_{j=1}^{N} p(z_j)p(\mathbf{o} \mid z_j)} \tag{1.1}$$

where $p(z_n \mid \mathbf{o})$ is the *posterior* (probability of a hidden state having value z_n given the observations $\mathbf{o}$), $p(\mathbf{o} \mid z_n)$ is the *likelihood* (probability of observing $\mathbf{o}$ given z_n) and $p(z_n)$ is the *prior* (probability of the hidden state having value z_n regardless of the sensory input). The denominator $\sum_{j=1}^{N} p(z_j)p(\mathbf{o} \mid z_j) = p(\mathbf{o})$, called *model evidence*, represents the overall probability of observing $\mathbf{o}$ and will be discussed in later sections.

This process involves representing information probabilistically (i.e. in the form of probability distributions). Going back to our footsteps example, the brain can solve the ambiguity by integrating its auditory (hearing footsteps) and visual (seeing your partner is not in bed) inputs with prior knowledge (regardless of footsteps, your partner needing the bathroom is a more likely scenario than your house being robbed). It can thus conclude that the footsteps are almost certainly caused by your partner and that you can safely go back to sleep.

Although for clarity purposes this example involved a higher level, conscious inference process, the same principles are valid for lower level unconscious inferences in primary sensory areas. These involve receiving noisy signals from sensory receptors and disambiguating them by integrating them with prior knowledge, giving rise to a clean percept. This entails a fundamental concept of the Bayesian brain, namely that perception is the result of unconscious inference (Knill and Pouget, 2004; Yuille and Kersten, 2006). Furthermore, these principles can be extended to sensorimotor control (Körding and Wolpert, 2004), with actions choice being the result of probabilistic inference integrating several sensory channels (i.e. multisensory integration, Ernst and Banks (2002)) with priors to estimate the outcomes of possible action sequences and selecting the most rewarding.

The fact that the brain uses prior knowledge means that it has an internal model of the environment. This model specifies (probabilistically) how the environment's hidden states generate sensory observations, and is thus called a *generative model*. Mathematically, the generative model represents the product of the prior and the likelihood (i.e. the numerator of the Bayes equation). As more evidence (in the form of sensory signals) is gathered throughout an individual's life, the generative model gets updated in light of it. In other words, the individual learns from experience. Formally, this means updating the parameters $\boldsymbol{\theta}$ of the generative model. We can thus distinguish between *inference* (estimating time-varying hidden states causing sensory observations) and *learning* (updating time-invariant generative model parameters). As the generative model predicts observations based on hidden states and model parameters, both inference and learning involve *Bayesian model inversion* (i.e. estimating hidden states and parameters from observations). If we include parameters update, we can rewrite

the Bayes rule as

$$p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{o}) = \frac{p(\mathbf{o} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta})}{\sum_{i=1}^{M} \sum_{j=1}^{N} p(z_j, \boldsymbol{\theta}_i) p(\mathbf{o} \mid z_j, \boldsymbol{\theta}_i)} \tag{1.2}$$

with $p(\mathbf{o}, \mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{o} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta})$ being the generative model and $\boldsymbol{\theta}_i$ one of the M possible set of values of the model's parameters. Note that up to this point we have only considered discrete-valued (i.e. with a finite number of possible values) hidden states and parameters, but the equations can be modified to deal with continuous variables by substituting the sum symbols with integrals.

This general approach has been extensively and successfully deployed to study a variety of phenomena. For example, several studies (Knill, 1998; Otten et al., 2017) point at the fact that visual perception is influenced by priors, a key prediction of the Bayesian brain hypothesis, as the experienced percept (the posterior) is thought to be the result of the combination of the sensory evidence (likelihood) and predictions based on prior knowledge (priors). This is true for low-level perceptual priors (Knill, 1998) as well as higher order ones (e.g. social priors, Otten et al. (2017)). Other fields of neuroscience that have been studied within this framework include 3D shape perception (Erdogan and Jacobs, 2017), multisensory integration (Ernst and Banks, 2002; Jacobs, 1999; Parise and Ernst, 2017), sensorimotor control (Kim, 2021; Körding and Wolpert, 2004; Todorov, 2004), and higher cognition (Baker et al., 2017; Steyvers et al., 2006; Tenenbaum et al., 2006).

1.2 Approximate inference

The Bayes rule provides a straightforward way of performing model inversion for simple problems where hidden states and model parameters can assume a manageable, discrete number of possible values. However, in a real-world scenario, the brain has thousands of separate, possibly continuously-valued stimulus features to make sense of at any given time, which would make the model evidence (i.e. the denominator in the Bayes rule) intractable. In fact, evaluating the model evidence would require the brain to sum (or integrate) over all possible values of all hidden states and model parameters, which is challenging (or impossible) even for fairly simple models (Gershman and Beck, 2017).

In statistics and machine learning, the problem of intractable exact Bayesian inference is solved by finding an approximate, tractable approximation of the exact solution. These methods take the name of *approximate inference*, and are broadly

divided into two main algorithmic families: *variatonal* and *sampling* (or *Monte Carlo*) methods, which have given rise to different probabilistic accounts of brain function (Gershman and Beck, 2017).

Briefly, Monte Carlo methods are based on the general idea that the posterior distribution over a latent variable can be approximated to an empirical point-mass function by drawing a set of samples from it. There are a variety of algorithm one can use to draw samples (Bishop, 2006), but all of them result in approximations of the posterior whose accuracy depends on the number of samples drawn (the bigger the sample size, the more accurate the approximation). As the number of samples approaches infinity, the approximate posterior will asymptotically approach the true posterior. One could thus say that Monte Carlo methods consist in performing *approximate inference on exact models* (i.e. the inference process is approximate, but the model is not altered or simplified in any way). These methods have inspired several different probabilistic theories of brain function (Aitchison and Lengyel, 2016; Gershman et al., 2012; Sanborn and Chater, 2016) and have been used to explain several cognitive phenomena (Chater et al., 2020; Vul et al., 2014). These will not be further discussed in this thesis as they are not the focus of the experimental work described in the next chapters.

Variational inference, on the other hand, is based on approximating the posterior to a distribution of a parametric family chosen a priori (Bishop, 2006). The parameters of the approximate posterior are then fitted to maximise a lower bound on model evidence (see next sections for more details). If the true and approximate posterior's parametric families are different, one will never end up performing exact inference. Nevertheless, these approximations are often sufficiently accurate and computationally cheaper than sampling (Gershman and Beck, 2017). These methods, contrary to Monte Carlo ones, simplify the model to make its inversion tractable. One could thus say that variational inference consists in performing *exact inference on approximate models*. Variational inference is at the core of the theoretical framework we refer to in this thesis, and has been used to build the computational models described in Chapters 4, 5 and 6. It will thus be described in more detail in the next section.

1.2.1 Variational inference

As mentioned, most of the work presented in this thesis relies heavily on variational approximations, so variational inference will be discussed in more detail (although for a full demonstration see Bishop (2006)). We will consider a generative model with I

continuous latent variables (we do not include parameter update for simplicity, but from a mathematical standpoint the same rules apply).

1.2.1.1 Variational Free Energy

The real posterior $p(\mathbf{z} \mid \mathbf{o})$ is approximated with a distribution $q(\mathbf{z})$ of some parametric family chosen a priori, and parameters are optimised to make the approximation as close as possible to the real posterior. Formally, this involves minimising the Kullback-Leibler (KL) divergence between the two:

$$KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right] = -\int q(\mathbf{z}) ln\left(\frac{p(\mathbf{z} \mid \mathbf{o})}{q(\mathbf{z})}\right) d\mathbf{z} \tag{1.3}$$

which can be reduced to 0 only when the true posterior belongs to the parametric family chosen for $q(z)$. Note that the KL divergence is not symmetrical, so $KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right] \neq KL\left[p(\mathbf{z} \mid \mathbf{o}) \mid\mid q(\mathbf{z})\right]$. In this thesis we will focus solely on algorithms minimising $KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right]$, but there are variational inference algorithms based on $KL\left[p(\mathbf{z} \mid \mathbf{o}) \mid\mid q(\mathbf{z})\right]$ minimisation as well (i.e. *expectation propagation*, Bishop (2006)).

Directly minimising the KL divergence between approximate and true posterior is still not tractable, as it is a function of the (unknown) true posterior. The solution to this is maximising a quantity known as *Variational Free Energy* (VFE):

$$VFE = lnp(\mathbf{o}) - KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right] \tag{1.4}$$

As both $lnp(\mathbf{o})$ and $-KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right]$ are always non-positive, maximising VFE is equivalent to minimising $KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right]$. Alternatively, VFE can be seen as a lower bound on model evidence, and can take the name of *evidence lower bound* ($ELBO$).

VFE is a tractable quantity as

$$\begin{aligned} VFE &= lnp(\mathbf{o}) - KL\left[q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{o})\right] \\ &= lnp(\mathbf{o}) + \int q(\mathbf{z}) ln\left(\frac{p(\mathbf{z} \mid \mathbf{o})}{q(\mathbf{z})}\right) d\mathbf{z} \\ &= \int q(\mathbf{z}) ln\left(\frac{p(\mathbf{z}, \mathbf{o})}{q(\mathbf{z})}\right) d\mathbf{z} \end{aligned} \tag{1.5}$$

is a function of the joint $p(\mathbf{z}, \mathbf{o}) = p(\mathbf{o} \mid \mathbf{z})p(\mathbf{z})$.

1.2.1.2 Mean field approximation

To make VFE optimisation tractable, we assume the various latent variables to be independent from each other, so that approximate posterior factorises as

$$q(z_1, ..., z_I) = \prod_{i=1}^{N} q(z_i) \tag{1.6}$$

This is often not the case for the true posterior, and therefore represents a further approximation (*mean field approximation*).

We now re-write VFE as

$$VFE = E_{q(\mathbf{z})}\left[ln\left(\frac{p(\mathbf{z},\mathbf{o})}{q(\mathbf{z})}\right)\right] \tag{1.7}$$

where the notation $E_{f(x)}[g(x)]$ indicates the expected value of $g(x)$ under the distribution $f(x)$ so that

$$E_{f(x)}[g(x)] = \int f(x)g(x)dx \tag{1.8}$$

for a continuous variable x.

If we apply the mean field approximation, it can be shown (Bishop, 2006) that

$$lnq^*(z_j) = E_{q(\mathbf{z}_{i\neq j})}\left[lnp(\mathbf{z},\mathbf{o})\right] + const \tag{1.9}$$

with $q^*(\cdot)$ representing the optimal approximate posterior and $E_{q(\mathbf{z}_{i\neq j})}\left[lnp(\mathbf{z},\mathbf{o})\right]$ the expected value of the log joint under all approximate posteriors except for $q(z_j)$ (which would end up being a function of z_j).

Variational inference is at the core of one the most popular probabilistic accounts of brain function, *predictive coding* (PC). This is going to be the reference theoretical framework for much of this thesis, and will therefore be discussed in more depth in the next section.

1.3 Predictive Coding

1.3.1 Prediction error minimisation

The core idea of predictive coding (Rao and Ballard, 1999) is that the brain is hierarchically organised (from low level sensory areas to high level associative areas), and that its fundamental function is to minimise prediction errors at each level of the hierarchy. Here prediction errors represent the discrepancy between predictions coming

from the level above (feedback signal) and the signal coming from the level below (feed-forward signal). In other words, at each level of the processing hierarchy the brain is trying to formulate predictions to "explain away" the signal coming from the level below, and only the portion of the signal predictions can't account for is passed on to the next level. In this context, perception is the result of short-term prediction error minimisation in low level sensory areas, while learning is the result of longer-term updating of predictions to better account for future incoming signal (i.e. stimuli).

This general idea was first proposed by Rao and Ballard (1999), who developed a model of visual processing based on these principles and used it to explain the extra-classical receptive field effects (Allman et al., 1985). In the following two decades this framework has been further developed and expanded, and today is perhaps the most influential and best worked-out (both in mathematical and neurobiological terms) unified account of brain function.

1.3.2 Predictive Coding as variational inference

In its modern and most widespread version, PC is formalised as mean-field variational inference with a generative model with Gaussian variables (Friston, 2005; Friston and Kiebel, 2009). From this premises, one can derive inference and learning as the result of prediction error minimisation.

To illustrate this, let's consider a very simple case, in which an agent is exposed to series of sequential (continuous) observations $\mathbf{o} = \{o_1, ..., o_T\}$ (which here we make one-dimensional for simplicity), which are assumed to be caused by hidden states $\mathbf{z} = \{z_1, ..., z_T\}$ so that

$$o_i = f(z_i) + \epsilon \tag{1.10}$$

with ϵ being Gaussian noise. Thus

$$p(o_t \mid z_t) = \mathcal{N}\left(o_t \mid f(z_t), \sigma^{(0)2}\right) \tag{1.11}$$

where the notation $\mathcal{N}(x \mid \mu, \sigma^2)$ denotes a Gaussian distribution over x with mean μ and variance σ^2. The superscript in $\sigma^{(0)2}$ indicates the hierarchical level.

The agent's prior over z_t is also Gaussian, so that

$$p(z_t) = \mathcal{N}\left(z_t \mid m, \sigma^{(1)2}\right) \tag{1.12}$$

where m and $\sigma^{(1)2}$ represent the sufficient statistics of the prior distribution over z. For simplicity, we assume hidden states (and thus observations) to be independent

from one another, so that

$$p(\mathbf{z}) = \prod_{t=1}^{T} p(z_t) \tag{1.13}$$

The agent is presented with the observations sequentially (as it would most likely be the case in a real-world environment). It therefore has to infer online the values of each hidden state z_t and update its beliefs about m. To perform such update (i.e. learning), we introduce a further Gaussian prior (hyperprior) over m, so that at time t

$$p(m) = \mathcal{N}(m \mid \mu_t, \sigma_t^{(2)2}) \tag{1.14}$$

The temporal indexing on μ and $\sigma^{(2)2}$ is necessary as the agent's priors change over time as a result of experience. On the other hand, we do not place temporal indices on $\sigma^{(0)2}$ and $\sigma^{(1)2}$, which here will be treated as fixed parameters and will be discussed in more depth in the next section.

Thus, as it is exposed to any observation o_t, it must infer the posterior distribution

$$\begin{aligned} &p(z_t, m \mid o_t, \sigma^{(0)2}, \sigma^{(1)2}, \mu_t, \sigma_t^{(2)2}) = \\ &\frac{\mathcal{N}(o_t \mid z_t, \sigma^{(0)2}) \mathcal{N}\left(z_t \mid m, \sigma^{(1)2}\right) \mathcal{N}\left(m \mid \mu_t, \sigma_t^{(2)2}\right)}{\int p(o_t \mid z_t, m) \mathcal{N}\left(z_t \mid m, \sigma_t^{(1)2}\right) \mathcal{N}\left(m \mid \mu_t, \sigma_t^{(2)2}\right) dz_t dm} \end{aligned} \tag{1.15}$$

Note that here we are assuming the agent to be learning completely online, and therefore have no memory of individual past observations (the limitations of this "memoryless" approach are discussed in Chapter 6). The agent thus relies on constantly updating the sufficient statistics of the prior distribution over the hidden states (here m).

As discussed earlier, the integral at the denominator (model evidence) is intractable, and the agent must therefore resort to approximate inference. As mentioned, there are several ways this could be done, and this diversity of methods is reflected by the different probabilistic accounts of brain function. Here we focus on PC, which in its most common form (Friston and Kiebel, 2009) is based on the optimisation of VFE (although see Spratling (2017) for alternative algorithms).

Our agent thus needs to optimise

$$VFE = E_{q(z_t)} \left[ln \frac{p(o_t, z_t, m)}{q(z_t, m)} \right] \tag{1.16}$$

where both the hidden state z_t and the parameter m receive the same mathematical treatment. The optimal variational posteriors $q^*(z_t)$ and $q^*(m)$ can be evaluated as in

equation 1.11. Thus

$$\begin{aligned} lnq^*(z_t) &= E_{q(m)}\left[lnp(o_t, z_t)\right] + const \\ &= ln\mathcal{N}\left(o_t \mid f(z_t), \sigma^{(0)2}\right) + E_{q(m)}\left[ln\mathcal{N}\left(z_t \mid m, \sigma^{(1)2}\right)\right] + const \end{aligned} \tag{1.17}$$

which, after taking out terms that do not depend on z_t, becomes

$$lnq^*(z_t) = -\frac{(o_t - f(z_t))^2}{2\sigma^{(0)2}} - \frac{E_{q(m)}\left[(z_t - m)\right]^2}{2\sigma^{(1)2}} + const \tag{1.18}$$

It is convenient to re-write this replacing variances with their inverse, precision, so

$$lnq^*(z_t) = -\frac{\lambda^{(0)}(o_t - f(z_t))^2}{2} - \frac{\lambda^{(1)}\left((z_t - \breve{\mu}_t^{(2)})^2 + \breve{\lambda}_t^{(2)-1}\right)}{2} + const \tag{1.19}$$

where $\lambda^{(0)-1} = \sigma^{(0)2}$, $\lambda^{(1)-1} = \sigma^{(1)2}$ and $\breve{\mu}_t^{(2)}$ and $\breve{\lambda}_t^{(2)}$ are mean and precision of the variational posterior $q^*(m) = \mathcal{N}(m \mid \breve{\mu}_t^{(2)}, \breve{\lambda}_t^{(2)-1})$, respectively (see below). These values are initially set to those of the priors $(\mu_t, \lambda_t^{(2)})$.
To get mean $\breve{\mu}_t^{(1)}$ and precision $\breve{\lambda}_t^{(1)}$ of the optimal posterior $q^*(z_t) = \mathcal{N}(z_t \mid \breve{\mu}_t, \breve{\lambda}_t^{-1})$, the agent makes use of the Laplace approximation, thus

$$\breve{\mu}_t^{(1)} = \underset{z_t}{argmax}(q^*(z_t)) \tag{1.20}$$

and

$$\breve{\lambda}_t^{(1)} = -\frac{\partial^2 q^*(z_t)}{\partial z_t^2}\left(\underset{z_t}{argmax}(q^*(z_t))\right) \tag{1.21}$$

To get $\breve{\mu}_t$, the agent makes use of gradient ascent, so it iteratively evaluates

$$\breve{\mu}_t^{(1)} \longleftarrow \breve{\mu}_t^{(1)} + \Delta\breve{\mu}_t^{(1)} \tag{1.22}$$

until convergence, where

$$\begin{aligned} \Delta\breve{\mu}_t^{(1)} &\propto -\frac{\partial q^*(z_t)}{\partial z_t} \\ &= \lambda^{(0)} f'(\breve{\mu}_t^{(1)})(o_t - f(\breve{\mu}_t^{(1)})) - \lambda^{(1)}(\breve{\mu}_t^{(1)} - \breve{\mu}_t^{(2)}) \end{aligned} \tag{1.23}$$

which illustrates the core idea of PC, namely hierarchical prediction error minimisation. In fact, this equation represents a trade-off between prediction errors at a lower $(o_t - f(\breve{\mu}_t^{(1)}))$ and higher $(\breve{\mu}_t^{(1)} - \breve{\mu}_t^{(2)})$ levels of the hierarchy, weighted by their respective precisions ($\lambda^{(0)}$ and $\lambda^{(1)}$), the role of which we will further discuss in later sections.

From a PC perspective, $f(\breve{\mu}_t^{(1)})$ and $\breve{\mu}_t^{(2)}$ represent the top down signals attempting to predict, or "explain away" (i.e. suppressing) the bottom up signals (o_t and $\breve{\mu}_t^{(1)}$, respectively). Here $f'(\breve{\mu}_t^{(1)})$ can be seen as a scaling term, which ensures the two prediction errors are evaluated on the same scale.

As for the posterior's precision,

$$\breve{\lambda}_t^{(1)} = \lambda^{(0)}(f''(z_t)o_t - f''(z_t)f(z_t) - (f'(z_t))^2) + \lambda^{(1)} \tag{1.24}$$

where again the term $(f''(z_t)o_t - f''(z_t)f(z_t) - (f'(z_t))^2)$ ensures proper scaling.

Now the agent has inferred a value for the hidden state z_t ($\breve{\mu}_t^{(1)}$) with a certain degree of confidence ($\breve{\lambda}_t^{(1)}$). Note that this entails a probabilistic (Gaussian) representation of z_t, in line with the general Bayesian brain hypothesis (Knill and Pouget, 2004). It can now use this information to update its beliefs and make more accurate predictions in the future. It does so by evaluating

$$\begin{aligned} q^*(m) &= E_{q(z_t)}\left[lnp(o_t, z_t, m)\right] + const \\ &= -\frac{\lambda^{(1)} E_{q(z_t)}\left[(z_t - m)^2\right]}{2} - \frac{\lambda_t^{(2)}(m - \mu_t)^2}{2} + const \end{aligned} \tag{1.25}$$

From here on the same principles outlined above apply, so

$$\breve{\mu}_t^{(2)} = \underset{m}{argmax}(q^*(m)) \tag{1.26}$$

$$\breve{\lambda}_t^{(2)} = -\frac{\partial^2 q^*(m)}{\partial m^2}\left(\underset{m}{argmax}(q^*(m))\right) \tag{1.27}$$

and, to perform gradient ascent

$$\breve{\mu}_t^{(2)} \longleftarrow \breve{\mu}_t^{(2)} + \Delta\breve{\mu}_t^{(2)} \tag{1.28}$$

where

$$\begin{aligned} \Delta\breve{\mu}_t^{(2)} &\propto -\frac{\partial q^*(m)}{\partial m} \\ &= \lambda^{(1)}(\breve{\mu}_t^{(1)} - \breve{\mu}_t^{(2)}) - \lambda_t^{(2)}(\breve{\mu}_t^{(2)} - \mu_t) \end{aligned} \tag{1.29}$$

Again, we notice that this is another trade-off between precision-weighted prediction errors. The posterior precision can then be calculated as

$$\breve{\lambda}_t^{(2)} = \lambda^{(1)} + \lambda_t^{(2)} \tag{1.30}$$

The evaluation of $q^*(z_t)$ and $q^*(z_t)$ can be carried out recursively, until the values of the posterior parameters converge. The agent's priors can then be updated, so that

$$\mathcal{N}(m \mid \mu_{t+1}, \lambda_{t+1}^{(2)-1}) = \mathcal{N}(m \mid \breve{\mu}_t, \breve{\lambda}_t^{(2)-1}) \tag{1.31}$$

The last two equations nicely illustrate the intuitive concept of confidence about one's beliefs increasing with experience, as the prior's updated precision is always the sum of its old precision and the posterior's precision (and precision is always positive).

The overall picture that emerges from this is that of a set of beliefs at various hierarchical levels being optimised to better account for and suppress the incoming sensory data. In the case of inference about hidden states this results in a flexible process, allowing to keep track of changing environmental variables. On the other end, learning relies on parameter update, which decreases in magnitude as evidence is accumulated, as the prior's precision always increases. Nevertheless, both processes stem from the same algorithm (variational inference) and optimise the same quantity (VFE), giving PC a unified and comprehensive mathematical framework. This has provided an elegant explanation for a range of perceptual phenomena, as bistable perception (Denison et al., 2011; Hohwy et al., 2008; Weilnhammer et al., 2017) and motion illusions (Watanabe et al., 2018), as well as neural ones, such as classical and extra-classical receptive field effects in the primary visual cortex(Rao and Ballard, 1999; Spratling, 2010), repetition suppression (Auksztulewicz and Friston, 2016), and electrophysiolgical responses to violation of expectations (Garrido et al., 2008, 2009a,b). Furthermore, theoretical work has also pointed to PC's biological plausibility (Bastos et al., 2012). PC has also been the object of theories of higher cognition, ranging from consciousness (Seth et al., 2012) to theory of mind (Koster-Hale and Saxe, 2013) and emotions (Seth, 2013).

It is worth mentioning that the core mathematical goal of VFE optimisation has been extended to formulate a unifying theory of perception, cognition and behaviour, the *Free Energy Principle* (Friston, 2010). This incorporates actions as well (Friston et al., 2009), but for the purposes of this thesis we will limit further discussions to PC only.

1.3.3 Precision

As illustrated in the previous section, precision acts as a learning rate, regulating the trade-off in prediction errors minimisation. The higher the precision, the more the associated prediction error will be weighted. Going back to the example discussed

above, a very precise prior about the hidden state z_t (high $\lambda^{(1)}$) combined with a very noisy input (low $\lambda^{(0)}$) will cause the agent's best guess about the value of the hidden state ($\breve{\mu}_t^{(1)}$) to closely resemble the prediction ($\breve{\mu}_t^{(2)}$) rather than the observation itself ($f^{-1}(o_t)$). The precisions are thus regulating the relative weight of sensory evidence and prior beliefs.

Let's now consider another situation, in which an agent must infer a single hidden state z from two different sensory streams, o and u.

$$o_t = f(z_t) + \epsilon^{(o)} \tag{1.32}$$

$$u_t = g(z_t) + \epsilon^{(u)} \tag{1.33}$$

with $\epsilon^{(o)}$ and $\epsilon^{(u)}$ being the Gaussian noise associated with o and u respectively. For brevity and clarity, let's set both $f(\cdot)$ and $g(\cdot)$ to correspond to the identity function. Thus

$$o_t = z_t + \epsilon^{(o)} \tag{1.34}$$

$$u_t = z_t + \epsilon^{(u)} \tag{1.35}$$

and

$$p(o_t, u_t \mid z_t) = \mathcal{N}(o_t \mid z_t, \lambda^{(o)-1})\mathcal{N}(u_t \mid z_t, \lambda^{(u)-1}) \tag{1.36}$$

where $\lambda^{(o)}$ and $\lambda^{(u)}$ are the respective precisions associated with the two sensory streams, replacing what was $\lambda^{(0)}$ in the previous section.

As before, the agent has a Gaussian prior over z

$$p(z_t) = \mathcal{N}(z_t \mid \mu, \lambda^{(1)}) \tag{1.37}$$

Here we will omit learning for compactness, and because it would result in a repetition of what discussed in the previous section. Therefore we do not use the Roman notation m for the hidden state's prior mean and we place no further prior over it.
Applying variational inference as before, we get

$$\Delta\breve{\mu}^{(1)} \propto \lambda^{(o)}(o_t - \breve{\mu}^{(1)}) + \lambda^{(u)}(u_t - \breve{\mu}^{(1)}) - \lambda^{(1)}(\breve{\mu}^{(1)} - \mu) \tag{1.38}$$

where $-\lambda^{(o)}$ and $-\lambda^{(u)}$ control the relative weight of the two prediction errors, and thus of the two sensory streams, while $-\lambda^{(1)}$ controls the relative weight of the prior about the value of z.

From a neuroscientific perspective, this illustrates how precision not only regulates information integration between different cortical hierarchical levels, but also between different channels converging on the same hierarchical level, multisensory integration being the obvious example (Crucianelli et al., 2019).

At lower hierarchical levels (i.e. sensory cortices) precision associated with different features of the environment (λ^0, λ^o and λ^u in our examples) can be thought as attention, or the amount of cognitive resources invested in them. If one focuses on a particular feature, this will drive inference (e.g. object identification) more than unattended ones (Feldman and Friston, 2010). At higher cognitive levels, precision can more simply reflect how reliable or salient a source of information is thought to be (Macaluso et al., 2016), or confidence about certain beliefs (Adams et al., 2014).

Of particular relevance for this thesis is the work carried out by Garrido et al. (2013), who showed how identical signals could elicit different brain responses depending on the precision of the participant's predictions. More specifically, an outlier stimulus has been shown to elicit a greater response if it violated a highly precise prediction, reflecting the low probability associated with it, or, equivalently, *surprise* (see Chapter 2). This study was the central inspiration for the experiment described in Chapter 4, where we replicate some of its findings.

Precision is thus a crucial quantity in predictive coding, and is the focus of several PC-inspired theories of psychopathology. For example, Van de Cruys et al. (2014) provided a PC account of autism spectrum disorders, suggesting that hyper-precise priors prevent autistic individuals from effectively suppress prediction errors, as the excessive precision prevents predictions to adjust to sensory evidence. This is thought to be at the core of the lack of flexibility exhibited by individuals with autism, and has been used to explain much of the autistic symptomatology (Van de Cruys et al., 2014). Furthermore, (Adams et al., 2013) suggested that abnormal encoding of precision is at the core of psychotic disorders, and offered a nerobiologically plausible account of this. Finally, Kube et al. (2020) offered a PC view of depressive disorders, attributing processing biases (i.e. giving more weight to information with negative valence) to excessive precision attributed to beliefs about negative events.

Despite the role of precision in cognitive and perceptual processes has received a lot of attention from academics, it tends to be modelled as a fixed quantity, as in the examples above. However, although precision to some extent might be hard-wired, biological agents often need to estimate it in order to let the most appropriate information sources guide action. Experimental work aimed at investigating how precision is estimated and tracked in real time would thus be of great importance to

the field, but is notably absent from the literature. This is the main focus of Chapter 4, where we present such an experiment.

1.3.4 Why Gaussians? Backwards derivation

In section 1.3.2 we showed how inverting a Gaussian generative model leads naturally to prediction error minimisation. To better clarify why PC needs Gaussian distributions, we now derive this backwards, showing how prediction error minimisation implies a Gaussian generative model.

Let's consider the general case of an agent trying to maximise the log probability L of some hidden variable z based on an observation o solely by minimising the prediction error $(o-z)$. As in the previous section we are setting $f(z)=z$ for simplicity. This corresponds to gradient ascent where

$$\frac{dL}{dz} = w(o-z) \tag{1.39}$$

where w is the learning rate, a constant governing the speed of gradient ascent. Integrating gives

$$L = -\frac{w(z^2-2oz)}{2} + const \tag{1.40}$$

where *const* is an arbitrary constant. Setting

$$const = \frac{wo^2}{2} - \frac{ln\left(2\pi w^{-1}\right)}{2} \tag{1.41}$$

gives

$$\begin{aligned} L &= -\frac{ln\left(2\pi w^{-1}\right)}{2} - \frac{w(o-z)^2}{2} \\ &= ln\mathcal{N}\left(o \mid z, w^{-1}\right) \end{aligned} \tag{1.42}$$

From here we note that the learning rate w corresponds to the precision of a Gaussian distribution with mean z. Thus minimising prediction error always corresponds to maximising the log probability of some Gaussian.

The same observation applies to the sorts of situation described in the previous sections and typically considered in theories of PC (Bogacz, 2017; Friston, 2005), in which the agent infers on z based on o and prior expectation μ by minimising the sum of the first $(o-z)$ and second level $(z-\mu)$ prediction errors, weighted by constants

$w^{(1)}$ and $w^{(2)}$. In such case

$$\frac{dL}{dz} = w^{(1)}(o-z) - w^{(2)}(z-\mu) \tag{1.43}$$

and

$$L = -\frac{w^{(1)}(z^2-2oz)}{2} - \frac{w^{(2)}(z^2-2z\mu)}{2} + const \tag{1.44}$$

Setting

$$const = \frac{w^{(1)}o^2}{2} - \frac{ln\left(2\pi w^{(1)-1}\right)}{2} + \frac{w^{(2)}\mu^2}{2} - \frac{ln\left(2\pi w^{(2)-1}\right)}{2} \tag{1.45}$$

gives

$$\begin{aligned} L &= -\frac{ln\left(2\pi w^{(1)-1}\right)}{2} - \frac{w^{(1)}(o-z)^2}{2} - \frac{ln\left(2\pi w^{(2)-1}\right)}{2} - \frac{w^{(2)}(z-\mu)^2}{2} \\ &= ln\mathcal{N}\left(o \mid z, w^{(1)-1}\right) + ln\mathcal{N}\left(z \mid \mu, w^{(2)-1}\right) \end{aligned} \tag{1.46}$$

Thus trading off first and second level prediction errors corresponds to combining two Gaussians based on their relative precisions.

In classical PC therefore all variables must be treated as having a Gaussian probability distribution (Friston, 2005; Friston and Kiebel, 2009). This constitutes a major assumption, which, if violated, can lead to suboptimal inference and learning. We discuss this in Chapter 5, where we consider a very evident violation of such assumption, namely bimodal distributions.

Chapter 2

Physiological signals for tracking human learning

2.1 Predictive coding in the brain

As anticipated in the previous chapter, with the seminal work from Rao and Ballard (1999) predictive coding was presented as a theory of cortical function. The core idea, which is shared with more modern accounts, was that of a hierarchically organised cortex, with predictions travelling down the hierarchy and ultimately attempting to suppress, or "explain away", the incoming sensory signals. The portion of signal predictions failed to account for (i.e. prediction errors) would instead travel up the hierarchy. At each hierarchical level feedback predictive signals would thus suppress upcoming, feed-forward error signals.

This original formulation (Rao and Ballard, 1999) was an attempt to model the functioning of the visual cortex only. Friston (2005) expanded the framework to the whole cortex, and formulated predictive coding as Variational Free Energy (VFE) maximisation, crucially introducing the concept of precision (or inverse variance). Furthermore, the author discussed a possible neural implementation of predictive coding as interplay between representation and error units at different hierarchical levels. For illustrating this mathematically we are going to follow Bogacz (2017) instead of the original paper (Friston, 2005), as the derivations and results are very similar, but significantly easier to understand. There are minor differences, which will be discussed.

2.1.1 Inference

Going back to the example discussed in the previous chapter, let's consider the generative model

$$p(o,z) = \mathcal{N}\left(o \mid f(z,\boldsymbol{\theta}),\sigma^{(0)2}\right)\mathcal{N}\left(z \mid \mu,\sigma^{(1)2}\right) \tag{2.1}$$

where $\boldsymbol{\theta} = \{\theta_1,...,\theta_I\}$ are the parameters of the function $f(\cdot)$ that maps the the hidden state z to the observation o. The variational free energy can thus be written as

$$\begin{aligned} VFE &= E_{q(z)}\left[ln\frac{p(o,z)}{q(z)}\right] \\ &= ln\mathcal{N}\left(o \mid f(E[z],\boldsymbol{\theta}),\sigma^{(0)2}\right)\mathcal{N}\left(E[z] \mid \mu,\sigma^{(1)2}\right) \\ &= -\frac{1}{2}\frac{(o-f(E[z],\boldsymbol{\theta}))^2}{\sigma^{(0)2}} - \frac{1}{2}\frac{(E[z]-\mu)^2}{\sigma^{(1)2}} - \frac{1}{2}ln\sigma^{(0)2} - \frac{1}{2}ln\sigma^{(1)2} - ln(2\pi) \end{aligned} \tag{2.2}$$

where $q(z)$ is set to be a Delta function, and thus its entropy is

$$\begin{aligned} \mathcal{H}\big(q(z)\big) &= \int q(z)lnq(z)dz \\ &= 0 \end{aligned} \tag{2.3}$$

and the expected value of z^2 is

$$E\left[f(z,\boldsymbol{\theta})\right] = f(E[z],\boldsymbol{\theta}) \tag{2.4}$$

Let's now set

$$\xi^{(0)} = \frac{o-\phi}{\sigma^{(0)2}} \tag{2.5}$$

$$\xi^{(1)} = \frac{\phi-\mu}{\sigma^{(2)2}} \tag{2.6}$$

which are going to are going to represent our *error neurons*, and

$$\phi = E[z] \tag{2.7}$$

which is going to represent our *representation neuron*. Note that the error neurons encode precision-weighted prediction errors (even though in this case we are using the variance notation, for reasons that will become clear later). This is where Bogacz

(2017) and Friston (2005) make different choices, as the latter formalised error units as

$$\xi^{(0)} = \frac{o - \phi}{1 + \gamma^{(0)}} \tag{2.8}$$

$$\xi^{(1)} = \frac{\phi - \mu}{1 + \gamma^{(1)}} \tag{2.9}$$

with

$$\gamma^{(0)} = \sigma^{(0)} - 1 \tag{2.10}$$

and

$$\gamma^{(1)} = \sigma^{(1)} - 1 \tag{2.11}$$

therefore weighting prediction errors by standard deviation instead of variance. Furthermore, the use of the γ parameters puts a lower bound on variance, and thus an upper bound on precision. This ensures the activity of $\xi^{(0)}$ and $\xi^{(1)}$ to converge reasonably fast and prevents a situation in which the values of the variances (or equivalently of the standard deviations) gets close to 0 (which would make the activity of the error neuron infinitely high). The value of 1 is arbitrary, but it makes sense to assume an irreducible amount of noise in any variable processed by the brain.

We can now write

$$VFE = -\frac{1}{2}\xi^{(0)2}\sigma^{(0)2} - \frac{1}{2}\xi^{(1)2}\sigma^{(1)2} - \frac{1}{2}ln\sigma^{(0)2} - \frac{1}{2}ln\sigma^{(1)2} - ln(2\pi) \tag{2.12}$$

Inference about the most likely value of ϕ can be performed maximising VFE by gradient ascent with respect to ϕ, evaluating

$$\begin{aligned} \dot{\phi} &= \frac{\partial VFE}{\partial \phi} \\ &= \xi^{(0)} f'(\phi, \boldsymbol{\theta}) - \xi^{(1)} \end{aligned} \tag{2.13}$$

which is equivalent to the precision-weighted prediction error trade-off discussed in the previous chapter. If we think about it in neural terms, we can see that the dynamic described by this differential equation is that of a representation neuron ϕ receiving feed-forward excitatory input from a lower-level error neuron $\xi^{(0)}$ undergoing some (possibly non-linear) transformation $f'(\phi, \boldsymbol{\theta})$, and feedback inhibitory input from a higher level error neuron $\xi^{(1)}$ with synaptic strength 1.

As for the error units

$$\dot{\xi}^{(0)} = o - f(\phi, \boldsymbol{\theta}) - \xi^{(0)2}\sigma^{(0)2} \tag{2.14}$$

$$\dot{\xi}^{(1)} = \phi - \mu - \xi^{(1)2}\sigma^{(1)2} \tag{2.15}$$

which can be understood by inspecting them after convergence (i.e. setting $\dot{\xi}^{(0)}$ and $\dot{\xi}^{(1)}$ to be equal to zero) and verifying that solving for $\xi^{(0)2}$ and $\xi^{(1)2}$ gives equations 2.5 and 2.6, respectively. Thus, after convergence these neurons represent variance-weighted prediction errors, receiving feed-forward excitatory input from their associated representation neurons (o and ϕ, respectively), feedback inhibitory input from the representations neurons at the higher hierarchical level (ϕ and μ, respectively) and recurrent inhibitory input from themselves with synaptic weight equal to their associated variance. In the case presented here we are only considering two hierarchical levels, and thus only feedback connections from ϕ undergo a (possibly nonlinear) transformation as determined by $f(\phi, \boldsymbol{\theta})$. However, this model can accommodate an arbitrary number of levels, each with a different mapping function (Bogacz, 2017; Friston, 2005).

2.1.2 Learning

In addition to providing a possible biological implementation of inference about hidden states (in this case z), this scheme can accommodate learning about the model's parameters through VFE maximisation.
It can be shown that

$$\frac{\partial VFE}{\partial \mu} = \xi^{(1)} \tag{2.16}$$

$$\frac{\partial VFE}{\partial \sigma^{(0)2}} = \frac{1}{2}\left(\xi^{(0)2} - \sigma^{(0)-2}\right) \tag{2.17}$$

$$\frac{\partial VFE}{\partial \sigma^{(1)2}} = \frac{1}{2}\left(\xi^{(1)2} - \sigma^{(1)-2}\right) \tag{2.18}$$

so one could set a learning rate ω and perform the following updates after the activity of all representation and error neurons have converged:

$$\mu \longleftarrow \mu + \omega\xi^{(1)} \tag{2.19}$$

$$\sigma^{(0)2} \longleftarrow \sigma^{(0)2} + \omega\frac{1}{2}\left(\xi^{(0)2} - \sigma^{(0)-2}\right) \tag{2.20}$$

$$\sigma^{(1)2} \longleftarrow \sigma^{(1)2} + \omega\frac{1}{2}\left(\xi^{(1)2} - \sigma^{(1)-2}\right) \tag{2.21}$$

These all make sense intuitively if one thinks about the case in which no update is needed. For the mean μ, this happens when $z = \mu$ and thus the value of the hidden state is estimated to exactly correspond to expectations (and thus the prediction error $z - \mu$ is equal to zero). For variances $\sigma^{(0)2}$ and $\sigma^{(1)2}$, no update occurs when $(o - z)^2 = \sigma^{(0)2}$

and $(z-\mu)^2 = \sigma^{(1)2}$ respectively. The meaning of this becomes clear if one considers that variance corresponds to the expected value of the squared error, so when the actual squared error and estimated variance coincide no update is needed.

The update equations for $\boldsymbol{\theta}$ would depend on the form of $f(\cdot)$, and some examples are discussed in Bogacz (2017). Furthermore, the paper includes models incorporating multiple features, and thus update rules for covariances. Here we omit all this for brevity, but the same principles apply.

2.1.3 Neurobiological implementation

Figure 2.1 illustrates this proposed neurobiological implementation of predictive coding. The values of hidden states and observations are encoded by the activity of neurons, reflecting their variability, while the values of the model parameters are encoded by the strength of the synapses, reflecting their consistency over time.

Importantly, this model exhibits several properties that point to its biological plausibility. Inference is performed by the activity of interacting error and representation neurons, with the latter converging to the peak of the posterior over hidden states. The activity of all neurons is influenced only by the activity of afferent neurons and the associated synaptic weight, a property that Bogacz (2017) calls *local computation*. Learning, on the other hand, occurs by means of synaptic plasticity driven only by the activity of the pre and post-synaptic neurons, thus exhibiting a property Bogacz (2017) calls *local plasticity*. This means synaptic strength (and thus beliefs about model parameters) changes only according to information the synapse itself has "access" to. Furthermore the learning rules described here are *Hebbian*, as synaptic plasticity is influenced by the product of the activity of the pre and post-synaptic neuron (Bogacz, 2017; Friston, 2005). The model discussed so far (Friston, 2005) only considers static stimuli (i.e. observations are stable over time). A dynamic model was later developed by Friston (2008) to expand this to hidden states associated with time-varying observations. This situation does not apply to any of the experiments described in the next chapters, so the model won't be discussed in detail here. Furthermore, Bastos et al. (2012) in a seminal paper considered several structural and functional properties of the canonical cortical microcircuit, conciliating these with the dynamic model similar to the one described in Friston (2008). This work will not be fully engaged with either, as the focus of this thesis rests on mathematical aspects of probabilistic inference and learning, rather than biological ones. It is nevertheless important to mention it to point out that predictive coding is an extremely well worked-out framework from both mathematical e neurobiological perspectives.

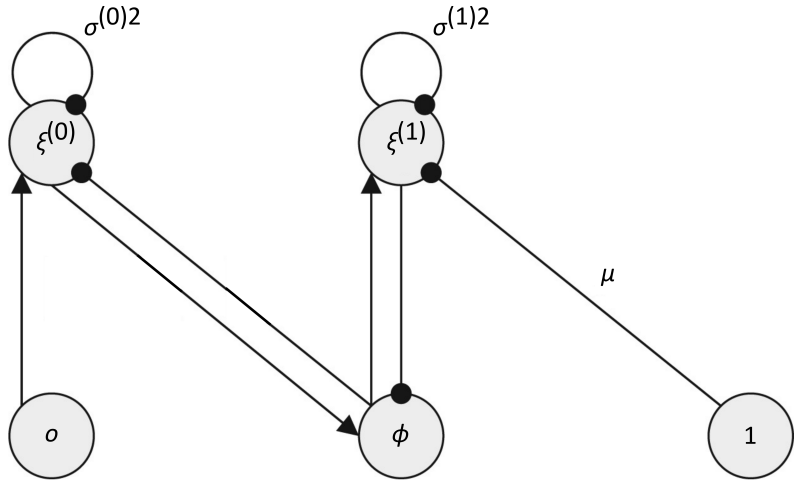

Fig. 2.1 Graphical representation of the neural model discussed in this chapter. Circles represent neurons, and lines represent synapses. Arrows represent excitatory synapses, while circles represent inhibitory ones (both arrows and circles are placed on receiving end of the synapse). Synaptic strength, when not specified, is equal to 1. Adapted from Bogacz (2017)

2.1.4 Considerations on biologically plausible models

There are fundamental differences between the models described in Chapter 1 and the ones described in this chapter. In the former, model parameters are treated as probability distributions (like hidden states), and thus have priors and their most likely value is updated to convergence. On the other hand, in the latter parameters are treated as single values and updated with a single gradient ascent step. These models thus are not strictly following variational inference where learning is concerned, as they reduce, and not minimise VFE (which is formulated as an expectation with respect to hidden states only). Therefore model parameters do not receive a fully probabilistic treatment, and neither do hidden states as, despite having priors, their posterior is approximated to be a delta distribution (i.e. only the maximum a posteriori value is estimated). This means that these models do not account for any uncertainty associated with their best guess about the value of hidden states, which is estimate with complete confidence (see Aitchison and Lengyel (2016) for an alternative neurobiologically plausible inference scheme where uncertainty is represented by neural oscillations).

In building the computational models described in Chapters 4, 5 and 6 we opted to adhere more to classical mean-field variational inference (with a minor further

approximation in Chapters 4), favouring a more fully probabilistic treatment of the variables of interest. The models discussed in this chapter (Bogacz, 2017; Friston, 2005) are however very relevant for formulating hypothesis about possible physiological markers of learning, as will become evident in the next sections.

2.2 Surprise

So far in this thesis predictive coding was presented as a family of algorithms aimed at reducing the discrepancy between some input and its predicted value. In this framework thus the fundamental goal of the brain is that of reduce (precision weighted) prediction errors, which, as discussed in the previous section, are thought to be encoded by a specific population of error neurons (Friston, 2005, 2008). Alternatively, one could say that predictive coding is an inference scheme aimed at reducing surprise. In fact, if we define surprise about a variable x as in Shannon (1948)

$$s(x) = -lnp(x) \tag{2.22}$$

and we apply it to a normally-distributed variable, we get

$$\begin{aligned} s(x \mid \mu, \sigma^2) &= -ln\mathcal{N}(x \mid \mu, \sigma^2) \\ &= \frac{1}{2} ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \end{aligned} \tag{2.23}$$

which can be minimised with respect to x by means of gradient descent, and thus updating

$$x \longleftarrow x - \Delta x \tag{2.24}$$

until convergence, with

$$\Delta x = \frac{(x-\mu)}{\sigma^2} \tag{2.25}$$

which is the familiar variance-weighted prediction error. Note that as this very simple model includes only one Gaussian (i.e. there is only one hierarchical level), it is possible to reduce prediction error all the way down to zero, which would not be the case in more complex situations.

Therefore with Gaussian generative models (which are assumed in predictive coding) minimising prediction error is equivalent to minimising surprise. This makes sense, as surprise is a measure of how unlikely an event is perceived to be, and according to predictive coding the brain is constantly adjusting its predictions and the internal

models of the world that generate them to better account for sensory stimuli (Clark, 2013). Inference (adjusting predictions) and learning (adjusting models) can thus be seen as short and long-term surprise minimisation, respectively.

Another interesting observation about surprise in normally distributed variables concerns its relationship with the Kullback-Leibler (KL) divergence. If we consider two variables x_1 and x_2 with distributions $\mathcal{N}(x_1 \mid \mu_1, \sigma_1^2)$ and $\mathcal{N}(x_2 \mid \mu_2, \sigma_2^2)$ it can be shown that

$$KL\left[\mathcal{N}(x_1 \mid \mu_1, \sigma_1^2) \,||\, \mathcal{N}(x_2 \mid \mu_2, \sigma_2^2)\right] = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 + ln\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{2.26}$$

which, like surprise, depends on a squared error divided by a variance. Keeping in mind that in probabilistic accounts of brain function variables are represented as probability distributions (Knill and Pouget, 2004), which are Gaussians in the case of predictive coding (Friston, 2005), one could see $\mathcal{N}(x_1 \mid \mu_1, \sigma_1^2)$ as a top-down prediction and $\mathcal{N}(x_2 \mid \mu_2, \sigma_2^2)$ as a bottom-up sensory signal, and their KL divergence as the information-theoretic discrepancy between the two. Surprise can then be seen in predictive coding as a measure of how a prediction is different from the actual signal it was trying to predict, and it can be easily shown in the example above that minimising the KL divergence with respect to μ_1 (i.e. the prediction) is equivalent to minimising surprise, and thus to minimising prediction error.

In a predictive coding context, considering the KL divergence instead of surprise allows a more fully probabilistic treatment of the problem at hand, as it allows to formalise both upcoming signal and top-down predictions as Gaussian distributions. However, for simplicity our models in the experiments presented in Chapters 4 and 5 do not include a low-level inference component (i.e. we are not modelling perception probabilistically), and thus perceived stimuli are represented as single real values. Therefore, in that work we opted for surprise as an inverse measure of how well a stimulus is predicted by the brain. Note that some work has been done to disentangle surprise and KL divergence in an experimental setting (Nour et al., 2018), but such distinction does not apply to the experiments described in this thesis.

If one were to track trial-by-trial surprise in an experiment, they would be able to estimate the probability participants associate to each individual stimulus. This probability is based on prior expectations, and thus measuring surprise allows to estimate the sufficient statistics of the priors, and track their changes over time. Therefore *measuring trial-by trial surprise allows to track trial-by-trial learning.*

Of course, surprise cannot be measured directly. However, if we consider the neurobiological implementation of predictive coding described above, one could reasonably expect the activity of error neurons to provide some measurable index of surprise (note that surprise and error neurons activity are monotonically related). In fact, neuroimaging studies (Alink et al., 2010; Kok et al., 2012) seem to support the presence of error signals in the brain, which can be suppressed where the stimuli are highly predictable, in line with the model outlined above.

In the next two sections we discuss two indexes of surprise, namely the mismatch negativity and pupil dilation. These are going to be the dependent variables in the empirical studies described in Chapters 4 and 5.

Note that in this section we only considered univariate distributions for simplicity and compactness, but the same principles can be extended to multivariate ones, even with non-diagonal covariance matrices.

2.2.1 The mismatch negativity

In cognitive neuroscience, one of the most practical and widely used techniques to capture brain signals during an experimental task is electroencephalography (EEG). Measuring the electrical activity of neurons with a set of electrodes placed on the participant's scalp, EEG allows to detect brain activity related to task events of particular scientific interest (e.g. the presentation of a stimulus or an action) with an extremely high temporal resolution. These electrical responses take the name of *event-related potentials* (ERP).

Decades of empirical work have allowed researchers to identify several different types of ERP *components*. These are particular portions of the recorded electrical signal that have been consistently linked with a specific cognitive phenomenon, like semantic mismatch in case of the N400 (Kutas and Federmeier, 2011) or preparation for a motor action in the case of the readiness potential (Libet et al., 1993). The abundance of scientific research on these components allows researchers to safely make relatively strong assumptions about their meaning when using them as the dependent variable in their experimental designs.

Of particular relevance for this thesis is the most common of the ERP experimental paradigms, the *oddball paradigm*. Typically this involve presenting participants with a set of stimuli, with two possible stimulus types: a more frequent one (*standard*) and a rarer one (*oddball* or *deviant*). The stimuli can be of various nature, but in the most common version of this paradigm, the *auditory oddball*, these are auditory tones varying in pitch. In this paradigm, researchers have consistently observed an

increased negativity in electrical potential after stimulus onset in deviant compared to standard stimuli (Lee et al., 2017; Schwartz et al., 2018). This ERP component takes the name of *mismatch negativity* (MMN), it typically peaks between 100 and 200 ms and is strongest in fronto-central electrodes (Wacongne et al., 2012). The MMN has been widely studied in the auditory modality, with experiments manipulating stimulus pitch (Bodatsch et al., 2011; Weber et al., 2020), but also duration (Bodatsch et al., 2011; Näätänen et al., 1989) and inter-stimulus interval (Ford and Hillyard, 1981). Furthermore, variants of the MMN have been found in the visual (Pazo-Alvarez et al., 2003), olfactory (Pause and Krauel, 2000) and somatosensory modalities (Shinozaki et al., 1998).

In sum, there is consistent empirical evidence showing that the MMN component can be reliably detected when contrasting brain responses to unlikely compared to likely events. For our purposes, a more useful way to frame the MMN is that of a response to a surprising (i.e. less likely) event, or to a mismatch between expectations (predictions) and actual stimulus (observations). This makes this component a promising candidate to track trial-by-trial surprise, and thus learning (see previous section). In fact, the MMN has been shown to scale proportionally with stimulus probability (Javitt et al., 1998; Koelsch et al., 2016), which, as discussed earlier, in monotonically related to surprise.

Another aspect that makes the MMN an attractive choice for tracking learning (for the purpose of this thesis at least) is its tight theoretical links with predictive coding. Earlier accounts (Sussman and Winkler, 2001; Winkler et al., 1996) suggested that the MMN arises from a temporo-frontal error detection system that allows the brain to adjust its internal model (*model adjustment hypothesis*). A competing theory was the *adaptation hypothesis* (Jääskeläinen et al., 2004), according to which repeated exposure to a standard auditory stimulus would cause feature-specific neurons in the auditory cortex to adapt to it and, as a consequence, suppress the N1 component, which is commonly associated with early auditory processing (Näätänen and Picton, 1987). The MMN would thus reflect a failure to suppress the N1 component in early sensory cortices, and not an internal model update in higher level brain areas. Coherently with the former, Friston (2005) presented the first predictive coding account of the MMN, suggesting that would be caused by a temporary failure to suppress prediction error (i.e. the error neurons $\boldsymbol{\xi}$ described above) in the presence of an unexpected stimulus. Later work (Garrido et al., 2009b) incorporated both the model adjustment and adaptation hypotheses within the predictive coding formulation, arguing that neither in itself was sufficient to account for the available empirical evidence. The

authors pointed at an earlier study (Garrido et al., 2008), in which model comparison on empirical data showed how a dynamical causal model (DCM) that combined the two theories into a more general predictive coding framework gave the best account for the MMN. Furthermore, Wacongne et al. (2012) built a detailed neuronal model performing predictive coding, showing how it could account for the main properties of the MMN. In this work the authors emphasised how the MMN is not driven by passive synaptic habituation (as proposed by the adaptation hypothesis), but by active predictions coming from higher cortical areas. Finally, Garrido et al. (2013) showed that the amplitude of the MMN in response to a deviant stimulus depends on the precision of the prior (i.e. the predictions), in line with the predictive coding formulation.

This body of work points at the suitability of the MMN to be used to investigate human learning in a probabilistic context. However, the classical auditory oddball task described above is almost always limited to a discrete number of stimulus types (most commonly 2, although see Garrido et al. (2013)), constraining the complexity of the applicable learning models. More specifically, discrete probability distributions make it impossible to apply predictive coding models as described in this and the previous Chapter, as they are all based on the inversion of Gaussian (which are continuous distributions) generative models. In Chapters 4 and 5 we describe an augmented version of this paradigm, where we make use of continuously-distributed auditory stimuli (varying in pitch) to investigate precision tracking (Chapter 4) and to test whether the brain can represent bimodal distributions (Chapter 5), contrary to what predictive coding would predict (see Chapter 1). In particular, we made use of the MMN in one of the experiments described in Chapter 5, using it as an index of surprise to probe the probability assigned to the stimuli, which in turn allowed us to infer what type of generative model participants were inverting throughout the experiment.

2.2.2 Pupil dilation

Another technique we made use of in the experiments discussed in this thesis is *pupillometry* (i.e. the measurament of pupil diameter). It is well-known that pupil size changes in response to luminance variations in the environment, optimising the amount of light reaching the retina (Laughlin, 1992). However, it has been shown that pupil size changes in condition of constant luminance can be related to a wide range of cognitive processes, as mental effort (van der Wel and van Steenbergen, 2018), attention (Kang et al., 2014; Wierda et al., 2012), decision-making (Cavanagh et al., 2014; de Gee et al., 2014), arousal (Reimer et al., 2014) and volatility (Browning et al., 2015), to mention a few.

Of particular significance here is the well-established link between pupil dilation in absence of luminance changes and surprise, which has been explored in several experimental paradigms. For example, both Lavín et al. (2014) and Preuschoff et al. (2011) measured pupil diameter in participants performing a gambling task in which they had to learn the reward probabilities associated with particular choices. Both studies found pupil dilation to be positively associated with surprise (or, equivalently, negatively associated with outcome probability). Furthermore, Kuchinke et al. (2007) investigated how pupil diameter varied in a lexical decision task, finding pupil dilation to negatively scale with the probability of word stimuli. Similarly, Reinhard and Lachnit (2002) found pupil dilation to be negatively associated with stimulus probability in a Go/NoGo task. The link between pupil dilation and surprise has also been found in auditory oddball experiments (Friedman et al., 1973; Hong et al., 2014; Korn and Bach, 2016; Liao et al., 2016; Murphy et al., 2011; Qiyuan et al., 1985), similarly to the MMN (see previous section).

In addition, Rajkowski (1993) showed a tight association between pupil dilation and locus coeruleus (LC, a brain area that produces noradrenaline) activity on monkeys, and later pharmacological manipulations in humans (Hou et al., 2005; Phillips et al., 2000) confirmed this, finding LC-suppressing drugs to inhibit pupil responses and LC-stimulating drugs to enhance them. LC activity has been in turn suggested to be related to surprise (Dayan and Yu, 2006), a claim supported by electrophysiological findings in monkeys (Rajkowski et al., 1994, 2004). This evidence led researchers to hypothesize that surprise-related pupil dilation is caused by noradrenergic activity in the LC (Lavín et al., 2014). Evidence for this was found in a study combining pupillometry and functional magnetic resonance imaging (fMRI), Murphy et al. (2014), in which LC activity and pupil diameter correlated in absence of any stimulation (i.e. at resting state) and responded similarly to surprising stimuli.

It should be noted that Zénon (2019) suggested a unified account of pupil dilation under constant luminance, relating the pupil response to all aforementioned cognitive phenomena to information gain, formalised as KL divergence between prior and posterior beliefs. As discussed in the previous section, the use of Gaussian distributions in our experiments does not allow to disentangle "Shannon" surprise (negative log probability, Schwartenbeck et al. (2016)) to "Bayesian" surprise (KL divergence, see Baldi and Itti (2010)).

As mentioned in the previous section, in the experiments described in Chapters 4 and 5 we extended the standard auditory oddball paradigm (which typically involves a Bernoulli distribution, i.e. a standard and a deviant stimulus type) to more complex

continuous distributions. As it has been shown that pupil size responds proportionally to probability density (Nassar et al., 2012; O'Reilly et al., 2013), we make use of pupillometry to index surprise to track learning with online probabilistic models (Chapter 4) aimed at testing and augmenting predictive coding, and the Bayesian brain framework in general.

A potential drawback of using pupil dilation as an index of surprise is the relatively slow time course of pupil responses. In fact, pupil size in response to a stimulus takes about 2 seconds to return to baseline and peaks at around 930 ms from stimulus onset (Hoeks and Levelt, 1993), making it difficult to disentangle responses to events with high temporal proximity. In Chapters 4 and 5 we address this by making use of a convolutional kernel, fitting a pupil response function (a gamma function) to each participant's pupil data to better account for individual variability (Denison et al., 2020). This, combined with a auto-regressive component to account for slow fluctuations in pupil diameter (Zénon, 2017), proved to be a very effective approach to study trial-by-trial learning (this is discussed in more detail in Chapter 4).

Chapter 3

Structure learning

3.1 General overview

In probabilistic theories of brain function perception and cognition are viewed as the result of Bayesian inference, which can take different forms (Aitchison and Lengyel, 2016; Friston and Kiebel, 2009; Gershman and Beck, 2017; Ma et al., 2006; Sanborn and Chater, 2016). Regardless of specific inference algorithms, this is achieved by inverting some generative model, and inferring the value of its parameters and latent variables (or, equivalently, hidden states) as a result. For this to happen, the specific form of the generative model must be known in advance, as in the predictive coding example outlined above. In a real-world scenario this is not realistic, and the form (or structure) of generative models must be acquired through experience. Learning (or assuming) the wrong model structure is likely to lead to suboptimal inferences (Beck et al., 2012), and thus non-adaptive behaviour. The process of learning such structure takes the name of *structure learning* (Griffiths and Tenenbaum, 2005).

In general, structure learning is seen as serving the purpose of acquiring internal models of the environment, thus simplifying and making sense of the wide variety and volume of observations one can be exposed to and allowing useful generalisations (Braun et al., 2010). However, internal models of the world can be very complex, which makes studying structure learning as a whole very challenging. In practice, researchers focus on a particular aspect of it. What follows is a brief overview (by no means exhaustive) of some of these different aspects.

Representation learning *Feature* or *representation learning* (Austerweil and Griffiths, 2008; Wu et al., 2021), consists in reducing highly dimensional observations to limited number of useful and interpretable features. These features (or representations)

can then be re-used in different contexts facilitating learning in different domains, a phenomenon called *transfer learning* (Menghi et al., 2021) or *learning to learn* (Braun et al., 2010).

Clustering To be able to generalise some knowledge acquired during a particular experience, one must be able to correctly infer what other contexts that experience is relevant to. To achieve this, humans spontaneously divide their observations into categories, in a process called *clustering* (Dasgupta and Griffiths, 2021; Sanborn et al., 2006), which is the main topic of Chapter 6 and will be discussed in greater detail in later sections.

Concept Learning The idea of *concept learning* (Lake et al., 2015; Smith et al., 2020) partially overlaps with clustering and representation learning, and can be seen as the most abstract version of both (i.e. learning about abstract categories and features). An interesting line of research in the field is concerned with studying how humans organise concepts in cognitive maps (Constantinescu et al., 2016) and learn their relational structure (Mark et al., 2020; Whittington et al., 2020).

Causal learning Finally, another form structure learning can take is *causal learning* (Gershman et al., 2017; Griffiths and Tenenbaum, 2005; Tenenbaum and Griffiths, 2001), consisting in learning the causal structure of a set of events (i.e. what causes/influences what) and the strength of these causal relationships.

3.2 Model comparison vs incremental models

Let's consider the simple case of an agent making T multivariate observations $\{\mathbf{o}_1, ..., \mathbf{o}_T\}$ and trying to infer the corresponding hidden states $\{\mathbf{z}_1, ..., \mathbf{z}_T\}$. As discussed, this requires a generative model with parameters $\boldsymbol{\theta}$. In absence of prior knowledge about the structure of this model, this too must be learned from the observations $\{\mathbf{o}_1, ..., \mathbf{o}_T\}$.

One way to achieve this is to generate a set of N hypotheses $\{\mathcal{G}_1, ..., \mathcal{G}_N\}$ about the form of the generative model, having thus a set of candidate models to compare and choose from. This poses the problem of the potentially infinite number of possible models. For a biological agent with limited cognitive resources, it would be necessary to have a contained hypothesis space. This can be solved by relying on more abstract, generalisable structural assumptions that constrain the process of hypothesis generation (Tenenbaum et al., 2011). This has been formalised with hierarchical Bayesian models,

which use a finite set of qualitatively different "building blocks" to generate the a set of hierarchically organised candidate models(Kemp et al., 2010, 2007; Kiebel et al., 2008; Lucas and Griffiths, 2010; Tenenbaum et al., 2006). In humans, hypothesis generation has been suggested to be influenced by the plausibility a certain model is believed to have (i.e. its prior probability), which has been shown to explain a number of cognitive biases (Dasgupta et al., 2017).

Once the agent has generate a set of models to evaluate (each with its own set of parameters $\boldsymbol{\theta}_n$, the number of which can vary between models), it can simply perform model comparison and select the winning model. There are several ways this can be achieved mathematically (Ward, 2008), but all these methods involve fitting parameters $\boldsymbol{\theta}_n$ and hidden states $\mathbf{z}$ to the observations $\mathbf{o}$ for each candidate model $\mathcal{G}_n$. Then a model-specific score can be obtained, involving a fitness metric to reward accuracy (e.g. log likelihood) and a complexity penalty to avoid overfitting (e.g. number of parameters). Additionally, as mentioned above, one could have a prior over the model space, so each candidate model is assigned a prior probability, which can be understood as a measure of how plausible that model (or, equivalently, hypothesis) is thought to be. This prior on models can influence both hypothesis generation (Dasgupta et al., 2017) and model comparison (Tenenbaum et al., 2006).

Once model comparison has been performed, the agent can select the model with the highest score and choose its policies accordingly.

Structure learning as model comparison has been the focus of interesting theoretical work (Friston et al., 2021; Kemp and Tenenbaum, 2008; Tenenbaum et al., 2011), and simulations have been shown to correctly predict human behaviour in structure learning tasks (Gershman, 2017; Kemp et al., 2010).

Despite this, this approach presents some major limitations. In fact, most of the aforementioned models assume structure learning to happen *offline*, meaning only after a certain number of observations have taken place. This is not only suboptimal in terms of immediate action selection in response to each observation (the agent is not performing inference *online*), but it also requires the agent to remember each individual observation (or at least a great number of them, especially for complex models). This does not mean it would be impossible to build an online model based on model comparison, but this would involve updating several models each time a new stimulus is encountered. In addition, if the agent had to predict the consequences of a particular policy in response to a stimulus, it would have to take the expected reward value over the whole model space (Gershman, 2017). This might be feasible when the hypothesis space is limited (Gershman, 2017; Tomov et al., 2018) (or when all

observations are presented at the same time or are available to be somehow revisited at will), but would put a biological agent living in complex and volatile environments under considerable cognitive strain, requiring considerable computational resources even for relatively simple problems.

Furthermore, the use of offline models does not allow to fit them to human behaviour (or physiological/brain data) based on online learning. In other words, these models cannot be used to study human structure learning on a trial-by-trial basis, and require a training phase followed by a test phase to compare model and human outputs, as in Kemp et al. (2010), Gershman (2017) and Orbán et al. (2008).

An alternative to generating different candidate models and comparing them after a certain number of observations (or after each observation) is to start with a single, very simple model, and augmenting it when necessary as new data are collected. In other words, as a new observation comes along which is not properly accounted for, the agent can increase the complexity of the generative model. This has the advantage of being computationally parsimonious, as the model is kept as simple as possible. It also happens *online*, making it more convenient where immediate responses to new stimuli are required, and does not require the agent to hold in memory a great number of observations. We call this type of models *incremental models*.

This type of approach has been deployed by Wu et al. (2021) to study representation learning for sequence, image, video and language data. Their simulations showed that learning a hierarchy of "chunks" (i.e. frequently occurring patterns) results in interpretable representations of the stimuli that can be reused in future tasks (i.e. they make transfer learning possible). Their model learned representations starting from very simple sequences, combining them hierarchically as new information became available. Similarly, (Gershman et al., 2010) used an incremental approach to give a structure learning-based interpretation of classical conditioning, suggesting animals infer *latent causes* behind the association between conditioned and unconditioned stimuli. Crucially, the number of possible latent causes is unknown, so the animal's internal models are posited to start simple and grow according to the complexity of the data. The same modelling framework was used also used to study transfer learning and generalisation in humans by Collins and Frank (2013, 2016). Furthermore, the idea that biological agents adopt (although not necessarily exclusively) incremental models for structure learning supported by evidence from animal studies (Wang et al., 2011), suggesting a link between increasingly complex internal models and structural brain growth.

On the other hand, it is worth considering the case of an agent mistakenly building an overly complex model (or transitioning to a simpler environment), which has not received as much attention from researchers. In such a situation it would be beneficial for the agent to prune away the unnecessary components of its internal models. This has been addressed by Smith et al. (2020), who built a model of concept learning capable of both increasing and decreasing its complexity.

The family of models whose complexity adapts to that of the data are known as *Bayesian non-parametric models* (Gershman and Blei, 2012), and will be discussed more in detail in the next section. In this thesis we will limit further discussion of structure learning to this modelling framework, as it is the one on which the simulation work described in Chapter 6 is based.

3.3 Clustering and the Chinese restaurant process

In this thesis we focus on a specific form of structure learning, namely online *clustering*, defined as unsupervised grouping of stimuli or events into useful categories. In other words, we discuss how an agent can augment its generative model online, adding new components (clusters) when necessary.

Let's again consider an agent making a series of observations $\{\mathbf{o}_1, ..., \mathbf{o}_T\}$. With no structural knowledge about their probability distribution, the agent has two "model-free" options: consider each observation as a completely separate entity, unrelated to any other, or consider all observations part of a common category, generalising whatever it can learn from one to all others. Both of these are clearly suboptimal in most situations. The former does not allow any generalisation whatsoever (e.g. learning that an individual lion is dangerous tells me nothing about other lions), and the latter leads to an over-generalisation (e.g. learning that an individual lion is dangerous tells me that all animals are). Obviously a more flexible, model-based solution is required.

The agent can assume that observations can be grouped into M clusters, so that knowledge about a member of a cluster can be generalised to all members of that cluster (e.g. learning that an individual lion is dangerous tells me that all lions are dangerous). If the value of M is known (and the shape of each individual cluster is assumed, e.g. Gaussian), clustering is a fairly straightforward problem and can be solved with Bayesian model inversion, similarly to what discussed in previous Chapters. In such a case, the structure of the generative model is known in advance (or assumed), and does not need to be learned. Conversely, if the number of clusters M is unknown, so is the structure of the generative model. In this case, the agent must learn the number

of clusters to build an effective model of its environment and generalise information about individual observations appropriately.

The most popular way to do this is using a method called the *Chinese restaurant process* (CRP), which belongs to the family of Bayesian non-parametric models (Gershman and Blei, 2012). Here for each new observation $\mathbf{o}_t$ the agent must infer its cluster assignment c_t (which us not directly observable, and is thus treated as an hidden state). It thus has to evaluate

$$p(c_t \mid o_t) \propto p(o_t \mid c_t)p(c_t) \tag{3.1}$$

where

$$p(c_t = n) = \begin{cases} \frac{v_n}{t-1+\alpha} & if \quad n \leq N \\ \frac{\alpha}{t-1+\alpha} & if \quad n > N \end{cases} \tag{3.2}$$

Here v_n is the number of observations previously assigned to cluster n, N is the number of clusters for which $v_n > 0$ and α is a concentration parameter regulating the agent's tendency to form new clusters.

Put more simply, the agent has popularity-based priors on cluster assignment, with popular clusters (i.e. clusters to which many previous observations were assigned) being considered a priori more likely. This implies that, as more stimuli are observed, the chances of them belonging to a previously unseen cluster decreases. This property makes psychological sense: after many years doing bird-watching, it will become less and less likely for me to spot a new species of bird (unless I travel to an unknown environment). Of particular importance is the concentration parameter α, which regulates how "conservative" the agent is. For high values of α the agent is going to be more likely to create a lot of clusters, whilst for low values it going to prefer a simpler model. At the two extremes, the agent is going to form a separate cluster for each observation (extremely high α) or assign all observations to a single cluster (extremely low α), reverting to the "model-free" approaches described above.

Therefore, one could see α as an index of how complex an agent is allowing its internal models to be, or, equivalently, of the cognitive resources the agent is willing to allocate to structure learning. This aspect has been tackled by Dasgupta and Griffiths (2021), who showed that placing a prior on cluster mappings penalising model complexity is equivalent to using CRP priors. Model complexity was formalised as the entropy $\mathcal{H}[p(\mathbf{c})] = \sum_{n=1}^{N} p(c_n) ln p(c_n)$ of the cluster mappings, a measure of the representational cost of the distribution $p(\mathbf{c})$ (Shannon, 1948).

The CRP has been deployed in some form in the study of structure learning (and, in particular, clustering) in a variety of contexts. For example, Gershman et al. (2017)

applied it to study social influence, and found that it was based on a cluster-like representation of individual preferences. The CRP was also used to build a latent cause theory of Pavlovian conditioning (Gershman et al., 2010; Gershman and Niv, 2012; Gershman et al., 2015b), , suggesting that animals assume their experiences to be influenced by hidden causes. Animals would thus cluster their experiences into a discrete set of hidden causes, and, as a new event comes along, inferring which hidden cause was behind it would determine which cluster needs to be updated. The authors managed to replicate a range of experimental findings on classical conditioning (e.g. conditioned response acquisition,extinction, renewal) with simulations based on this principle (Gershman et al., 2010; Gershman and Niv, 2012; Gershman et al., 2015b).

Collins and Frank (2013) built a CRP-based model of the interaction between executive functions and learning, and used to show that human participants spontaneously learn and generalise latent task rules (which can be seen as latent causes regulating the relationship between actions and outcomes), even when not cued to do so. They later used a similar models and replicated their findings, backing them up with electrophysiological eveidence (Collins and Frank, 2016).

Despite this relatively well-established modelling framework, and the aforementioned benefits of incremental models (namely being online) for studying human cognition, very little work has been done to apply these to investigate trial-by-trial structure learning. In fact, similarly to the model comparison framework discussed above, work with incremental models in the cognitive sciences have largely been limited to simulations (Dasgupta and Griffiths, 2021; Franklin and Frank, 2018) and comparing their outputs with existing (Gershman et al., 2010; Gershman and Niv, 2012; Gershman et al., 2015b) or new (Gershman et al., 2017) experimental data. There have only been a handful of empirical studies fitting these models to trial-by trial participants data (Collins and Frank, 2013, 2016; Davis et al., 2012), leaving this avenue largely unexplored.

Unfortunately, the pandemic did not allow us to directly tackle this with an empirical study. We did however put the basis for future experimental work by building a novel clustering task loosely based on Davis et al. (2012), as well as a novel clustering model with a strong focus on intelligent use of computational resources (in this case working memory), or computational rationality (Gershman et al., 2015a). Our model can be fit to predict online human behaviour, and provides a series of trial-by-trial metrics that could prove instrumental for future neuroimaging studies. Finally, our simulations provide a proof of concept for the importance of revisiting and re-interpreting past events stored working memory (i.e. making inferences about the past) for building optimal internal models.

Chapter 4

Pupil dilation indexes automatic and dynamic inference about the precision of stimuli distributions

4.1 Abstract

Learning about the statistics of one's environment is a fundamental requirement of adaptive behaviour. In this experiment we probe whether pupil dilation in response to brief auditory stimuli reflects automatic statistical learning about the underlying stimulus distributions. Specifically, we consider whether pupil dilation reflects automatic (task-irrelevant) learning about the precision of Gaussian distributions of pitch in a sequence of tones. We provide clear evidence, both by comparing responses to perceptually identical probe tones in low and high precision blocks, and using a novel model-based analysis, that subjects did indeed track the precision of the stimulus distribution. This extends previous work looking at electrophysiological effects of precision (or, equivalently, variance) learning, and provides new evidence that the putatively noradrenergic processes underlying pupil dilation reflect rapidly updated information about distributions of sensory stimuli. In addition, our study represents a validation of our model-based approach to analysing pupillometry data, which we believe has considerable promise for future studies.

4.2 Introduction

Sensitivity to the statistics of one's environment is a key requirement for adaptive behaviour. Correspondingly, statistical learning has been a fertile area of research (Kirkham et al., 2002; Saffran et al., 1996, 1999), often in conjunction with probabilistic theories of cognition (Fiser et al., 2010; Tenenbaum et al., 2011; Turk-Browne et al., 2010). Here, we consider statistical learning to solve the specific problem of forming beliefs about the precision (or, equivalently its inverse, the variance) of the underlying probability distributions that govern incoming sensory stimuli. From a normative perspective, accurately estimating precision is extremely important, as it governs key features of learning and inference, such as how quickly to update one's beliefs (Behrens et al., 2007; Mathys et al., 2011) and how to weight incoming sensory information against prior beliefs (Friston, 2008), and aberrant precision estimation is widely believed to play a key role in psychopathology (Adams et al., 2013; Fletcher and Frith, 2009; Friston et al., 2014; Lawson et al., 2014).

To explore learning about precision, we make use of the widely replicated finding that non-luminance related pupil dilation indexes the surprise associated with incoming sensory stimuli (Alamia et al., 2019; Damsma and van Rijn, 2017; De Berker et al., 2016; Friedman et al., 1973; Kloosterman et al., 2015; Lavín et al., 2014; Nassar et al., 2012; O'Reilly et al., 2013; Preuschoff et al., 2011; Qiyuan et al., 1985; Raisig et al., 2010; Reinhard and Lachnit, 2002). Note that here we define surprise as the negative log probability of an event occurring, though see Baldi and Itti (2010); Schwartenbeck et al. (2016); Zénon (2019) for an important alternative. This permits one to make inferences about participants' implicit beliefs about the statistics of their environment, without the necessity for an explicit probe or decision, and thus provides a mean to characterise statistical learning processes (Alamia et al., 2019; Vincent et al., 2019). Specifically, where predictions are more precise, subjects should be more surprised by objectively identical stimuli, leading to a greater dilation response.

Learning about precision has previously been tested using reaction time and magnetoencephalography (MEG) data by Garrido et al. (2013). Here, subjects performed a modified version of the auditory oddball task, where on separate blocks tones were drawn from either high or low precision Gaussian distributions in log frequency space (Fig. 4.1). The present work seeks to extend this approach, using pupillometry data. This is important, given the tight link between pupil dilation and noradrenergic activity in the locus coeruleus (Murphy et al., 2011), since it provides the opportunity to better understand the function of the noradrenergic system (Dayan and Yu, 2006). Pupillometry data is also much cheaper and easier to acquire than neuroimaging data,

and its use to characterise statistical learning in different groups (Browning et al., 2015; Montague et al., 2012) is thus attractive from a practical perspective.

We thus collected pupillometry data whilst subjects performed a slightly modified version of the paradigm used by Garrido et al. (2013). We used the data to perform two types of analysis. The first was closely modelled on that used in previous work (Garrido et al., 2013), and directly compared responses to probe tones during high and low precision blocks. Our main hypothesis was that a deviant sound (probe tones at 2000 Hz) would be more surprising, and therefore elicit a bigger pupil dilation, in high precision compared to low precision blocks, as the associated probability density would be lower (Fig. 4.1). In the second we combined agent-based modelling (O'Doherty et al., 2007) with a novel convolution-based approach to analysing pupillometry data (Denison et al., 2020), which we believe has considerable promise for exploring automatic statistical learning in humans. We hypothesised that pupil dilation responses would reflect dynamic updating of beliefs about the precision of the stimulus distribution.

4.3 Methods

4.3.1 Participants

16 participants (12 females), aged 18 to 34 (mean = 21.1) took part in the study and they all gave informed consent.

4.3.2 Stimuli and Procedure

Participants were asked to look at a fixation cross in the centre of a computer screen while listening to a series of tones through headphones. Their task consisted only in pressing the space bar when the sound came only from one of the two headphones' speakers (i.e. when they heard it coming only from one side). Importantly, the pitch of the tones was entirely task irrelevant, deconfounding outlier and target (unilateral) tones, which is a concern with several previous studies (Hong et al., 2014; Liao et al., 2016), as well as indexing automatic, rather than task dependent statistical learning processes. Furthermore, the absence of a visual target avoided luminance changes that could alter pupil diameter.

The experiment was divided into 4 sessions, during each of which subjects were presented with 800 pure tones, each lasting 50 ms, with an interstimulus interval of one second. 4 blocks (2 high precision blocks and 2 low precision blocks) were present in each session, with 200 tones each and no breaks between blocks. The order of the

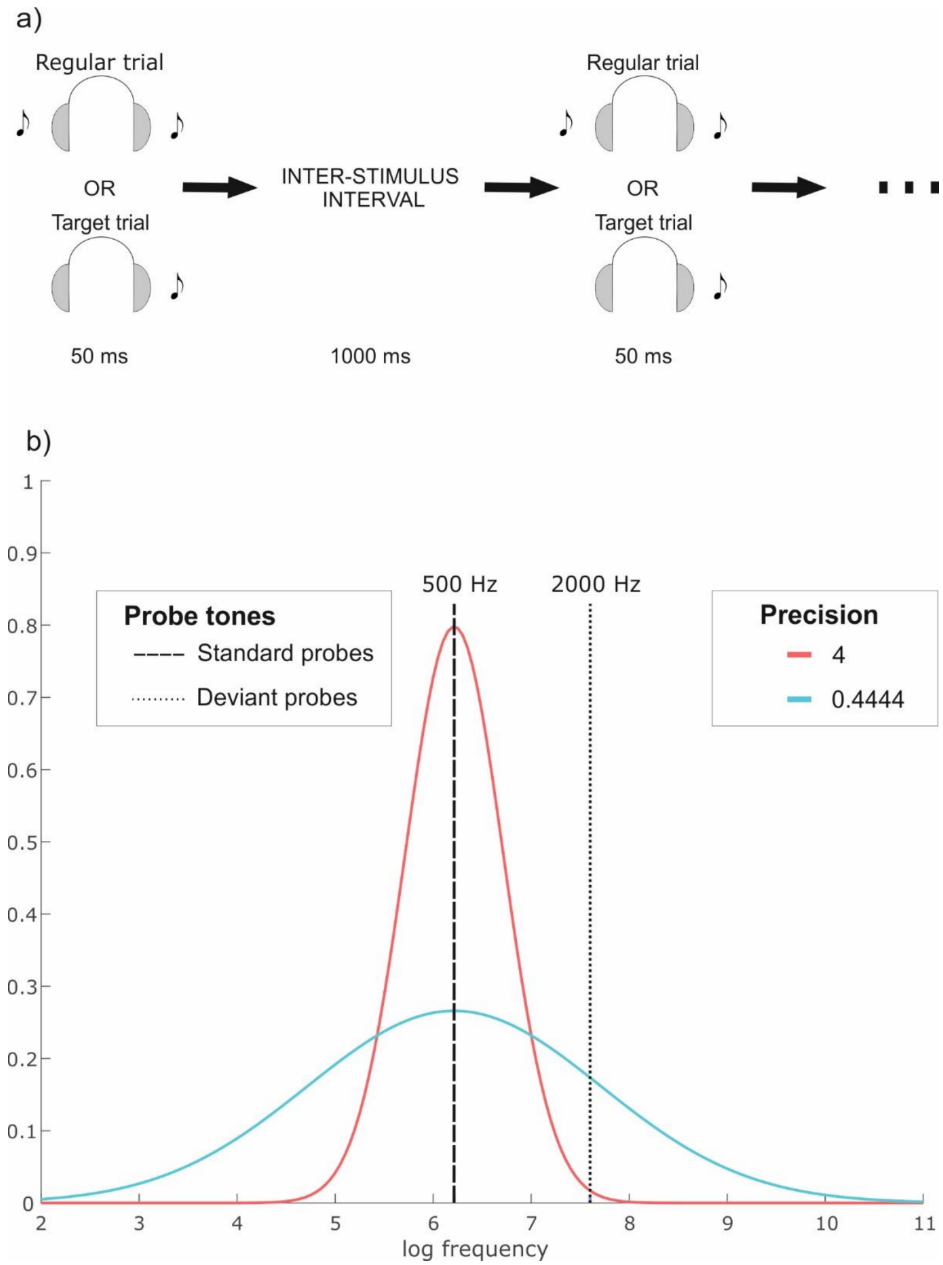

Fig. 4.1 Illustration of the experimental paradigm (a). Participants were exposed to a series of tones (800 per session, 3200 in total) and were asked to press the space bar when they heard the sound coming only from one speaker (i.e. only from one side). The pitch of the tones was sampled from two different probability distributions (b), alternating between high and low precision blocks. In line with previous work (Garrido et al., 2013) probes were added at 500 Hz and 2000 Hz, slightly distorting the distribution.

blocks was counterbalanced and participants were not aware of the presence of different blocks within each session. Stimuli were selected to be similar to those used in a prior

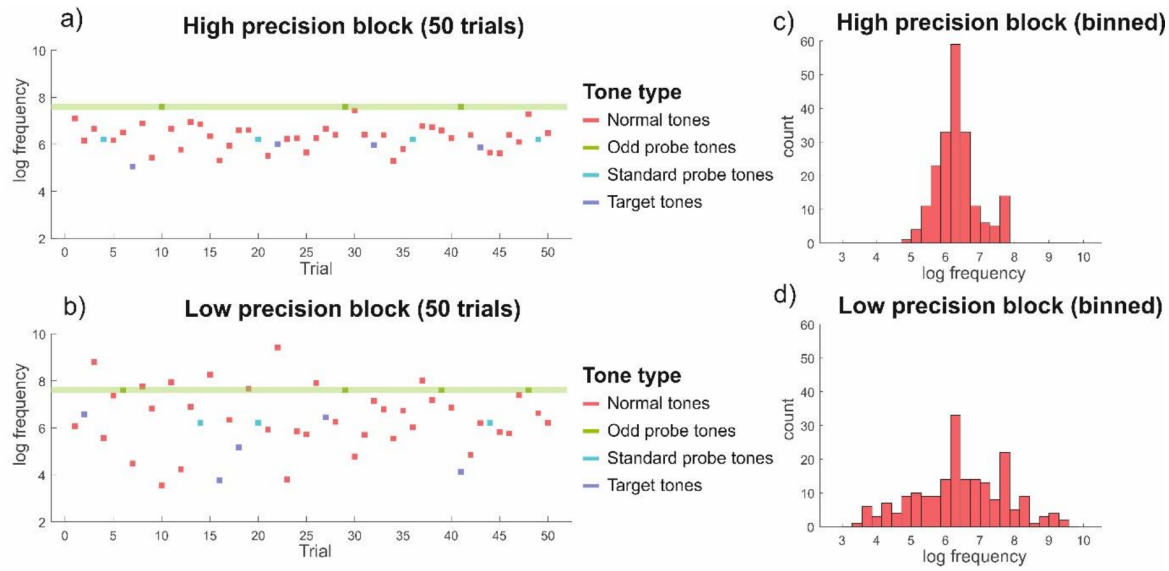

Fig. 4.2 Graphical representation of 50 trials (corresponding to 50 tones) in a high (a) and low (b) precision block. Deviant probes (2000 Hz) are highlighted in green to evidence how they stand out more in high compared to low precision blocks. These and standard probes (500 Hz) slightly distorted the stimuli distribution, as shown in binned distributions (1 bin = 2/7 of an octave) of a sample high (c) and low (d) precision block (200 trials per block).

study (Garrido et al., 2013). Specifically, for each session the frequency of 688 out of 800 tones was sampled from a Gaussian distribution in log-frequency space, with mean $\mu = 500$ Hz and standard deviation $\sigma_h = 0.5$ octaves for high precision blocks and σ_l =1.5 octaves for low precision blocks. (These values correspond to precisions of 4 and 0.4444 respectively). Out of the 112 remaining tones, 56 were standard probes (500 Hz, corresponding to the mean of the distribution) and 56 were deviant probes (2000 Hz, two octaves above the mean), which slightly distorted the probability distribution, adding two point-masses of 7% probability each. The number of unilateral, target tones varied across sessions (82, 80, 75, and 78 respectively). Both probes and targets were pseudo-randomly inserted in the stream, and targets were made to never coincide with a probe, or occur immediately after it. This was done because targets are very likely to elicit a strong, long lasting pupil response, which would confound the effect of surprise.

See Fig. 4.2 for a graphical representation of the tone sequence in low and high precision blocks, and of how probe tones distorted the distributions.

4.3.3 Pupillometry data recording and preprocessing

Pupillometry data were recorded at 500 Hz using an EyeLink 1000 eye-tracking device whilst subjects sat in a moderately lit room (Fig. 4.3 shows 4 seconds of raw pupil size data). As we were interested in learning effects rather than precise psychophysics, we did not directly measure the luminance of the screen or the fixation cross. Similarly, tones were presented at a constant volume, at a level that was comfortable but clearly audible for subjects, but these levels were not recorded. Critically, because these quantities were held constant for each subject throughout the duration of the experiment, purely physical properties of the stimuli or environment cannot explain the results we describe below.

Linear interpolation was used to remove artefacts relating to eyeblinks and saccades, and the data were low pass filtered at 20 Hz. Data were recorded from both eyes simultaneously, and averaged prior to further analysis. Additionally, for the model-based analysis, time series were downsampled to 50 Hz, and were then mean-corrected and normalised to unit variance, to standardise responses across subjects.

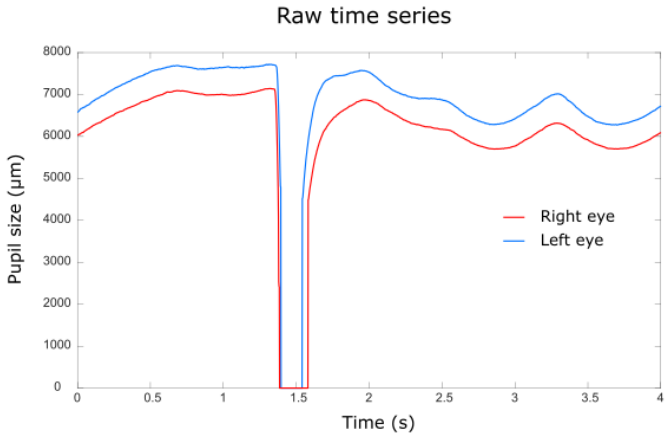

Fig. 4.3 4 seconds of raw pupil size data from both eyes from a single participant (subject 2). This is a representative sample of how the data looked before any preprocessing took place. Most noticeably, this time window includes an eye blink (with pupil size values briefly falling to zero).

4.3.4 Probe tones analysis

We first analysed our data using a classical model-free approach, where we averaged the responses to the probe tones in each precision condition for each subject. We then averaged all data-points from 900 ms to 1000 ms after stimulus onset and performed

a two-way repeated measures ANOVA with precision (high vs low) and probe type (deviant vs standard) as within-subjects factors. The 900-1000 ms time window was averaged to avoid multiple comparisons while trying to capture the peak of the pupil response, which typically occurs around 930 ms from stimulus onset (Hoeks and Levelt, 1993).

4.3.5 Model-based analysis

4.3.5.1 Time series analysis

Building on previous work analysing fMRI time series (Penny et al., 2003), we developed a model for analysing pupillometry data combining a General Linear Model (GLM) incorporating a convolution kernel, and an autoregression (AR) component, in order to model fluctuations not captured by the convolution model (which could come from a variety of sources) (Zénon, 2017). For simplicity, we only consider a first order AR model, but this approach can naturally be extended to include higher orders (Penny et al., 2003). For related approaches, see Korn and Bach (2016); Vincent et al. (2019); Zénon (2017).

In this GLM-AR model, data $\mathbf{z} = \{z_1, ..., z_T\}$ is modelled in terms of a $[T \times K]$ design matrix $\mathbf{X}$, a $[K \times 1]$ vector of regressor coefficients $\mathbf{w}$ and a $[T \times 1]$ vector of errors $\mathbf{e}$:

$$\mathbf{z} = \mathbf{X}\mathbf{w} + \mathbf{e} \tag{4.1}$$

with

$$e_t = \begin{cases} a(z_{t-1} - \mathbf{x}_{t-1}\mathbf{w}) + \iota_t & \text{if} \quad t > 1 \\ r & \text{if} \quad t = 1 \end{cases} \tag{4.2}$$

Here e_t is modelled as a combination of the "prediction error" at the previous time point, weighted by the AR coefficient a, and ι_t which is drawn from $\boldsymbol{\iota}$, a $[T \times 1]$ vector of Independent Identically Distributed (IID) errors. The model also includes an additional session-specific parameter r to model the error at $t = 1$.

The design matrix $\mathbf{X}$ is generated by convolving a $[T \times K]$ input matrix $\mathbf{U}$ with a gamma kernel to model the slow time course of pupil responses (Denison et al., 2020; Hoeks and Levelt, 1993; Korn and Bach, 2016). The kernel is parameterised using three parameters: shape (h) and scale (l) parameters governing the properties of the Gamma distribution, and a delay parameter (d) introducing a temporal delay. Thus:

$$\mathbf{x}_k = \mathbf{u}_k * g \tag{4.3}$$

with

$$g = \Gamma(t - d \mid h, l) \tag{4.4}$$

where Γ indicates the probability density function of the gamma distribution. Our approach thus naturally accounts for between-subject variability in the time course of pupil responses, since the shape of the gamma distribution is fitted to individual responses (Denison et al., 2020). Note that in our experiment there were no luminance changes, and we thus needed only to model pupil dilation responses. The GLM-AR approach is equally applicable to data involving pupil constriction (Korn and Bach, 2016), and we will consider this in future work.

4.3.5.2 Cognitive modelling

To model cognitive processes during the task, we devised a simple agent where at each trial i, trial-specific mean (m_i) and log-precision (k_i) estimates were generated using a predictive coding algorithm (Friston and Kiebel, 2009; Rao and Ballard, 1999; Shipp, 2016; Spratling, 2017), augmented to allow for dynamic estimation of precision. This does not imply any strong claim about the actual neuronal inference mechanisms employed by subjects on the task, since other schemes could undoubtedly provide similar predictions (Aitchison and Lengyel, 2016; Friston et al., 2017; Ma et al., 2006). It does, however, provide a simple and parsimonious method for modelling task performance, which is grounded in computational modelling of cortical function.

In this model, the agent infers on the log joint probability of current hidden states and observations, given all current and previous observations $\mathbf{y}_{1:i} = \{y_1, ..., y_i\}$, prior beliefs about the initial hidden states (m_0 and k_0) and two fixed parameters which govern how quickly the mean ($\eta^{(m)}$) and log precision ($\eta^{(k)}$) are expected to change, often known as their volatility (Behrens et al., 2007; Mathys et al., 2011). Note that these parameters were fitted to individual subjects' data, meaning that we can capture a broad spectrum of beliefs about volatility.

Assuming the Markov property, we can write

$$lnp(m_i, k_i \mid \mathbf{y}_{1:i}, m_0, k_0, \eta^{(m)}, \eta^{(k)}) = E_{lnp(m_{i-1}, k_{i-1})}\left[p(m_i, k_i \mid y_i, m_{i-1}, k_{i-1}, \eta^{(m)}, \eta^{(k)})\right] \tag{4.5}$$

and estimated recursively, which represents standard Bayesian filtering. Here $E_{h(\theta)}[f(x \mid \theta)]$ represent the expected value of $f(x \mid \theta)$ under $h(\theta)$, so that

$$\begin{aligned} E_{h(\theta)}[f(x \mid \theta)] &= \int f(x \mid \theta) h(\theta) d\theta \\ &= f(x) \end{aligned} \tag{4.6}$$

The model has the conditional independence properties that

$$p(y_i \mid \mathbf{y}_{1:i-1}, \mathbf{m}_{1:i}, \mathbf{k}_{1:i}, m_0, k_0, \eta^{(m)}, \eta^{(k)}) = p(y_i \mid m_i, k_i) \tag{4.7}$$

and

$$p(m_i, k_i \mid \mathbf{y}_{1:i-1}, \mathbf{m}_{1:i-1}, \mathbf{k}_{1:i-1}, m_0, k_0, \eta^{(m)}, \eta^{(k)}) = p(m_i \mid m_{i-1}, \eta^{(m)}) p(k_i \mid k_{i-1}, \eta^{(k)}) \tag{4.8}$$

meaning that observations depend only on the current hidden states, and each sequence of hidden states is a separate Markov chain.

More specifically, each continuous-valued observation y_i is sampled from a normal distribution with mean m_i and log-precision k_i

$$p(y_i \mid m_i, k_i) = \mathcal{N}(y_i \mid m_i, e^{-k_i}) \tag{4.9}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

Similarly, the mean and log-precision of the distribution are treated as independent, zero-mean, Gaussian random walks, with fixed precisions (volatilities), given by

$$p(m_i \mid m_{i-1}, \eta^{(m)}) = \mathcal{N}(m_i \mid m_{i-1}, \eta^{(m)-1}) \tag{4.10}$$

$$p(k_i \mid k_{i-1}, \eta^{(k)}) = \mathcal{N}(k_i \mid k_{i-1}, \eta^{(k)-1}) \tag{4.11}$$

The log joint distribution can now be written as:

$$\begin{aligned} lnp(y_i, m_i, k_i \mid \mathbf{y}_{1:i-1}, m_0, k_0, \eta^{(m)}, \eta^{(k)}) = & lnp(y_i \mid m_i, k_i) + lnE_{p(m_{i-1})}\left[p(m_i \mid m_{i-1}, \eta^{(m)})\right] \\ & + lnE_{p(k_{i-1})}\left[p(k_i \mid k_{i-1}, \eta^{(k)})\right] \end{aligned} \tag{4.12}$$

To avoid the need to estimate joint probability distributions, the agent performs variational inference (Beal, 2003), and approximates the log joint with a distribution

$q(m_i, k_i)$ which factorises such that:

$$q(m_i, k_i) = q(m_i)q(k_i) \tag{4.13}$$

This gives a variational lower bound on the log model evidence

$$VFE = E_{q(m_i,k_i)}\left[ln\left(\frac{p(y_i, m_i, k_i \mid \mathbf{y}_{1:i-1}, m_0, k_0, \eta^{(m)}, \eta^{(k)})}{q(m_i, k_i)}\right)\right] \tag{4.14}$$

To generate a predictive distribution over m_i, the agent uses the variational posterior $q(m_{i-1})$ generated on the previous trial. Thus

$$\begin{aligned} p(m_i \mid \mathbf{y}_{1:i-1}, m_0, \eta^{(m)}) &\approx E_{q(m_{i-1})} p(m_i \mid m_{i-1}, \eta^{(m)}) \\ &= \mathcal{N}(m_i \mid \tilde{\mu}_i^{(m)}, \tilde{\tau}_i^{(m)-1}) \end{aligned} \tag{4.15}$$

where

$$\tilde{\mu}_i^{(m)} = \breve{\mu}_{i-1}^{(m)} \tag{4.16}$$

and

$$\tilde{\tau}_i^{(m)} = \left(\eta^{(m)-1} + \breve{\tau}_{i-1}^{(m)-1}\right)^{-1} \tag{4.17}$$

where $\breve{\mu}_{i-1}^{(m)}$ and $\breve{\tau}_{i-1}^{(m)}$ are the sufficient statistics (mean and precision) of $q(m_{i-1})$, and $\tilde{\mu}_i^{(m)}$ and $\tilde{\tau}_i^{(m)}$ are the sufficient statistics of the predictive distribution. (At $i = 1$, $\tilde{\mu}_1^{(m)} = m_0$ and $\tilde{\tau}_1^{(m)} = \eta^{(m)}$).

Similarly, the predictive distribution over k_i used by the agent is given by

$$\begin{aligned} p(k_i \mid \mathbf{y}_{1:i-1}, k_0, \eta^{(k)}) &\approx E_{q(k_{i-1})} p(k_i \mid k_{i-1}, \eta^{(k)}) \\ &= \mathcal{N}(k_i \mid \tilde{\mu}_i^{(k)}, \tilde{\tau}_i^{(k)-1}) \end{aligned} \tag{4.18}$$

where

$$\tilde{\mu}_i^{(k)} = \breve{\mu}_{i-1}^{(k)} \tag{4.19}$$

and

$$\tilde{\tau}_i^{(k)} = \left(\eta^{(k)-1} + \breve{\tau}_{i-1}^{(k)-1}\right)^{-1} \tag{4.20}$$

(at $i = 1$, $\tilde{\mu}_1^{(k)} = k_0$ and $\tilde{\tau}_1^{(k)} = \eta^{(k)}$).

Thus for the agent

$$\begin{aligned} VFE =& E_{q(m_i,k_i)}\left[ln\mathcal{N}(y_i \mid m_i, e^{-k_i})\right] + E_{q(m_i)}\left[ln\mathcal{N}(m_i \mid \tilde{\mu}_i^{(m)}, \tilde{\tau}_i^{(m)-1})\right] \\ &+ E_{q(k_i)}\left[ln\mathcal{N}(k_i \mid \tilde{\mu}_i^{(k)}, \tilde{\tau}_i^{(k)-1})\right] + \mathcal{H}(m_i) + \mathcal{H}(k_i) \end{aligned} \tag{4.21}$$

where $\mathcal{H}(\cdot)$ denotes the entropy of a probability distribution.

The optimal solution $q^*(\cdot)$ for variable m_i can now be derived simply using the properties of the Gaussian distribution and standard properties of the variational inference (see Bishop, 2006 for a fuller exposition), and is given by

$$\begin{aligned} lnq^*(m_i) =& E_{q(k_i)}\left[ln\mathcal{N}(y_i \mid m_i, e^{-k_i})\right] + ln\mathcal{N}(m_i \mid \tilde{\mu}_i^{(m)}, \tilde{\tau}_i^{(m)-1}) + const \\ >& ln\mathcal{N}(y_i \mid m_i, e^{-E[k_i]}) + ln\mathcal{N}(m_i \mid \tilde{\mu}_i^{(m)}, \tilde{\tau}_i^{(m)-1}) + const \\ =& -\frac{1}{2}ln(2\pi) - \frac{1}{2}\breve{\mu}_i^{(k)} - \frac{e^{\breve{\mu}_i^{(k)}}(y_i - m_i)^2}{2} \\ & -\frac{1}{2}ln(2\pi) - \frac{1}{2}ln\left(\tilde{\tau}_i^{(m)}\right) - \frac{\tilde{\tau}_i^{(m)}(m_i - \tilde{\mu}_i^{(m)})^2}{2} + const \end{aligned} \tag{4.22}$$

Here we use the fact that for any linear function $E_{\mathcal{N}(\mu,\sigma^2)}[\theta x] = \mu$. We also make use of Jensen's inequality, according to which in convex functions

$$E[f(x)] > f(E[x]) \tag{4.23}$$

Thus, as the non-linear function we take expectations of (e^{-k_i}) is convex, we can use this result as a lower bound for the optimal solution. We applied this approximation for the sake of cleaner and more interpretable update equations, which can be more easily related with our reference theoretical framework (Predictive Coding), with very marginal influence on the end result.

Similarly

$$\begin{aligned} lnq^*(k_i) =& E_{q(m_i)}\left[ln\mathcal{N}(y_i \mid m_i, e^{-k_i})\right] + ln\mathcal{N}(k_i \mid \tilde{\mu}_i^{(k)}, \tilde{\tau}_i^{(k)-1}) + const \\ >& ln\mathcal{N}(y_i \mid E[m_i], e^{k_i}) + ln\mathcal{N}(k_i \mid \tilde{\mu}_i^{(k)}, \tilde{\tau}_i^{(k)-1}) + const \\ =& -\frac{1}{2}ln(2\pi) - \frac{1}{2}k_i - \frac{e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2}{2} \\ & -\frac{1}{2}ln(2\pi) - \frac{1}{2}ln\tilde{\tau}_i^{(k)} - \frac{\tilde{\tau}_i^{(k)}(k_i - \tilde{\mu}_i^{(m)})}{2} + const \end{aligned} \tag{4.24}$$

where once again we use Jensen's inequality as the non-linear function we take the expectation of (m_i^2) is convex.

Typically, in variational inference one makes use of conjugate prior distributions, which ensure that prior and posterior distributions are of the same type, and furnish straightforward update equations that can be iteratively evaluated (Bishop, 2006; Blei et al., 2017). This is not possible here, due to the use of a (non-conjugate) Gaussian

prior for the log precision, and the agent thus makes use of gradient ascent, combined with the Laplace approximation, to derive estimates of the posterior mean and variance for each variable (Bishop, 2006; Friston et al., 2007). The use of gradient ascent is particularly attractive here, as it is employed in classical formulations of predictive coding (Friston and Kiebel, 2009). Thus

$$\breve{\mu}_i^{(m)} = \underset{m_i}{argmax}(q^*(m_i)) \tag{4.25}$$

$$\breve{\mu}_i^{(k)} = \underset{k_i}{argmax}(q^*(k_i)) \tag{4.26}$$

$$\breve{\tau}_i^{(m)} = -\frac{\partial^2 q^*(m_i)}{\partial m_i^2}\left(\underset{m_i}{argmax}(q^*(m_i))\right) \tag{4.27}$$

and

$$\breve{\tau}_i^{(k)} = -\frac{\partial^2 q^*(k_i)}{\partial k_i^2}\left(\underset{k_i}{argmax}(q^*(k_i))\right) \tag{4.28}$$

Differentiating $q^*(m_i)$ twice with respect to m_i gives:

$$\frac{\partial q^*(m_i)}{\partial m_i} = e^{\breve{\mu}_i^{(k)}}(y_i - m_i) - \tilde{\tau}_i^{(m)}(m_i - \tilde{\mu}_i^{(m)}) \tag{4.29}$$

$$\frac{\partial^2 q^*(m_i)}{\partial m_i^2} = -e^{\breve{\mu}_i^{(k)}} - \tilde{\tau}_i^{(m)} \tag{4.30}$$

The first equation is the familiar core of predictive coding, which trades off prediction errors at the first $(y_i - m_i)$ and second $(m_i - \tilde{\mu}_i^{(m)})$ levels, weighted by their precisions.

Similarly, for the log precision k_i

$$\frac{\partial q^*(k_i)}{\partial k_i} = \frac{1}{2} - \frac{e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2}{2} - \tilde{\tau}_i^{(k)}(k_i - \tilde{\mu}_i^{(k)}) \tag{4.31}$$

$$\frac{\partial^2 q^*(m_i)}{\partial m_i^2} = -\frac{e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2}{2} - \tilde{\tau}_i^{(k)} \tag{4.32}$$

Gradient ascent is performed using Newton's method of optimisation, in which, for function f at iteration n, variable x is updated such that

$$\Delta x = -\frac{f'(x)}{f''(x)} \tag{4.33}$$

This yields the following coupled equations, which the agent iteratively evaluates:

$$\breve{\mu}_i^{(m)} \longleftarrow \breve{\mu}_i^{(m)} - \frac{e_i^{\breve{\mu}^{(k)}}(y_t - \breve{\mu}_i^{(m)}) - \tilde{\tau}_i^{(m)}(\breve{\mu}_i^{(m)} - \tilde{\mu}_i^{(m)})}{-e^{\breve{\mu}_i^{(k)}} - \tilde{\tau}_i^{(m)}} \tag{4.34}$$

$$\breve{\mu}_i^{(k)} \longleftarrow \breve{\mu}_i^{(k)} - \frac{\frac{1}{2} - e^{\breve{\mu}_i^{(k)}} \frac{(y_t - \breve{\mu}_i^{(m)})^2}{2} - \tilde{\tau}_i^{(k)}(\breve{\mu}_i^{(k)} - \tilde{\mu}_i^{(k)})}{-e^{\breve{\mu}_i^{(k)}} \frac{(y_t - \breve{\mu}_i^{(m)})^2}{2} - \tilde{\tau}_i^{(k)}} \tag{4.35}$$

For reasons of computational expedience, in the analyses presented here, we fixed the number of iterations to sixteen, rather than explicitly evaluating convergence.

4.3.5.3 Design matrix specification

The vectors of trial-by-trial mean and log precision estimates $\hat{\mathbf{m}} = \breve{\boldsymbol{\mu}}^{(m)}$ and $\hat{\mathbf{k}} = \breve{\boldsymbol{\mu}}^{(k)}$ were used to estimate a trial-specific surprise regressor $\mathbf{s}$ as follows:

$$s_i = -ln\mathcal{N}(y_i \mid \hat{m}_{i-1}, e^{-\hat{k}_{i-1}}) \tag{4.36}$$

We modelled behaviour using four versions of this agent. In the first (M1: 'full') model, both the mean and precision of the distribution were dynamically estimated as described above, using the priors given in Table 1. In the second (M2: 'precision only') model, belief updating about the mean was effectively prevented by fixing its prior variance to be 10^{-6}. In the third (M3: 'mean only') model by contrast, belief updating about precision was effectively prevented by fixing its prior variance to be 10^{-6}. In the fourth (M4: 'fixed') model, both sorts of belief updating were prevented in a similar fashion. (Note that, where belief updating is prevented, between-subject variability in fixed estimates of the mean or precision is still allowed, since m_0 and k_0 are still fitted as free parameters).

All four GLM-AR models included binary regressors to model tone and target presentation, a regressor encoding stimulus number (to model linear drifts in the pupil responses across a session), and one encoding $\mathbf{y}$ (frequency in log space). For completeness, we also tested models that included a "response" regressor indicating which stimuli subjects responded to, rather than one indicating target presentation. However, these models provided markedly inferior fits, so we do not discuss them further. In addition, we included a regressor encoding the absolute difference in frequency between each tone and the one immediately preceding it $|y_i - y_{i-1}|$, to account for simple adaptation effects.

Finally, models M1 and M2 additionally included a regressor encoding $e^{\hat{k}}$ (the trial-by-trial precision estimate). Identical weak (zero mean) priors were set for each regressor, as specified in Table 2.

In addition to these four inference-based models, we included a null (M0) model which included only the AR component of our scheme. This allows us to assess whether any of our models do better than a simple AR process.

4.3.5.4 Model fitting and comparison

Model-fitting was performed using variational Laplace (VL) (Daunizeau, 2017; Friston et al., 2007), with prior distributions as specified in Table 2.1. This is a fast and powerful approach to model-fitting, which furnishes an estimate of the model evidence that is typically more accurate than measures such as the AIC and BIC (Penny, 2012) but requires that model parameters be treated as Gaussian. We thus transformed our model parameters where necessary, as specified in Table 2.1. In addition to the model-specific parameters described above, VL also estimates a noise log-precision parameter ν for each subject, which we also report in Table 2.1. Model comparison was based on the negative variational free energy for each model and subject derived during model fitting.

4.3.5.5 Model checking and visualisation

In order to visually compare the accuracy of our model predictions for key task events against observed responses, we preprocessed each subject's pupillometry data by regressing out the AR component of our model, epoching (between -0.2 s and 2.5 s) and baseline correcting (for the interval -0.2 s to 0 s). We then performed identical preprocessing on the predicted time series for each subject, and plotted responses to probe and target tones.

To directly visualise the nature of pupil responses to surprise, and allow comparison with the responses predicted by our modelling, we performed a time-point by time-point regression analysis on the epoched data described above. The regression model contained a constant term, the surprise estimated for each trial from the full model for that subject, and a regressor encoding target trials. This was used to compare predicted and observed responses with each other, as well as with the gamma kernels produced by the GLM-AR modelling.

4.4 Results

4.4.1 Behavioural data

Behavioural data were not relevant to our hypothesis, and therefore were only analysed to ensure participants were paying attention to the tones they were exposed to and to check for unexpected effects of block type. Participants had an average hit rate (number of responses to targets over total number of targets) of 0.84 (range 0.56-0.98) and an average false alarm rate (number of responses to non-targets over total number of non-targets) of 0.008 (range 0.001-0.023). These data suggest that all participants paid attention to the tones, as, despite sometimes missing them, they almost exclusively responded to targets.

We also carried out three paired samples t-tests to investigate whether the type of block (high vs. low precision) influenced the hit rate, the false alarm rate and the reaction times. We found no significant difference in hit rate between high ($\mu = 0.85$, $\sigma^2 = 0.015$) and low ($\mu = 0.83$, $\sigma^2 = 0.031$) precision blocks ($t(15) = 0.80$, $p = 0.437$), nor in false alarm rate ($\mu = 0.008$, $\sigma^2 = 0.00004$ for high precision blocks and $\mu = 0.009$, $\sigma^2 = 0.00009$ for low precision blocks, $t(15) = -0.76$, $p = 0.460$). Likewise, we did not find a significant difference in reaction times between high ($\mu = 559$ms, $\sigma^2 = 8391$) and low ($\mu = 574$ms, $\sigma^2 = 9701$) precision blocks, though there was some evidence of a trend ($t(15) = -2.01$, $p = 0.063$). In sum, we found no clear evidence for differences in behaviour on the task between blocks.

4.4.2 Probe tones analysis

We first analysed responses to the standard and deviant probe tones, using a classical model-free analysis. A two-way repeated measures ANOVA revealed no main effect of precision ($F(1,15) = 3.35, p = 0.087$), nor of probe type ($F(1,15) = 3.74, p = 0.072$). On the other hand, the interaction was significant ($F(1,15) = 24.71, p < 0.001$), with the difference in pupil response between deviant probe and standard probe trials being bigger in high precision blocks than in low precision ones (Fig. 4.4). This clearly suggests that subjects learnt about the precision of stimulus distributions, even though these were task irrelevant.

Post-hoc analyses revealed no significant difference between the standard probe trials in high ($\mu = -2.7 \times 10^{-3}$, $\sigma^2 = 0.05 \times 10^{-3}$) and low ($\mu = -0.9 \times 10^{-3}$, $\sigma^2 = 0.03 \times 10^{-3}$) precision blocks ($t(15) = -0.81$, $p = 0.423$), nor between standard ($\mu = -0.9 \times 10^{-3}$, $\sigma^2 = 0.03 \times 10^{-3}$) and deviant ($\mu = 2.0 \times 10^{-3}$, $\sigma^2 = 0.021 \times 10^{-3}$) probe trials in low

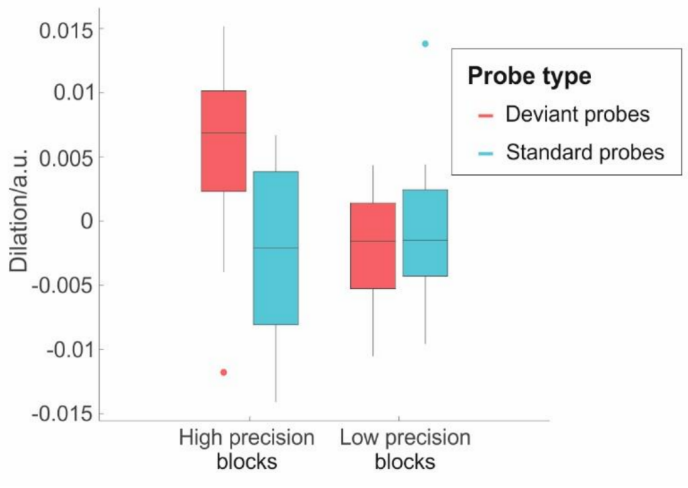

Fig. 4.4 Results of the model-free analysis, illustrating the change in pupil diameter in response to standard (500 Hz) and deviant (2000 Hz) probe tones in the high and low-precision conditions. Greater pupil dilation was observed for deviant compared with standard probes in the high precision condition (t(15) = 3.24, p = 0.022, Bonferroni-corrected), but not in the low precision condition (t(15) = -0.665, p = 0.516). This demonstrates that subjects tracked the current precision of the distribution of tones. (Data points between 900ms and 1000ms from stimulus onset were averaged for this analysis. Data are displayed using a Tukey boxplot, with points outside the whisker ranges additionally plotted).

precision blocks ($t(15) = -0.67$, $p = 0.516$). On the other hand, deviant probe tones elicited significantly bigger pupil dilation in high ($\mu = 5.7 \times 10^{-3}$, $\sigma^2 = 0.05 \times 10^{-3}$) compared to low ($\mu = -2.0 \times 10^{-3}$, $\sigma^2 = 0.021 \times 10^{-3}$) precision blocks ($t(15) = 4.97$, $p < 0.001$, Bonferroni-corrected). Finally, deviant probes ($\mu = 5.7 \times 10^{-3}$, $\sigma^2 = 0.05 \times 10^{-3}$) resulted in a larger pupil response compared to standard ones ($\mu = -2.7 \times 10^{-3}$, $\sigma^2 = 0.05 \times 10^{-3}$) in high precision blocks, ($t(15) = 3.24$, $p = 0.022$, Bonferroni-corrected). These results are illustrated in Fig. 4.4. This model-free analysis suggests that larger responses were associated with deviant probes when embedded in a narrower distribution, in keeping with the prediction that these events are more surprising, and consistent with previous work considering electrophysiological and behavioural responses (Garrido et al., 2013).

Our model-free approach, like those in many previous studies of the oddball paradigm, is restricted to consideration of carefully specified probe tones. This is inefficient, because it only considers a small subset of all experimental stimuli, and introduces restrictions that may preclude consideration of more subtle or complex statistical learning effects using pupillometry. Consequently, we performed a complementary model-based analysis, which forms the principle focus of this paper.

4.4.3 GLM-AR modelling

Model comparison was performed using model-space averaging (FitzGerald et al., 2019), a refinement of random-effects Bayesian model selection (Stephan et al., 2009) which automatically (and optimally) mitigates the dilutionary effect of including inferior models in the model-space. This analysis strongly favoured M1 and M2 (the "dual estimation" and "precision only" models, Table 2.2) as compared to the other models, but did not clearly distinguish between them. These were assigned posterior probabilities of 0.403 and 0.427, and protected exceedance probabilities (FitzGerald et al., 2019; Rigoux et al., 2014) of 0.441 and 0.533. This provides clear evidence that subjects tracked the precision of the tone distribution, even though it was task-irrelevant, but does not settle the issue of whether they also tracked the mean of the distribution. This is perhaps unsurprising, given that the mean actually remained constant across the experiment, and could usefully be investigated in future work. We thus selected M2 for use when performing analyses of model performance as described below, on the basis that this seemed the "conservative" option.

One possible alternative explanation for our precision-tracking results is that responses might reflect a similarity effect of recent stimuli, which might be expected to differ between conditions (in the high precision condition recent stimuli tend to be more similar in frequency to the current stimulus than in the low precision condition). We control for this in our main analysis through the use of an adaptation regressor encoding the absolute difference between the current and previous stimulus, but, as pointed out by a reviewer of the journal article reporting this work (Silvestrin et al., 2021), this might not be sufficient if the similarity effect involved multiple recent stimuli. To rule this out, we carried out an extra check analysis in which we included separate regressors encoding the absolute difference between the current stimulus and each of the seven preceding ones. We used this augmented approach to compare the "dual estimation" and "mean only" models, to see if there was evidence in favour of precision tracking even when this fuller stimulus history was accounted for. Reassuringly, this also provided strong evidence in favour of the full model, which had an exceedance probability of 0.994.

Quality of model fits, as estimated using simple percentage variance explained, was excellent ($\mu = 0.999$, range: 0.997-1.000). However, this includes the effects of the AR process, which is not of interest here, so it is also useful to assess how much variance is explained by the GLM itself. To assess this, we regressed out the predictions of the AR component from each subject's time series, and then calculated percentage variance explained solely by the GLM. This also showed a good fit with the data

Parameter	**Prior mean (variance)**	**Group-level posterior (variance)**	***p*-value (corrected)**
$ln(h)$	$ln(3)(2)$	1.61(0.271)	-
$ln(l)$	$ln(3)(2)$	1.59(0.536)	-
$ln(d)$	$ln(0.2)(2)$	$-3.09(3.390)$	-
a	1(2)	$1.00(< 0.001)$	-
w_{event}	0(4)	0.10(0.114)	1.000
$w_{precision}$	0(4)	0.06(0.027)	1.000
$w_{surprise}$	0(4)	0.04(0.002)	0.004
w_{pitch}	0(4)	$0.00(< 0.001)$	1.000
w_{target}	0(4)	0.54(0.111)	< 0.001
w_{drift}	0(4)	$-0.15(0.044)$	0.027
w_{adapt}	0(4)	$0.00(< 0.001)$	1.000
$ln(\eta^{(m)})$	$ln(100)(2)$	5.72(13.49)	-
$ln(\eta^{(k)})$	$ln(100)(2)$	1.31(2.93)	-
m_0	$ln(500)(2)$	9.12(18.15)	-
k_0	$ln(1)(2)$	5.95(3.39)	-
ν	4(4)	6.82(0.475)	-

Table 4.1 Summary statistics of the prior and group-level posterior distributions over each parameter. (Posterior distributions are based on weighted average single-subject parameter estimates). Parameters were transformed where appropriate to enable use of a Gaussian prior distribution, as required by the VL algorithm (ln indicates the natural logarithm). Non-parametric p-values calculated using permutation testing, Bonferroni-corrected for seven comparisons. These provide clear evidence that both surprise and target presentation were consistently associated with pupil dilation, and for a progressive decrease ("drift") in dilation responses over the session.

($\mu = 0.848$, range: 0.154-0.990), as illustrated by Figs. 4.5 and 4.6. Inspection of the convolution kernels derived from our modelling suggested that these provided plausible pupil dilation responses both when compared with existing literature (Denison et al., 2020; Hong et al., 2014; Knapen et al., 2016; Korn and Bach, 2016), and when compared with the waveform generated when regressing estimated beliefs about surprise onto

Model	$\sum VFE$	$\sum VFE - \sum VFE_{null}$	**Posterior probability**	**Protected exceedance probability**
M1 ("Dual estimation")	5358538	49684	0.403	0.441
M2 ("Precision only")	5358392	49538	0.427	0.533
M3 ("Mean only")	5357825	48971	0.153	0.019
M4 ("Fixed")	5357787	48933	0.011	0.004
M0 ("Null")	5308854	0	0.007	0.003

Table 4.2 Model comparison strongly favoured the models in which subjects dynamically updated their beliefs about the precision of the stimulus distribution (M1 and M2), but do not clearly distinguish these two, thus providing no clear evidence about whether subjects also inferred on the mean of the distribution. Model comparison was performed using model-space averaging, as described in FitzGerald et al. (2019).

epoched data (Fig. 4.6). This suggests that our GLM-AR modelling approach was appropriate for analysing these data, and supports its use in future studies.

In addition, we explored how closely the predictions made by the fitted models tracked the true precision of the stimulus distributions (Fig. 4.7). Analysis using Spearman's rank correlation coefficient showed a strong positive correlation between trial-by-trial precision estimated and the true task contingencies ($\overline{\rho} = 0.77$, $\sigma^2 = 0.01$), suggesting that our cognitive model performed adequately on the task.

To test for group-level effects of the factors in the GLM, we calculated an average of the maximum a posteriori (MAP) parameter estimates for each subject, weighted by the posterior probability assigned to each model in that subject during the model comparison. Inference was then performed using a permutation test in which we flipped the signs of the parameters in a randomly selected subset of subjects 100,000 times and used the resultant surrogate data to provide surrogate two-tailed p-values for each variable. These were corrected for seven comparisons using a Bonferroni correction. Clear evidence of positive dilation responses to surprise ($\overline{w} = 0.040$, $p = 0.004$) and target presentation ($\overline{w} = 0.538$, $p < 0.001$) were found, as well as a negative effect of trial order ($\overline{w} = -0.154$, $p = 0.026$), suggesting a progressive decrease in the size of dilation responses across the course of the experiment (Fig. 4.8). No statistically

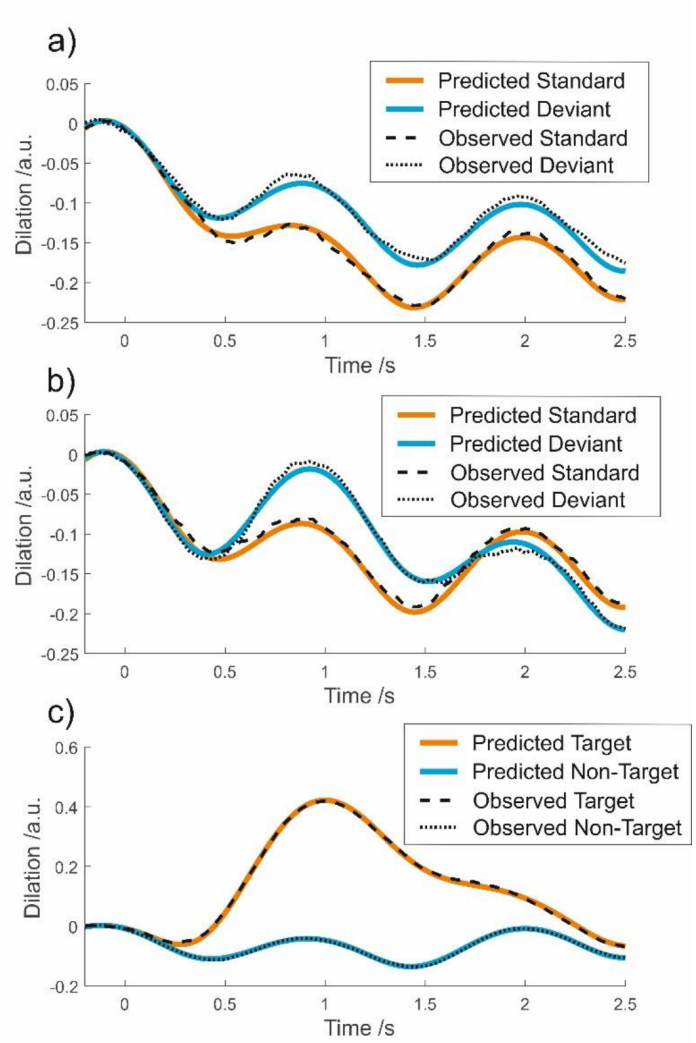

Fig. 4.5 Illustration of the accuracy of GLM-derived predictions for a single representative subject (subject two, percentage variance explained = 0.87, similar to the group mean). Each plot shows epoched, baseline corrected and averaged waveforms for the predicted (blue/orange) and observed (black) responses to key task conditions. (The predictions of the AR component of the model were regressed out prior to epoching, and these thus solely reflect how well the GLM predicts the data). The top plot (a) illustrates probe tones in the low precision condition, the middle plot (b) illustrates probe tones in the high precision condition, and the bottom plot (c) illustrates target and non-target tones (across both conditions). For all waveforms there is a close correspondence between predicted and observed data, reflecting the accuracy of the model fits.

significant effects were observed for precision itself, log frequency, tone presentation,

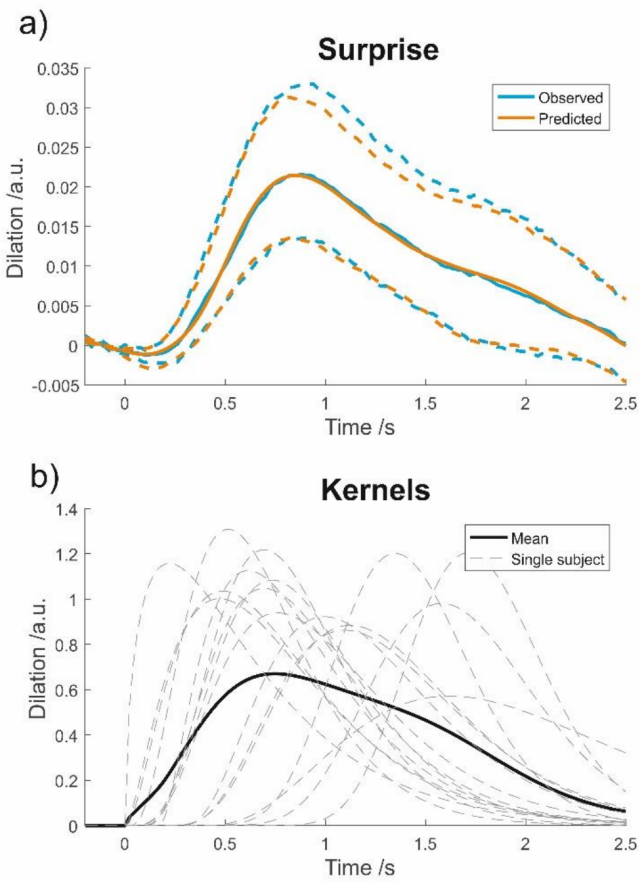

Fig. 4.6 (a) Surprise waveform estimated from our regression analysis (see Methods for more details). Observed dilation responses to surprise (blue) peak at roughly one second and then return to baseline. Predicted responses to surprise derived from our GLM-AR modelling (orange) show a close correspondence to observed responses (dashed lines indicate bootstrapped 95% confidence intervals). (b) Gamma kernels modelling pupil dilation derived from the GLM-AR model. (Single subject responses in grey, and the mean in black). These strongly resemble both the surprise waveform derived in our regression analysis and averaged responses from tasks using slower designs (for example Hong et al., 2014).

or the frequency separation between successive tones. (See Table 2 for full results). Additionally, we tested for between-subject correlations in the regression coefficients, using a partial correlation approach to control for non-specific differences in coefficient magnitude. These might be caused, for example, by quality of model fit, or the subject-specific shape of the gamma kernel. This showed no evidence for statistically significant correlations between coefficients, and since we have no clear hypotheses about such relationships, we do not discuss them further.

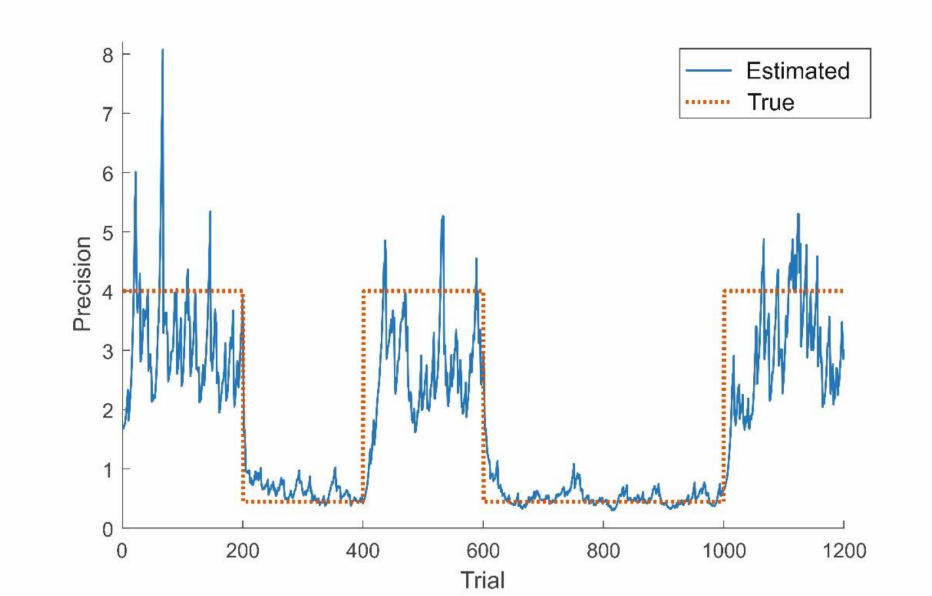

Fig. 4.7 Illustration of how trial-by-trial estimates of precision (blue) track the true precision of the distribution used to generate the non-probe tones (orange). Data is plotted from the first 1200 trials in a single representative subject (subject two). The tendency to underestimate precision in the high-precision blocks most likely reflects the distorted (and lower precision) probability distributions induced by the use of probe tones.

In line with previous work (Garrido et al., 2013), the stimulus distributions that we used in this experiment are distorted to introduce probe tones that can be compared across conditions (Fig. 4.2). This introduces the possibility that these tones are treated differently by subjects, and might be responsible for driving our results. To rule this out, we repeated the model comparison described above, using models that ignored probe trials. Reassuringly, these results were very similar, with M1 and M2 assigned posterior probabilities of 0.382 and 0.443 respectively, and protected exceedance probabilities of 0.370 and 0.604. This suggests that our key results were not driven by responses to the probe tones.

4.5 Discussion

The results obtained in this study provide clear evidence that pupil dilation reflects automatic and dynamically updated beliefs about the precision of stimulus distributions, in keeping with theories of probabilistic cognition (Aitchison and Lengyel, 2016; Friston, 2010; Ma et al., 2006; Tenenbaum et al., 2006). This extends previous work showing

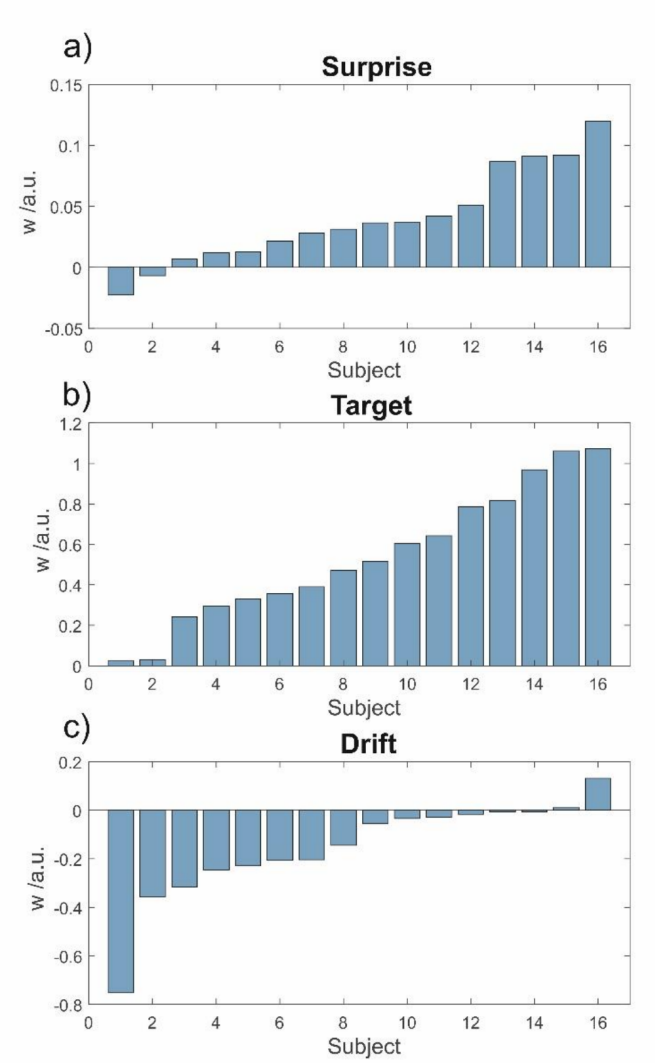

Fig. 4.8 Ordered parameter estimates of single subject regression weights for the Surprise regressor ($w_{surprise}$, a), Target regressor (w_{target}, b), and linear drift (w_{drift}, c) derived using weighted averaging (see Methods for further details). Positive pupil dilation responses to both surprise and target presentation were highly consistent across subjects, as was a progressive decrease in the size of responses over time.

evidence for an effect of the precision of stimulus distributions on reaction times and MEG responses (Garrido et al., 2013), and suggests that pupillometry can be a useful tool for examining statistical learning about higher order properties of stimulus distributions (Alamia et al., 2019), something we will consider further in future work. In addition, we use these data to demonstrate the potential for analysing pupillometry

data using a GLM-AR approach, which allowed highly accurate prediction of observed responses in our data (Figs 4.5, 4.6).

Our study complements an existing body of work using pupillometry to explore learning and related processes (Denison et al., 2020; Ebitz and Moore, 2019; Mathôt, 2018). Perhaps the simplest manifestation of this is in the literature on pupil dilation in response to perceptual oddballs (Friedman et al., 1973; Hong et al., 2014; Korn and Bach, 2016; Liao et al., 2016; Murphy et al., 2011; Qiyuan et al., 1985; Steinhauer and Zubin, 1982), but a similar approach has been adopted to explore response during gambling and learning tasks (Hämmerer et al., 2019; Lavín et al., 2014; Preuschoff et al., 2011), change-point detection (Nassar et al., 2012), the role of risk and learning about transition probabilities between discrete states (Alamia et al., 2019), and responses to volatility (Browning et al., 2015; Vincent et al., 2019), as well as surprise in other contexts (Kloosterman et al., 2015; Knapen et al., 2016; O'Reilly et al., 2013). As such, the principle contribution of our findings is to provide new information about the sort of cognitive processes that are reflected in pupil dilation responses, and contribute to the growing literature linking them specifically to statistical learning (Alamia et al., 2019).

A key limitation of many previous studies using oddball paradigms is the conflation of deviant and target stimuli (Hong et al., 2014; Liao et al., 2016; Murphy et al., 2014, 2011; Rajkowski et al., 1994, 2004; Steinhauer and Zubin, 1982), though see Wetzel et al. (2016) for studies which avoid this. This conflation makes it difficult to attribute pupil/LC effects unequivocally to surprise rather than other task-related processes. We obviated this by making pitch irrelevant to the task, so that evoked pupil dilation could be directly associated with stimulus probability (and, therefore, surprise). Importantly, having pitch be task-irrelevant allowed us to explore automatic, and possibly implicit, learning processes. This provides a complement to paradigms in which learning is directly relevant for behaviour, and indexes what is likely to be an important form of learning for behaviour in ecological settings.

It should be noted that a recent review (Zénon, 2019), in an attempt to give a unified explanation to the pupil effects of a wide range of cognitive processes (e.g. mental effort, attention, exploration/exploitation trade-off, decision making, surprise), related pupil dilation to information gain. This is formalised as the Kullback–Leibler divergence between prior and posterior distributions often called "Bayesian surprise" (Baldi and Itti, 2010; Schwartenbeck et al., 2016; Zénon, 2019), as opposed to information-theoretic ("Shannon") surprise (negative log probability). Bayesian surprise has the advantage that it quantifies the meaningful information present in a stimulus, as opposed to

simply how unexpected it is, and it is thus a plausible candidate to play a role here. However, our paradigm is not designed to separate these two quantities, and thus evaluating them as competing explanations for pupil dilation responses goes beyond the scope of our study. Similarly, given the close relationship between surprise and dynamic beliefs about volatility (Silvetti et al., 2013), it is conceivable that the dilation responses we observe more directly index beliefs about volatility, but this falls outside the scope of our study to test.

A diverse set of evidence points to a tight link between pupil diameter and noradrenergic activity in locus coeruleus (LC). This was first observed in monkeys, with an electrophysiological study showing that pupil diameter tracked LC tonic activity (Rajkowski, 1993). Pharmacological evidence confirmed this finding in humans, with LC-suppressing drugs decreasing pupil diameter and LC-stimulating drugs enhancing it (Hou et al., 2005; Phillips et al., 2000). In addition, theoretical work (Dayan and Yu, 2006) has linked LC activity with surprise, as electrophysiological data on monkeys seem to suggest (Rajkowski et al., 1994, 2004). Murphy and colleagues carried out an experiment linking everything together, showing how LC BOLD activity correlates with pupil diameter in resting state and how they respond similarly to deviant stimuli in an oddball task (Murphy et al., 2014), supporting the already popular idea (Lavín et al., 2014; Preuschoff et al., 2011) that surprise-related pupil dilation occurs as an effect of phasic noradrenergic activity. Assuming that this link holds here, our study thus makes the novel contribution that noradrenergic function, and thus the cognitive processes it subserves (Dayan and Yu, 2006), are sensitive to dynamically updated beliefs about stimulus precision, and may even play a role in this process.

Use of auditory oddball paradigms to index automatic statistical learning has considerable practical attractions, not least the fact that it requires minimal subject compliance (Boly et al., 2011). Combining this with pupillometric data collection is attractive, as such data is relatively simple and cheap to collect, particularly when compared with neuroimaging modalities such as MEG and fMRI. The general approach here thus has potential for exploring cognitive changes related to statistical learning in patient groups (Browning et al., 2015), as well as during healthy ageing (Hämmerer et al., 2019).

The GLM-AR approach that we adopt here, though originally motivated by the strong similarities between pupillometry and fMRI data (Penny et al., 2003), relates closely to a number of existing approaches. The use of a convolution kernel for analysing pupillometric data was first proposed by (Hoeks and Levelt, 1993), and similar approaches have been adopted by various authors subsequently (de Gee et al.,

2014; Denison et al., 2020; Knapen et al., 2016; Korn and Bach, 2016; Vincent et al., 2019; Wierda et al., 2012). We note though that many of these studies make use of a canonical pupil response, which is likely to be suboptimal, given the evidence for individual variability (Denison et al., 2020). However, such approaches do not account for the strong slow fluctuations in pupil diameter (Zénon, 2017). Typically, the effects of these fluctuations are mitigated via epoching, baseline-correction, and averaging, but this requires many repetitions of a particular trial type, which makes it difficult to capture phenomena such as learning (as we do here). We address this issue through use of an AR model, as has been proposed recently (Alamia et al., 2019; Zénon, 2017). However, to our knowledge the GLM-AR approach we propose is the first to combine the advantages of both individually tailored convolution kernels and AR modelling. A rigorous assessment of the significance of this is beyond the scope of this study, but we will explore it in future work.

A further, (and critical for our purposes), aspect of the GLM-AR approach that we use is that it allows us to fit the parameters of behavioural models to pupillometric responses, an approach more typically confined to analysis of behavioural data (Daw et al., 2011; O'Doherty et al., 2007; Schwartenbeck et al., 2015; Smittenaar et al., 2013). This "doubly model-based" aspect is important, as it allows us to use model comparison to establish which behavioural models best account for pupillometric data (for example, one might use it to compare different models of learning). In addition, it permits, in principle, the characterisation of between-subject variability in processes such as learning and inference, which may be of particular interest for understanding pathology (Browning et al., 2015; Huys et al., 2016; Krystal et al., 2017; Montague et al., 2012). However, the extent to which applying the GLM-AR approach to pupillometric data in practice permits such inferences is unclear, and future work will be necessary to establish this.

In sum, at a cognitive level, our work represents a contribution both to understanding task-irrelevant human statistical learning processes, and to characterising the computational mechanisms underlying pupil dilation responses to surprising stimuli. In addition, we believe that the GLM-AR approach that we propose has considerable potential for increasing the accuracy and flexibility of pupillometry data analysis, something we will explore in future work.

Chapter 5

Physiological responses to surprising stimuli violate classical predictive coding

5.1 Abstract

Predictive coding is perhaps the most-widely held account of probabilistic inference in the brain, and provides a powerful and elegant framework for explaining a range of experimental data. A core and obligatory feature of predictive coding is the use of Gaussian probability distributions, which makes the strong (if counterintuitive) prediction that the brain will encode all variables using Gaussians, even when the true distributions are radically non-Gaussian. We explored this prediction using a variant of the classic auditory oddball task, in which tones were drawn from a bimodal probability distribution. Strikingly, we found clear evidence that subjects treated the distribution of tones as being bimodal, in violation of classical predictive coding theories. Our findings thus suggest a need to augment existing predictive coding models, or else replace them with a more flexible scheme.

5.2 Introduction

In the rich landscape of probabilistic theories of brain function, predictive coding (PC) is perhaps the most popular, and has provided an elegant framework for explaining a range of perceptual (Denison et al., 2011; Hohwy et al., 2008; Watanabe et al., 2018), cognitive (Seth, 2013; Seth et al., 2012) and neural (Auksztulewicz and Friston, 2016;

Garrido et al., 2008; Kok et al., 2012) phenomena. According to this framework (Clark, 2013; Friston, 2005), the brain is constantly trying to anticipate the incoming bottom-up sensory signal via top-down predictions, effectively "explaining away" incoming sensory information. The discrepancy between top-down predictions and bottom-up inputs, called prediction error, is the only portion of the sensory signal that gets retained for further processing. This implies a hierarchical organisation of information processing in the brain, with each level of the hierarchy receiving bottom-up prediction errors from the level below and top-down predictions from the level above. The general purpose of the brain in this context is that of making top-down predictions and incoming sensory signals match, or, in other words, minimising prediction error (Friston, 2005).

In practice, to generate predictions the brain must have a model of its environment which specifies how sensory observations are generated. This takes the name of *generative model*. It can be shown (see Chapter 1, Friston (2005); Mathys et al. (2011)) that Bayesian model inversion through variational inference (Bishop, 2006) does result in prediction error minimisation if the generative model is Gaussian. Here we call *inference* the estimation of time-varying hidden states (i.e. latent variables causing sensory observations) through model inversion, and *learning* the update of model parameters, which in a PC setting are limited to the mean of the Gaussian priors over hidden states (updating beliefs about precision through prediction error minimisation is trickier, see Chapter 4).

One could thus see inference as reducing prediction error in the short term by optimising beliefs about hidden states and learning as reducing it in the long term, by putting oneself in the condition of making better predictions in the future.

This Gaussian formulation constitutes a foundational principle of PC, providing an elegant mathematical framework which can be used to build neurobiologically plausible models of cortical function (Bastos et al., 2012). However, it also constitute a significant constraint, as it predicts the brain to encode any continuous variable as a Gaussian distribution, even when that's not the case (see Chapter 1 for a mathematical demonstration of this).

In this work we test this somewhat counterintuitive prediction with two experiments. Experiment 1 builds on our previous work (see Chapter 4 and Silvestrin et al. (2021)), using pupil dilation as an index of surprise (as discussed more in depth in Chapter 4) in a modified version of the classic auditory oddball paradigm (Bodatsch et al., 2011; Weber et al., 2020). Here participants were presented with a series of tones whose log frequency was drawn from a bimodal probability distribution (a mixture of two Gaussians). We chose a bimodal distribution so that a Gaussian encoding would yield

very contrasting predictions, as a unimodal approximation would "miss" both peaks of the real distribution (see Fig. 5.1). In Experiment 2 we repeated Experiment 1 with exactly the same procedure and stimuli, but using EEG instead of pupillometry. Our component of interest here was Mismatch Negativity (MMN), a largely documented event-related potential (ERP) with higher negativity for surprising events compared to unsurprising ones (Garrido et al., 2008, 2009b; Lee et al., 2017; Schwartz et al., 2018). As pupil dilation, MMN amplitude was used as an index of surprise.

In both experiments the distributions were slightly distorted (see Fig. 5.2 for an illustration of such distortion) by the addition of probe tones (which we call Probe 1, 2 3 and 4), which we used for a model-free analysis, as in Garrido et al. (2013) and Silvestrin et al. (2021) (i.e. the experiment described in Chapter 4). The probe tones were chosen so that unimodal and bimodal encoding of the stimuli would result in opposite predictions, with two (Probe 1 and 3) at the peaks of the (real) bimodal distribution (which we call *standard probes*, one (Probe 2) in between the two peaks and one (Probe 4) in a low-probability area higher than the second peak *deviant probes*. With this setup, classical PC would predict Probe 2 to be the least surprising, as it would be at the peak of a unimodal (mis)representation. We predicted that both pupillometry and EEG would violate classical PC, with deviant probes (Probe 2 and 4) eleciting bigger surprise (indexed by pupil dilation and MMN) than standard probes (Probe 1 and 3).

In addition to this model-free analysis we also carried out a model-based one, directly comparing comparing competing generative models (i.e. unimodal Gaussian vs multimodal mixture of Gaussians).

5.3 Methods

5.3.1 Participants

20 participants (15 females) aged 18 to 29 (mean = 20.9) took part in Experiment 1 and 25 participants (19 females) aged 18 to 44 (mean = 22.0) took part in Experiment 2. They had all normal or corrected-to-normal vision, with no history of neurological or psychiatric disorder (including substance abuse) nor hearing problems. Written informed consent was obtained from all participants.

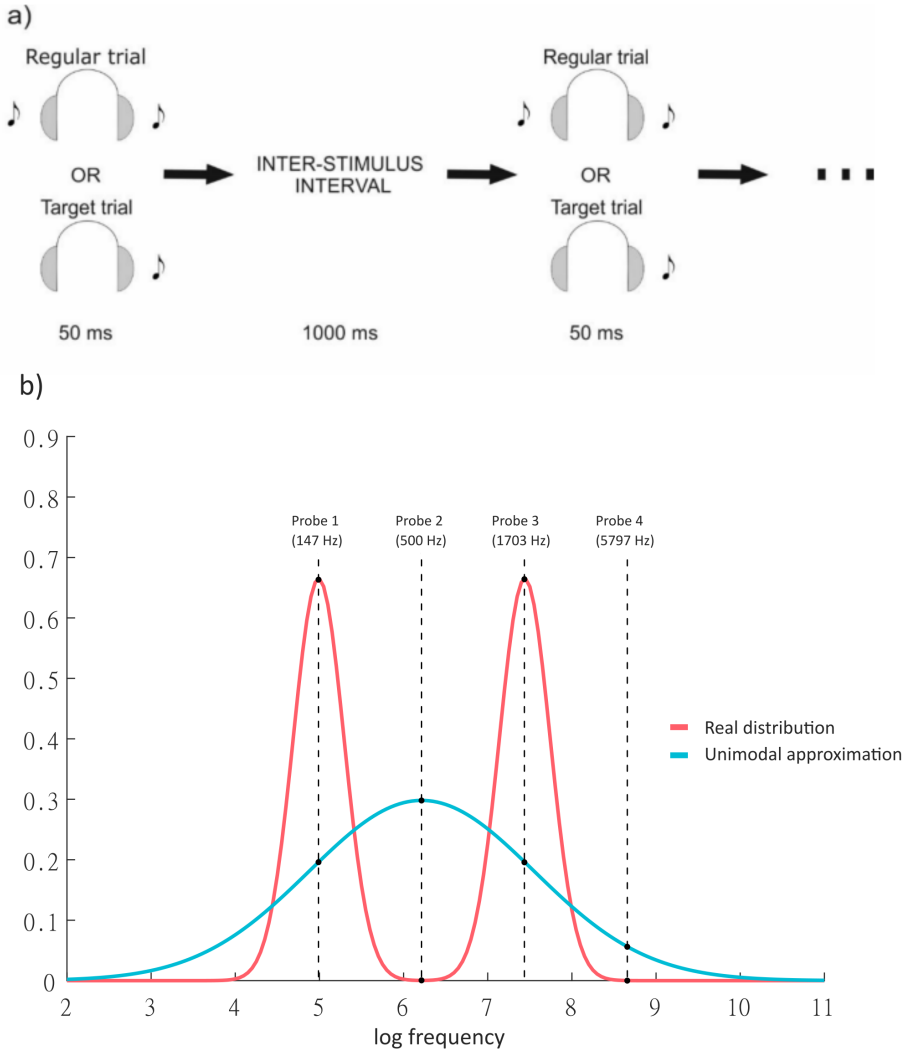

Fig. 5.1 (a) Illustration of the experimental paradigm. Participants were exposed to a series of tones (800 per session, 3200 in total) and were asked to press the space bar when they heard the sound coming only from one speaker (i.e. only from one side). (b) The pitch of the tones was sampled from a bimodal distribution (red), which, according to predictive coding, the brain would misrepresent as a unimodal Gaussian. In line with previous work (Garrido et al., 2013; Silvestrin et al., 2021) 4 probe tones were added. The two distributions illustrate the different probabilities associated with each probe under the assumption of a unimodal or bimodal generative distribution, yielding very different predictions. If participants represented the distribution as unimodal, Probe 2 would be the least surprising, as it is exactly at the peak of the distribution. Conversely, a bimodal encoding of the stimuli would cause Probe 2 to be much more surprising than Probes 1 and 3.

5.3.2 Materials

Both experiments were programmed with the MatLab package Psychophysics Toolbox (Brainard and Vision, 1997). Pupillometry data for Experiment 1 were collected with an EyeLink 1000 eye-tracking device, whilst EEG data for Experiment 2 were collected with a 64 electrodes BrainProduct actiCAP EEG device.

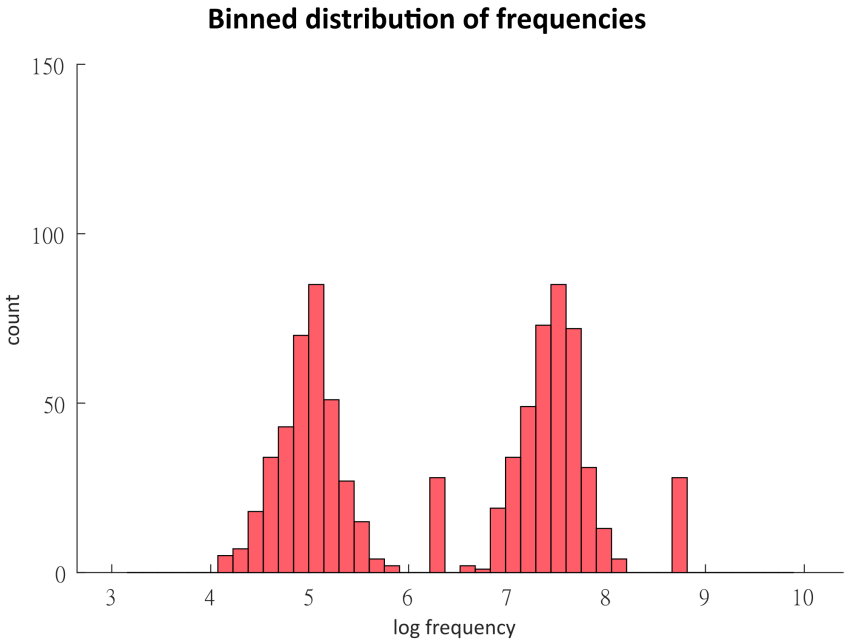

Fig. 5.2 Binning the distributions reveals the slight distortion caused by the addition of the probe tones. The binning was made so that all probe tones would fall exactly in the centre of a bin.

5.3.3 Task

The task and stimuli were identical in Experiments 1 and 2. Participants were asked to listen to a series of tones through headphones while looking at a fixation cross in the centre of a computer screen. The only thing required of them was to press the space bar whenever a tone was presented unilaterally (i.e. from only one of the two speakers, see Fig. 5.1).

Both experiments were divided into 4 sessions, with participants being allowed to a short break in between. During each session participants were presented with 800 pure

tones, each lasting 50 ms and with one second between tones, for a total duration of 14 minutes per session. Both experiments lasted approximately one hour.

In all sessions, the log frequency of the tones was sampled from a mixture of Gaussians with two modes (i.e.bimodal distribution). Therefore for each of the 4 sessions, 400 tones were sampled from one of two Gaussian distributions and then randomly shuffled together. The first Gaussian had mean $\mu_1 = 4.99$ (equivalent to 147 Hz) and the second had mean $\mu_2 = 7.44$ (equivalent to 1703 Hz). Both had standard deviations of $\sigma = 0.3$ (equivalent to 0.3 octaves). 112 of these tones were then pseudo-randomly replaced with the 4 types of probe tones (Probe 1, 2 3 and 4; 147 Hz, 500 Hz, 1703 Hz and 5797 Hz respectively, each 1.225 octaves apart from the next). Each session therefore contained 688 tones sampled from the bimodal distribution and 28 probe tones for each probe type, for a total of 800. The distribution was thus slightly distorted by 4 point-masses of 3.5% probability each.

The number of unilateral, target tones varied across sessions (78, 79, 79, and 77 respectively). These were pseudo-randomly inserted in the stream, and were made to never coincide with a probe, or occur immediately after it This was done because targets are very likely to elicit a strong, long lasting pupil response, which would confound the effect of surprise in Experiment 1.

5.3.4 Data acquisition and preprocessing

5.3.4.1 Experiment 1

Pupil data from both eyes were recorded at 500 Hz while participants performed the task in a moderately lit room.

Saccades and eyeblinks artifacts were removed by linear interpolation and a low pass filter of 20 Hz was applied to eliminate high frequency noise. The average pupil size between the two eyes was considered for all subsequent analyses. For the model-based analysis (but not for the probe tones analysis) data were additionally detrended, downsampled to 50 Hz, mean-corrected and normalised to unit variance, to have more comparable responses across subjects.

5.3.4.2 Experiment 2

EEG electrodes organised according to standard 64-channel-Arrangement, with FT9 used as a horizontal electro-oculogram (hEOG) to monitor eye movement and eyeblinks and FT10 as Iz. EEG signal was recorded with a sampling rate og 1000 Hz after ensuring all electrode impedances were under 25 kΩ.

EEG data were preprocessed with the Fieldtrip MatLab toolbox (Oostenveld et al., 2011). A 0.1 Hz high-pass filter was applied, and and data were re-referenced to the overall average. Data were then epoched from 50 ms before to 1050 ms after stimulus onset, and these epochs were visually inspected to identify bad channels, which (where present) were interpolated, and bad trials, which were discarded. After this, fast Independent Component Analysis (fastICA) was performed, and the resulting components were inspected to discard eyeblinks, eye movement, mechanical noise and high-frequency noise. Finally, data were downsampled to 256 Hz, baseline corrected (using a 50 ms pre-stimulus window) and 30 Hz low-pass filter was applied.

5.3.5 Data analysis

5.3.5.1 Experiment 1

Probe tones analysis As in the previous chapter, we first adopted a model-free approach to analyse pupil response to probe tones only. All data points between 900 and 1000 ms were averaged to capture the evoked response's peak (Hoeks and Levelt, 1993) and a one-way repeated measures ANOVA was carried out with probe type as the only 4-level within-subjects factor.

Model-based analysis To compare the competing generative models (underlying unimodal and multimodal representations) we deployed the same methods described in the previous chapter, with a General Linear Model incorporating a convolution kernel and an autoregression (AR) component. This GLM-AR model is described in detail in Chapter 4 and in Silvestrin et al. (2021).

However, unlike in Chapter 4, here we make use of static distributions. This means our models do not incorporate learning (i.e. generative model update), but rather assume that the distributions' statistics are known in advance. We thus treated such statistics as model parameters and we fitted them to individual data. We reasoned that, as stimuli are drawn from the same distribution for all 3200 trials, significant belief update would arguably occur only at the beginning. Therefore, there would be very little to be gained from incorporating learning into our models, and considerable added complexity. We thus deemed a simple comparison of competing (static) generative models sufficient to test our main hypothesis, namely that human participants are capable of representing multimodal distributions, in contrast with classical PC.

As in Chapter 4 (where this is explained in more detail), we fitted these models using Variational Laplace (Daunizeau, 2017; Friston et al., 2007), and used the resulting

variational free energy to perform model comparison through model-space averaging (FitzGerald et al., 2019).

Cognitive Modelling To model both unimodal and multimodal distributions, we use a Gaussian mixture model (GMM, for a fuller discussion see Bishop (2006)). To do so, we introduce a set of latent variables $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_I\}$ where each $\mathbf{c}_i$ is a 1-of-I binary vector indicating which Gaussian observation y_i is drawn from. Thus the number of modes of the distribution corresponds to I, giving a unimodal distribution, which we take to correspond to PC, when $I = 1$. Assuming (in this case accurately) that there is no temporal structure to $\mathbf{C}$, the conditional distribution over $\mathbf{c}_i$ is given by

$$p(\mathbf{c}_i \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_n^{c_{i,n}} \tag{5.1}$$

where $\boldsymbol{\pi}$ denotes the mixing coefficients (how likely it is for a randomly-selected sample to be drawn from each Gaussian).

The conditional distribution of observations is given by:

$$p(y_i \mid \mathbf{c}_i, \mathbf{m}, \boldsymbol{\lambda}) = \prod_{n=1}^{N} \mathcal{N}(y_i \mid m_n, \lambda_n^{-1})^{c_{i,n}} \tag{5.2}$$

Note that here the estimated sufficient statistics of the stimuli distribution ($\mathbf{m}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$), contrary to the previous chapter, are not trial-specific (the distribution is assumed to be static), and therefore have no temporal index.

For our analysis we considered 4 different Gaussian mixture models, differing only in the number of components (1, 2, 3 and 4, which we call M1, M2, M3 and M4).

5.3.5.2 Experiment 2

Probe tones analysis As for pupillometry data, we analysed electrophysiological responses to probe tones with a repeated-measures ANOVA with a single 4-level within subjects factor (probe type). To deal with multiple comparisons, we adopted a cluster-based non-parametric approach, as described in Maris and Oostenveld (2007). The evoked response of interest here was the mismatch negativity (MMN), which in similar auditory oddball paradigms (Garrido et al., 2013; Näätänen et al., 2004) has been shown to peak at around 150 ms from stimulus onset. We therefore restricted our analysis to electrophysiological activity ranging from 100 ms to 200 ms from stimulus onset.

5.4 Results

5.4.1 Pupillometry

5.4.1.1 Probe tones analysis

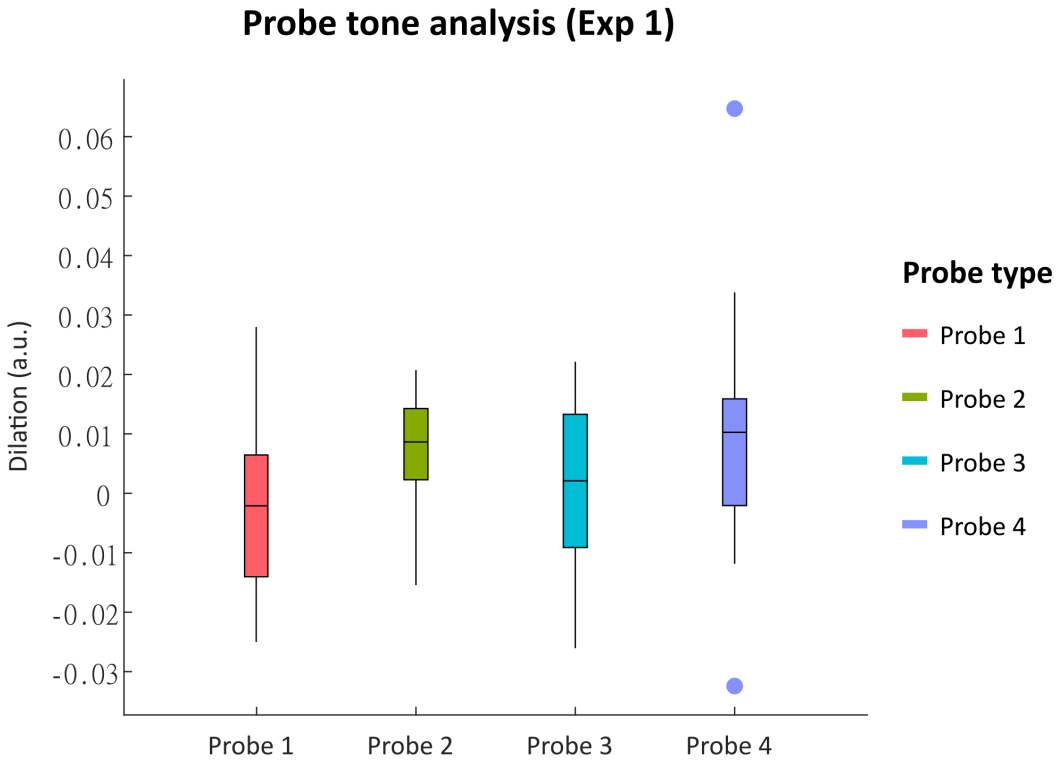

Fig. 5.3 Boxplot summarising the results of the model-free analysis. Greater pupil dilation was observed for deviant compared with standard probes, with Probe 2 eliciting more dilation than both Probe 1 ($t(19) = 2.56$, $p = 0.010$) and 3($t(19) = 2.01$, $p = 0.029$) and Probe 4 similarly eliciting more dilation than Probe 1 ($t(19) = 2.23$, $p = 0.019$) and 3 ($t(19) = 1.89$, $p = 0.037$). These results suggest participants did learn that the frequencies of the tones were bimodally distributed.
Data points between 900ms and 1000ms from stimulus onset were averaged for this analysis. Data are displayed using a Tukey boxplot, with points outside the whisker ranges additionally plotted.

The one-way repeated measures ANOVA we carried out on the data averaged from 900 to 1000 ms after stimulus onset revealed a significant main effect of probe type ($F(3,57) = 3.54$, $p = 0.020$). As we were specifically predicting the deviant probes

(Probe 2 and 4) to elicit a bigger pupil dilation than the standard ones (Probe 1 and 3) we carried out additional one-tailed paired-samples t tests, which revealed Probe 2 ($\mu = 7.9 \times 10^{-3}$, $\sigma^2 = 0.08 \times 10^{-3}$) to elicit a bigger pupil response than Probe 1 ($\mu = -2.7 \times 10^{-3}$, $\sigma^2 = 0.22 \times 10^{-3}$, $t(19) = 2.56$, $p = 0.010$) and 3 ($\mu = 1.3 \times 10^{-3}$, $\sigma^2 = 0.20 \times 10^{-3}$, $t(19) = 2.01$, $p = 0.029$). Similarly, Probe 4 ($\mu = 9.7 \times 10^{-3}$, $\sigma^2 = 0.36 \times 10^{-3}$) elicited a stronger response when compared with Probe 1 ($\mu = -2.7 \times 10^{-3}$, $\sigma^2 = 0.22 \times 10^{-3}$, $t(19) = 2.23$, $p = 0.019$) and 3 ($\mu = 1.3 \times 10^{-3}$, $\sigma^2 = 0.20 \times 10^{-3}$, $t(19) = 1.89$, $p = 0.037$), just as predicted. We further contrasted all the deviant probes (Probe 2 and Probe 4, $\mu = 8.8 \times 10^{-3}$, $\sigma^2 = 0.11 \times 10^{-3}$) with all the standard probes (Probe 1 and Probe 3, $\mu = -0.7 \times 10^{-3}$, $\sigma^2 = 0.14 \times 10^{-3}$), and found a significantly greater pupil dilation in the former ($t(19) = 2.87$, $p = 0.005$). These results are illustrated in Fig. 5.3.

5.4.1.2 Model-based analysis

Model comparison was performed using model-space averaging (FitzGerald et al., 2019), which strongly favoured M4 (i.e. the mixture of Gaussian with 4 components, see Table 5.1). This suggests that the distribution distortion caused by the probe tones (see Fig. 5.2) ended up leading participants to estimate two extra components, corresponding to Probe 2 and Probe 4. Nevertheless, this results clearly indicate that participants managed to form multimodal probabilistic representations of the stimuli, in contrast with classical PC's predictions.

As in Chapter 4, we tested for group-level effects of the GLM predictors with a permutation test by flipping the signs of the maximum a posteriori parameter estimates in a random subset of participants 100000 times and using the resultant surrogate data to provide surrogate two-tailed p-values for each variable (on which we applied a Bonferroni correction for multiple comparisons). We found a significant effect of surprise, drift (i.e. progressive decrease in dilation over time) and target (i.e. the stimulus being a target), replicating the findings described in Chapter 4 (see table 5.2 for details).

Model	$\sum VFE$	$\sum VFE - \sum VFE_{M1}$	**Posterior probability**	**Protected exceedance probability**
M1 (Unimodal Gaussian)	6857610	0	0.0515	0.0118
M2 (GMM with 2 components)	6858000	390	0.1939	0.0254
M3 (GMM with 3 components)	6858191	581	0.1956	0.0261
M4 (GMM with 4 components)	6858296	686	0.5589	0.9368

Table 5.1 In Experiment 1, model comparison strongly favoured the Gaussian mixture model (GMM) with 4 components, suggesting participants represented a distributions with 4 modes, violating classical PC. Model comparison was performed using model-space averaging, as described in FitzGerald et al. (2019).

5.4.2 EEG

5.4.2.1 Probe tones analysis

For this analysis we specified a design matrix as such

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{5.3}$$

with the four columns representing Probe 1, Probe 2, Probe 3 and Probe 4 respectively. We were thus contrasting all standard probes to all deviant probes. As we had a clear hypothesis (i.e. observing greater negativity in odd probes compared to standard probes, in accordance with the MMN and auditory oddball literature) all our tests were one-tailed.

The cluster permutation analysis testing for a MMN in the interval between 100 ms and 200 ms from stimulus onset revealed a positive effect ($p = 0.002$), particularly on fronto-central electrodes, as expected. We then investigated the single contrasts with the same cluster-based permutation approach, which revealed a greater negativity in

response to Probe 2 than Probe 1 ($p = 0.018$) and in response to Probe 4 than Probe 1 ($p = 0.012$) and Probe 3 ($p = 0.017$), suggesting that participants found Probe 2 to be more surprising than Probes 1 and 3, in line with our hypothesis and in contrast with classical PC. On the other end, contrary to our hypothesis, no significant difference was found between Probe 2 and Probe 3, although a nonsignificant negative cluster was identified by the analysis ($p = 0.177$). We also tested for a difference between all standard (Probe 1 and Probe 3) and all deviant probes (Probe 2 and Probe 4) and found greater negativity in deviant probes compared to standard ($p = 0.002$), suggesting, as expected, that deviant probes elicited greater surprise compared to standard ones. All these differences (both significant and non-significant) were stronger at fronto-central electrodes at around 150 ms from stimulus onset (see Fig. 5.4 for a scalp topography representation and Fig. 5.5 for a time-series representation), in accordance with the MMN literature (Garrido et al., 2009b; Näätänen et al., 2004).

5.5 Discussion

Taken together, the results of our model-free analyses suggest that participants were more surprised by (and thus assigned a lower probability to) deviant probes, which in turn suggests they managed to learn the real (bimodal) distribution of the stimuli, violating classical PC. This was clearer in Experiment 1, in which all the contrasts of interest resulted significant, then in Experiment 2, where the contrast between Probe 3 and Probe 2 did not reach significance. However, the analysis still revealed a nonsignificant MMN effect, which, taken together with the other (significant) contrasts, still supports the hypothesis of a bimodal stimulus encoding over a unimodal one. Taken together, the model-free analyses support our hypothesis, but the evidence is not terribly robust, as evidence by borderline (Experiment 1) or non-significant (Experiment 2) p-values.

Our model-based analysis provides an explanation for this. Model comparison strongly favoured a mixture of Gaussians over a unimodal Gaussian, but, somewhat surprisingly, the winning model was not the Gaussian mixture model with 2 components, as we expected, but the one with 4. This is almost certainly a result of the addition of the probe tones, which distorted the distribution by adding four point-masses, which participants represented as two additional components. This likely contributed to weaken the result of the model-free analysis in both experiments, as Probe 2 and Probe 4 were probably represented as the peak of two (smaller) Gaussian components. Despite this unexpected turn of events, our model-based analysis provides strong evidence of

the fact that our participants were capable of representing multimodal distributions, in violation with classical PC.

Whilst discussions about the precise neuronal inference scheme (or schemes) used in the brain may seem arcane, they have profound implications for understanding both human cognitive capacities and neurobiology. PC is perhaps the best worked-out probabilistic framework in neurobiological terms, and several aspects of these proposals have been highly influential.

Firstly, PC models based on prediction error minimisation have been applied to cortical computation (Bastos et al., 2012), providing an interpretation of cortical layers dynamics within and between cortical columns. These models do not account for non-normally distributed variables, which are inevitably misrepresented as Gaussians. Our results provide preliminary evidence that this is not the case in practice, suggesting these models should be augmented to better capture more complex (in our case multimodal) feature distributions.

Second, a core feature of theories based on PC is the central role placed on the encoding of precision (or inverse variance), and its putative signalling by the classical neuromodulators (Lawson et al., 2014; Moran et al., 2013) and other mechanisms (Bastos et al., 2012; Hovsepyan et al., 2020; van Pelt et al., 2016). In particular, abnormalities in precision have been used to provide elegant explanations for a range of psychopathology (Adams et al., 2013; Kube et al., 2020; Van de Cruys et al., 2014). From the perspective of classical PC this emphasis on precision makes a great deal of sense. Precision, and in particular the relative precision of different quantities within a model, is of fundamental importance within predictive coding, and is encoded multiplicatively (the role one would expect to be played by a neuromodulator). In addition, for Gaussian probability distributions, precision and entropy are monotonically related to one another, making precision encoding an ideal way of quantifying uncertainty. However, if the elegant structures placed upon neuronal function by classical PC are removed, it is less clear that precision signalling will play a key role in explaining cognitive processes and their dysfunction. In the first place, depending upon how neuronal inference is believed to occur, direct encoding of precision may play little or no role in the brain's inferential machinery (Aitchison and Lengyel, 2016). Moreover, for multimodal distributions like those considered here, precision and entropy are no longer monotonically related, and precision itself is much less useful as a way of quantifying uncertainty (to see this, compare the distributions in Fig. 5.1). This second observation leads to a clear experimental prediction – that where entropy and precision can be clearly dissociated,

neural encoding of uncertainty will largely reflect the former quantity. We intend to test this prediction in future work.

Parameter	**Prior mean (variance)**	**Group-level posterior (variance)**	***p*-value (corrected)**
$ln(h)$	1.6(0.5)	1.51(0.080)	-
$ln(l)$	1.6(0.5)	1.54(0.210)	-
$ln(d)$	$-3(4)$	$-3.87(4.064)$	-
a	1(0.1)	$1.00(< 0.001)$	-
w_{event}	0(2)	0.22(0.111)	0.004
$w_{surprise}$	0(2)	0.02(0.004)	0.006
w_{pitch}	0(2)	$0.01(< 0.001)$	0.961
w_{target}	0(2)	0.47(0.134)	< 0.001
w_{drift}	0(2)	$-0.22(0.046)$	< 0.001
m_1	$ln(500)(4)$	4.76(0.054)	-
m_2	$ln(500)(4)$	5.74(0.274)	-
m_3	$ln(500)(4)$	6.89(0.209)	-
m_4	$ln(500)(4)$	8.05(0.237)	-
$ln(\lambda_1)$	$ln(5)(4)$	2.35(1.100)	-
$ln(\lambda_2)$	$ln(5)(4)$	2.42(1.453)	-
$ln(\lambda_3)$	$ln(5)(4)$	2.68(0.678)	-
$ln(\lambda_4)$	$ln(5)(4)$	1.63(1.030)	-
$ln(\pi_2) - ln(\pi_1)$	0(4)	$-0.16(0.886)$	-
$ln(\pi_3) - ln(\pi_1)$	0(4)	0.09(0.677)	-
$ln(\pi_4) - ln(\pi_1)$	0(4)	$-0.11(0.291)$	-

Table 5.2 Summary statistics of the prior and group-level posterior distributions over each parameter (Experiment 1). Parameters were transformed where appropriate to enable use of a Gaussian prior distribution, as required by the Variational Laplace algorithm (ln indicates the natural logarithm). Non-parametric p-values calculated using permutation testing, Bonferroni-corrected for five comparisons. These provide clear evidence that both surprise and target presentation were consistently associated with pupil dilation, and for a progressive decrease ("drift") in dilation responses over the session.

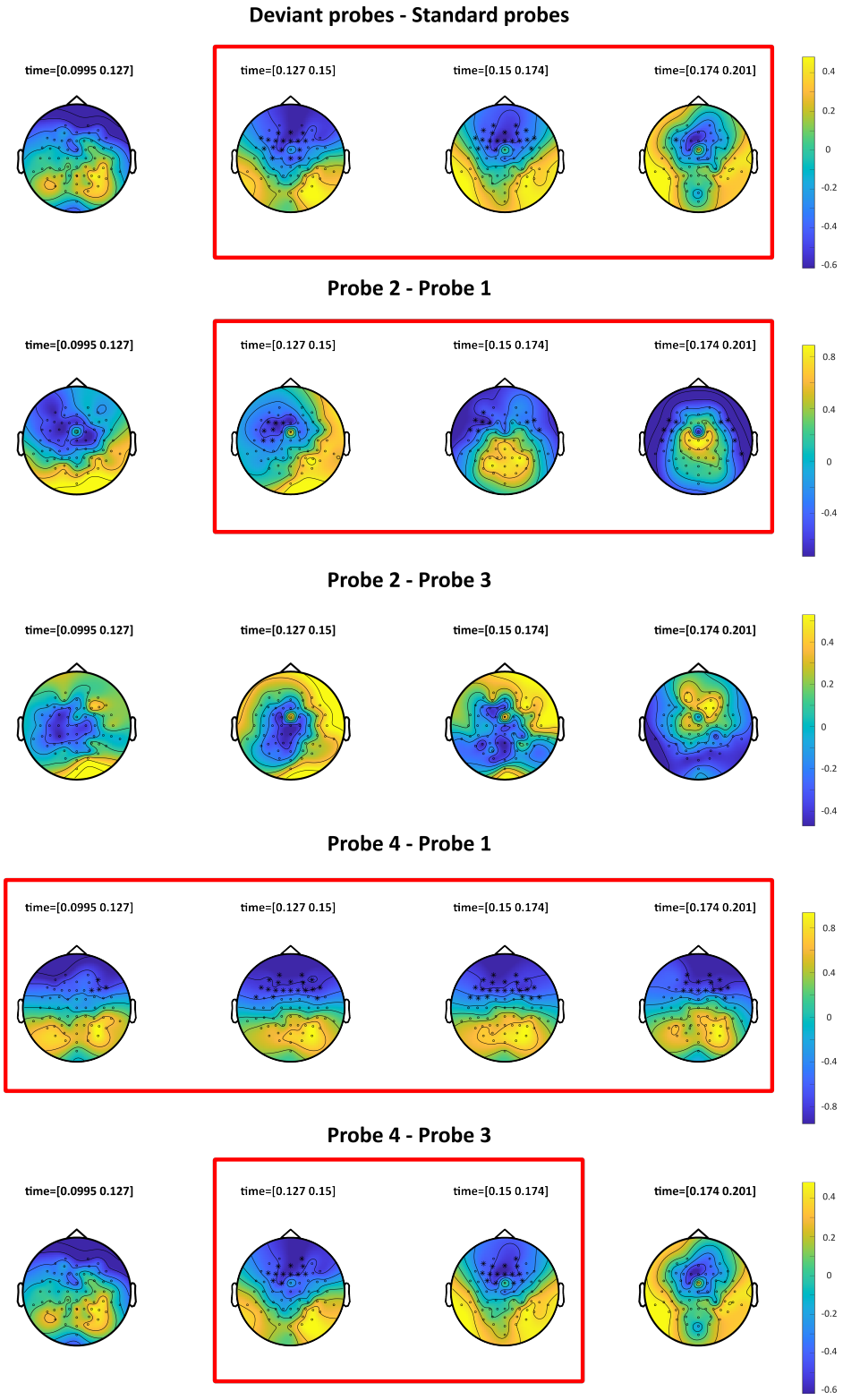

Fig. 5.4 This figure illustrates the scalp topography in various contrasts with the colour indicating the average electric potential difference (in μV) within the time window considered. Asterisks indicate electrodes that were in a significant cluster for at least half of the timepoints included within that particular time window (time windows including such electrodes are highlighted with a red rectangle). The significant negativities in fronto-central electrodes suggest a MMN effect.

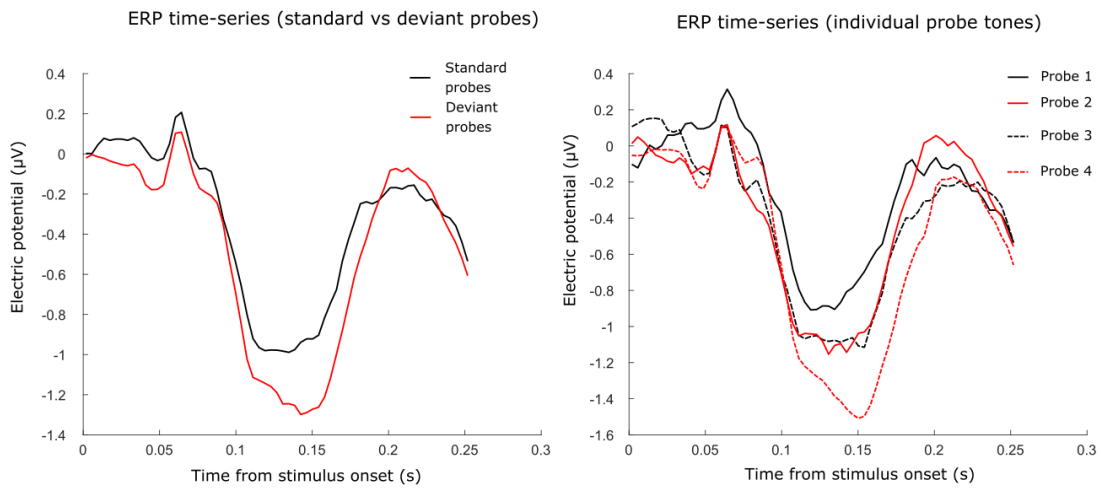

Fig. 5.5 This figure represents the ERP as a time series, contrasting responses from standard and deviant probes on the left and all 4 individual probes on the right. These are average responses across all relevant trials and across all the electrodes that were part of the cluster that reached statistical significance in the cluster-based permutation analysis contrasting the standard and deviant probe trials (see Fig. 5.4).

Chapter 6

Effects of retrospective inference on structure learning: a simulation study

6.1 Abstract

The ability to flexibly learn the structure of one's surroundings (structure learning) is crucial for adaptive behaviour. Use of an inaccurate model of the environment can lead to incorrect inferences, and thus maladaptive actions. Despite this, relatively little is understood about how structure learning occurs in human cognition. As a first step towards addressing this, we built on existing approaches to create an online clustering algorithm, in which we included a working memory component, allowing belief update about past stimuli (retrospective inference) in contrast with the more widespread fully online approach. We used this model to simulate behaviour on a novel structure learning task, where optimal performance required estimating the number and properties of discrete clusters of continuous stimuli. In this work we show how our algorithm outperforms a parametric one (i.e. with fixed number of clusters). We further demonstrate how retrospective inference benefits structure learning, with performance increasing with working memory capacity. We finally discuss trial-by trial measures that can be derived from our model, which provide testable predictions for future empirical studies.

6.2 Introduction

6.2.1 Structure Learning

Animals and humans live in diverse and complex environments, which they must navigate and act upon in order to survive. Certain behaviours might be advantageous in some situations and dangerous in others, making internal models of one's surrounding crucial for action selection. As more experience is gathered these models are updated (i.e. the individual learns) and action selection gets closer to optimal.

This idea has been formalised in probabilistic accounts of brain functions (Gershman and Beck, 2017; Knill and Pouget, 2004), according to which the brain does not have direct access to the states of the environment, but must infer them from the perturbations they cause in its activity. The brain is thus seen as an inference machine, combining prior knowledge with (noisy) sensory evidence to estimate the environment's (hidden) states. Formally, prior knowledge takes the form of a generative model, which specifies how the world's hidden states generate sensory observations. To infer the value of (time-varying) hidden states and update the generative model's (time-invariant) parameters the brain must use sensory evidence to perform Bayesian model inversion, which takes different forms depending on the specific theoretical framework.

There is a growing body of work investigating inference and learning as Bayesian model inversion in humans (De Berker et al., 2016; Diaconescu et al., 2017; FitzGerald et al., 2017; Mathys et al., 2011; Silvestrin et al., 2021). However, in all this work strong assumptions are made about the generative model. In particular, while the computational models allow parameters to be updated, the mathematical form (or structure) of these models is assumed to be known by the participants. This is unlikely to always be the case for human agents in real-world situations, where the underlying structure of the task at hand might need to be learned from scratch. We call the acquisition of such structure as a result of experience *structure learning* (Braun et al., 2010; Tenenbaum et al., 2011).

Structure learning is a general term that can be applied to many, more specific cognitive processes. One of these is representation or feature learning, consisting in reducing highly dimensional observations to a more contained number of useful and meaningful features (Austerweil and Griffiths, 2008; Wu et al., 2021). Another is causal learning, consisting in learning about the causal structure of a set of events and the strength of causal relationships (Gershman et al., 2017; Griffiths and Tenenbaum, 2005; Tenenbaum and Griffiths, 2001). Related to these is concept learning, or the acquisition of abstract categories and features (Lake et al., 2015; Smith et al., 2020)

and the relationship between these (Constantinescu et al., 2016; Mark et al., 2020; Whittington et al., 2020).

In this Chapter we focus on *clustering*, defined as unsupervised (i.e. in absence of feedback) categorisation of observations or events. For this to be a structure learning problem, the number of categories (clusters) is not known in advance, and the individual must therefore not only learn the characteristics of the clusters (i.e. its parameters), but also their number. In other words, the range of possible (discrete) values hidden states (which here represent cluster membership) can take must be learned. Therefore the number of components of the generative model, and thus its structure, is unknown, and when encountering a stimulus an agent must always consider the possibility of it belonging to a completely new cluster, effectively growing (i.e. adding a component) the generative model.

In a naturalistic setting this can happen in a variety of situations. For instance, in an unexplored environment an individual might encounter new species of animals or plants, and clustering them would be crucial to make generalisations about their most salient characteristics.

Past work on clustering focused mostly on giving new interpretations to known empirical phenomena (Gershman et al., 2010; Gershman and Niv, 2012; Gershman et al., 2017) and for investigating transfer learning (Collins and Frank, 2013, 2016; Franklin and Frank, 2018), mostly using stimuli with a discrete number of possible values. Apart from some notable exceptions (Collins and Frank, 2013, 2016; Davis et al., 2012), these studies were not concerned with fitting trial-by-trial participant data to online models, limiting the investigation of the learning process itself.

Here we build a modelling and experimental framework aimed at studying trial-by trial structure learning (in the form of clustering). Specifically we present a clustering task and a clustering model to simulate the behaviour of an artificial agent. The model contains a working memory component which we use to illustrate the benefits of retrospective inference fro structure learning (see next section). The model provides trial-by-trial behavioural outputs, so it could be fit to behavioural data of participants performing the task.

Specifically, we use mushrooms as an example, as when picking which ones to eat a certain knowledge of the different species is essential to select the edible ones and avoid the poisonous ones. One must thus group similar mushrooms into clusters in order to be able to generalise information about a single mushroom to all other mushrooms of the same species. If a porcino is delicious, all porcini will be.

The fact that the total number of clusters (mushroom species) is not known makes the task considerably more challenging. On one extreme, one might form a separate cluster for each individual mushroom they encounter (overfitting), leading to a loss of generalisation (the fact that one mushroom is tasty/poisonous says nothing about other individual mushrooms). On the other, one might cluster together all mushrooms (underfitting), leading to over-generalisation (if one mushroom is tasty/poisonous, then all mushrooms are tasty/poisonous). It is therefore crucial for embodied agents to be equipped with a cognitive apparatus that allows them to learn the most convenient model structure.

6.2.2 Retrospective Inference

In absence of feedback, one will never be completely certain of having assigned an item to the right cluster. In our mushrooms example, this is especially true for inexperienced individuals, who do not have a clear idea of how many mushrooms species they might encounter nor of the characteristics of the various species. Some mushrooms might very easily be misclassified, which would result in distorted representations, as cluster parameters would be updated with irrelevant data. Too many clusters might be formed, resulting in overfitting, or too few, resulting in underfitting. A memoryless approach (i.e. updating the generative model and forgetting individual stimuli) might therefore come at a cost for structure learning. At the other extreme, holding in memory every single mushroom one has ever seen allows to continuously rethink cluster membership, resulting in optimal inference and learning, but it quickly becomes cognitively infeasible as the number of encountered mushrooms increases.

As a reasonable compromise, one might hold in working memory a finite amount of stimuli, and re-evaluate them in light of new information, making cluster assignment and, as a consequence, structure learning, more accurate. This has been formalised (FitzGerald et al., 2020) as Finite Retrospective Inference (FRI), an online modelling approach which allows agents a certain working memory capacity, which they use to re-perform inference on the most recent hidden states (cluster membership in our case) using new evidence. This can be thought of as a sliding cognitive window containing the last a encountered mushrooms (with a being the maximum amount of stimuli one can or is willing to hold in working memory), with only the oldest item y_{i-a+1} being used for updating the model's parameters at each trial. After the update, this item is "forgotten", disappearing from memory and effectively "making room" for a new stimulus. The clear advantage of this approach is that hidden states of a stimulus y_i (in our case cluster responsibilities $\mathbf{r}_i$) are evaluated for the last and definitive time

at trial $i+a-1$, thus having more information at one's disposal (see Fig. 6.1) and subsequently making inference and learning more accurate (FitzGerald et al., 2020).

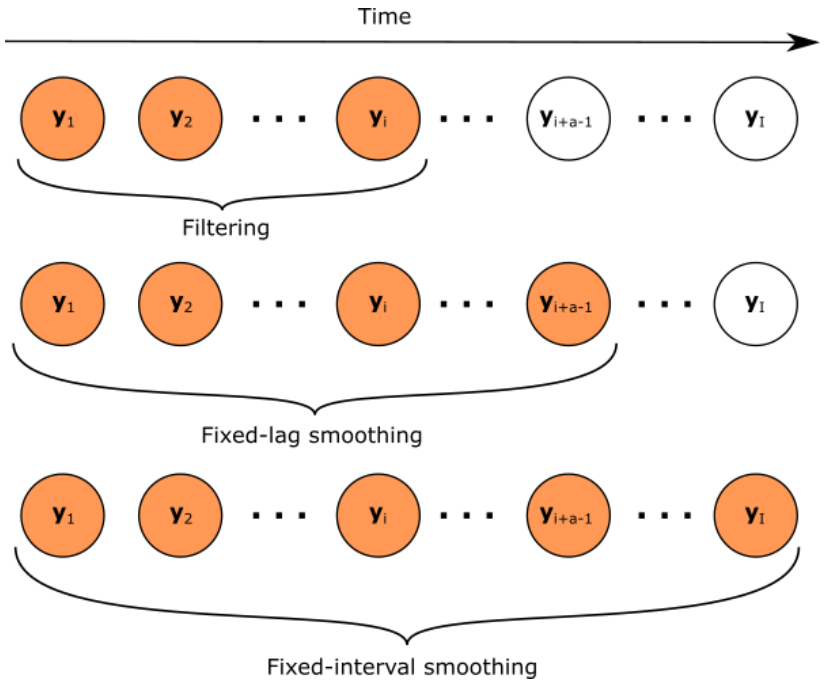

Fig. 6.1 Schematic representation of filtering, fixed-lag smoothing and fixed-interval smoothing. In the filtering model (top) cluster responsibilities for a stimulus $\mathbf{y}_i$ are evaluated using only information coming from $\mathbf{y}_{1:i}$. In fixed-lag smoothing (middle) the algorithm keeps updating cluster responsibilities of stimulus $\mathbf{y}_i$ until it slides out of its cognitive window, meaning final responsibilities are evaluated taking into account $\mathbf{y}_{1:i+a-1}$. Finally, in fixed-interval smoothing the algorithm remembers all data points individually, and updates their cluster responsibilities (including e.g. for $\mathbf{y}_i$) until it sees the last stimulus $\mathbf{y}_I$, therefore using the whole dataset $\mathbf{y}_{1:I}$.

In this paper we describe a novel computational model of clustering which we hope to deploy in empirical settings in the future. We utilise a non-parametric approach (Gershman and Blei, 2012) for priors about cluster membership and incorporate FRI. We then perform simulations on a novel task to illustrate the importance of learning the generative model's structure, as well as the benefits of FRI. We finally describe a few useful trial-by-trial metrics that can be derived from our model, which provide testable predictions for empirical investigations. What follows is a formal description of our algorithm.

6.3 Methods

6.3.1 Modelling

The algorithm we describe here is an online non-parametric adaptation of the mean-field Variational Mixture of Gaussians describerd in Bishop (2006) incorporating retrospective inference on a fixed-length window. We compare its behaviour to a simpler filtering model (i.e. fully online, without retrospective inference) and a parametric equivalent (i.e. with a fixed number of clusters).

Note that the choice of mean-field Variational Inference is not intended to imply a claim about the specific computations the brain would perform in an analogous situation. We used a variational algorithm for the sake of continuity with the rest of the work presented in this thesis, but similar results could be obtained with different approaches, such as sampling methods (Mackay, 1998) or expectation propagation (Minka, 2013).

6.3.1.1 Generative model

Our model assumes each stimulus is sampled from one of several clusters with Gaussian form, but remains agnostic about the number of possible clusters. In other words, the generative model is a mixture of Gaussians with an unknown number of components. In our simulation the stimuli are univariate, but here we describe the multivariate generalization of our algorithm.

In Gaussian mixture models (GMM) the likelihood function of any stimulus $\mathbf{y}$ is

$$p(\mathbf{y}) = \sum_{n=1}^{N} \pi_n \mathcal{N}(\mathbf{y} \mid \mathbf{m}_n, \mathbf{\Lambda}_n^{-1}) \tag{6.1}$$

with N being the (unknown) total number of clusters, $\boldsymbol{\pi}$ the mixture components, $\mathbf{m}$ the means and $\mathbf{\Lambda}$ the precision matrices.

We now introduce a new set of binary variables $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_I\}$. These represent the time-varying hidden state (cluster membership) that must be inferred. Its value must satisfy $c_{i,n} \in \{0, 1\}$ and $\sum_{n=1}^{N} c_{i,n} = 1$ and its probability distribution is specified as:

$$p(\mathbf{c} \mid \boldsymbol{\pi}) = \prod_{i=1}^{I} \prod_{n=1}^{N} \pi_n^{c_{i,n}} \tag{6.2}$$

with I being the total number of trials.

We can now write the conditional distribution of $\mathbf{y}$ as

$$p(\mathbf{y} \mid \mathbf{c}, \mathbf{m}, \mathbf{\Lambda}) = \prod_{i=1}^{I} \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_{i,n} \mid \mathbf{m}_n, \mathbf{\Lambda}_n^{-1})^{c_{i,n}} \tag{6.3}$$

In a Variational Inference setting it is convenient to make use of conjugate prior distributions, ensuring that prior and posterior have the same form. Thus the prior probability over the mixture components $\boldsymbol{\pi}$ is given by a Dirichlet distribution

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \tag{6.4}$$

with $\boldsymbol{\alpha}$ being a vector of parameters with N elements.

Similarly, the prior over $\mathbf{m}$ and $\mathbf{\Lambda}$ is given by a Gaussian-Wishart distribution (see Bishop, 2006 for details)

$$p(\mathbf{m}, \mathbf{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\big(\mathbf{m}_n \mid \boldsymbol{\mu}_n, (\beta_n \mathbf{\Lambda}_n)^{-1}\big) \mathcal{W}(\mathbf{\Lambda}_n \mid \mathbf{W}_n, v_n) \tag{6.5}$$

Thus the full joint can be written as

$$\begin{aligned} p(\mathbf{y}, \mathbf{c}, \boldsymbol{\pi}, \mathbf{m}, \mathbf{\Lambda}) =& p(\mathbf{y} \mid \mathbf{c}, \mathbf{m}, \mathbf{\Lambda}) p(\mathbf{c} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\mathbf{m} \mid \mathbf{\Lambda}) p(\mathbf{\Lambda}) \\ =& Dir(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} \Big\{ \mathcal{N}\big(\mathbf{m}_n \mid \boldsymbol{\mu}_n, (\beta_n \mathbf{\Lambda}_n)^{-1}\big) \mathcal{W}(\mathbf{\Lambda}_n \mid \mathbf{W}_n, v_n) \\ & \prod_{i=1}^{I} \pi_n^{c_{i,n}} \mathcal{N}(\mathbf{y}_{i,n} \mid \mathbf{m}_n, \mathbf{\Lambda}_n^{-1})^{c_{i,n}} \Big\} \end{aligned} \tag{6.6}$$

A graphical representation of the generative model can be found in Fig. 6.2 (which includes the models inverted by both filtering and retrospective inference agents, see below).

6.3.1.2 Filtering

Using equation 6 for inference and learning requires holding in memory all individual stimuli up to the current one (i.e. fixed-interval smoothing). This entails optimising trial-by trial Variational Free Energy (VFE) formulated as

$$VFE = E_{q(\mathbf{c}_{1:i}, \boldsymbol{\pi}, \mathbf{m}, \mathbf{\Lambda})} \left[ln \left(\frac{p(\mathbf{y}_{1:i}, \mathbf{c}_{1:i}, \boldsymbol{\pi}, \mathbf{m}, \mathbf{\Lambda})}{q(\mathbf{c}_{1:i}, \boldsymbol{\pi}, \mathbf{m}, \mathbf{\Lambda})} \right) \right] \tag{6.7}$$

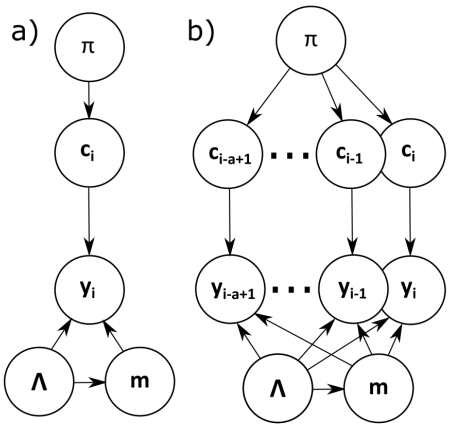

Fig. 6.2 Graphs representing causal dependencies in the generative model. Specifically, (a) represents the generative model to be inverted by agents performing simple filtering (see Section 6.3.1.2), while (b) represent the generative model to be inverted by agents performing retrospective inference with a working memory of size a (see Section 6.3.1.3). The only difference between the two is the number of data points simultaneously considered during inference (1 for the former and a for the latter). Arrow direction specifies the directionality of the causal relationship.

with $q(\cdot)$ being the approximate posteriors. The algorithm would thus need to revisit all the previous stimuli at each trial, which is in practice infeasible in a naturalistic setting.

A common solution to this is modelling learning as an online update of the priors, with the posterior at trial t becoming the prior at trial $t+1$. The quantity to optimise then becomes

$$VFE = E_{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_i,\mathbf{c}_t,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] \tag{6.8}$$

This approach is known as filtering, and does not require the agent to have any memory of previous individual stimuli, which are all encapsulated in the priors.

In our model VFE is maximised with an EM (expectation maximization) algorithm, based on recursive approximate posteriors update until convergence. In a setting where the number of clusters is known, the algorithm would simply evaluate $q^*(\mathbf{c}_i)$ (*E step*) to obtain cluster responsibilities $\mathbf{r}_i$, which would in turn be used to evaluate $q^*(\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})$(*M step*). Note that here $q^*(\cdot)$ represents the optimal solution for the approximate posteriors. The algorithm would then go back to the E step and update responsibilities, and use those for a new M step, and repeat the cycle until VFE converges. As in our case the agent does not know the number of clusters, we added two additional steps to the algorithm: one before the EM cycle (*Cluster formation*)

and one inside of it (*Cluster pruning*), between the E and M steps (see Fig. 6.5 for a graphical overview).

Cluster formation As a new stimulus $\mathbf{y}_i$ is presented, the model forms a new candidate cluster with index N. Here N always represents the index of the most recently formed cluster.

$$N \longleftarrow N+1 \tag{6.9}$$

Its posterior parameters $\breve{\boldsymbol{\theta}}_{i,N}$ are initialised as:

$$\breve{\boldsymbol{\mu}}_{i,N} = \mathbf{y}_i \tag{6.10}$$

$$\breve{\beta}_{i,N} = \beta_0 \tag{6.11}$$

$$\breve{\mathbf{W}}_{i,N} = \mathbf{W}_0 \tag{6.12}$$

$$\breve{v}_{i,N} = v_0 \tag{6.13}$$

$$\breve{\alpha}_N = \alpha_0 \tag{6.14}$$

Before any stimulus is presented $N = 0$, therefore the first cluster will be centred around the first stimulus.

E step In the E step the algorithm evaluates

$$lnq^*(\mathbf{c}_i) = E_{q(\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_i,\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] + const \tag{6.15}$$

From here it can be shown (Bishop, 2006) that

$$\begin{aligned}\breve{r}_{i,n} &= \frac{\breve{\rho}_{i,n}}{\sum_{j=1}^{N}\breve{\rho}_{i,j}} \\ &= E_{q(\pi_n,\mathbf{m}_n,\boldsymbol{\Lambda}_n)}[c_{i,n}]\end{aligned} \tag{6.16}$$

where

$$\begin{aligned}ln\breve{\rho}_{i,n} =& \psi(\breve{\alpha}_{i,n}) - \psi\left(\sum_{j=1}^{N}\breve{\alpha}_{i,j}\right) + \sum_{d=1}^{D}\psi\left(\frac{\breve{v}_{i,n}+1-d}{2}\right) \\ &+ Dln2 + ln|\breve{\mathbf{W}}_{i,n}| - \frac{D}{2\breve{\beta}_{i,n}} - \frac{\breve{v}_{i,n}}{2}(\mathbf{y}_i - \breve{\boldsymbol{\mu}}_{i,n})^T\breve{\mathbf{W}}_{i,n}(\mathbf{y}_t - \breve{\boldsymbol{\mu}}_{i,n})\end{aligned} \tag{6.17}$$

Here $\psi(\cdot)$ is the digamma function, D is the dimensionality of the stimulus $\mathbf{y}_i$ and $\breve{\boldsymbol{\theta}}_i$ are the posterior cluster parameters evaluated at trial i, which are set to the prior values $\tilde{\boldsymbol{\theta}}_i$ in the first iteration of the E step. By $\breve{\mathbf{r}}$ we denote the cluster responsibilities which are going to be used to evaluate $\breve{\boldsymbol{\theta}}$ (this will become relevant once retrospective inference is introduced, see section 6.3.1.3 below).

Cluster pruning At every iteration, between the E step and the M step, the algorithm goes through a pruning function, cutting any cluster n for which

$$\frac{\breve{\alpha}_{i,n}}{\sum_{j=1}^{N}\breve{\alpha}_{i,j}} < \varepsilon \tag{6.18}$$

or

$$\breve{r}_{l,n} \neq max\big(\mathbf{r}_l\big) \quad \forall l \in \{1,...,i\} \tag{6.19}$$

with ε being a threshold probability which we set at .02.

Put more simply, the algorithm eliminates all clusters with a very low mixing component or clusters which no encountered stimulus would be assigned to with the highest probability. Note that in a filtering model all the latter condition does is taking care of the cluster formed at trial i (i.e. cluster $N+1$) whenever it is unnecessary, and does not need to revisit responsibilities of previous trials, as they are never re-evaluated. This is not the case for retrospective inference models (see below).

M step In the M step the algorithm updates the parameters of the surviving clusters. It thus evaluates

$$lnq^*(\boldsymbol{\pi}) = E_{q(\mathbf{c}_i,\mathbf{m},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_i,\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] + const \tag{6.20}$$

$$lnq^*(\mathbf{m} \mid \boldsymbol{\Lambda}) = E_{q(\mathbf{c}_i,\boldsymbol{\pi},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_i,\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] + const \tag{6.21}$$

and

$$lnq^*(\boldsymbol{\Lambda}) = E_{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m})}\left[ln\left(\frac{p(\mathbf{y}_i,\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_i,\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] + const \tag{6.22}$$

from which the following update equations can be derived (Bishop, 2006)

$$\breve{\alpha}_{i,n} = \tilde{\alpha}_{i,n} + \breve{r}_{i,n} \tag{6.23}$$

$$\breve{\beta}_{i,n} = \tilde{\beta}_{i,n} + \breve{r}_{i,n} \tag{6.24}$$

$$\breve{\boldsymbol{\mu}}_{i,n} = \frac{\tilde{\beta}_{i,n}\tilde{\boldsymbol{\mu}}_{i,n} + \breve{r}_{i,n}\mathbf{y}_i}{\breve{\beta}_{i,n}} \tag{6.25}$$

$$\breve{\mathbf{W}}_{i,n}^{-1} = \widetilde{\mathbf{W}}_{i,n}^{-1} + \frac{\tilde{\beta}_{i,n}\breve{r}_{i,n}}{\tilde{\beta}_{i,n} + \breve{r}_{i,n}}(\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_{i,n})(\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_{i,n})^T \tag{6.26}$$

$$\breve{v}_{i,n} = \tilde{v}_{i,n} + \breve{r}_{i,n} \tag{6.27}$$

These updated parameters will then be used in the next iteration of the E step for evaluating responsibilities.

Update The algorithm iteratively carries out the E step, Cluster pruning and M step until the value of VFE (rounded to the 6th decimal place) converges. The updated cluster parameters obtained at the end of this process will then become the priors for trial $i+1$.

$$\tilde{\alpha}_{i+1,n} = \breve{\alpha}_{i,n} \tag{6.28}$$

$$\tilde{\beta}_{i+1,n} = \breve{\beta}_{i,n} \tag{6.29}$$

$$\tilde{\boldsymbol{\mu}}_{i+1,n} = \breve{\boldsymbol{\mu}}_{i,n} \tag{6.30}$$

$$\tilde{\mathbf{W}}_{i+1,n} = \breve{\mathbf{W}}_{i,n} \tag{6.31}$$

$$\tilde{v}_{i+1,n} = \breve{v}_{i,n} \tag{6.32}$$

6.3.1.3 Retrospective Inference

Compared to fixed-interval smoothing, filtering decreases the computational burden, but it comes with a cost in inference and learning accuracy.

A mid-way alternative has been proposed (FitzGerald et al., 2020), called fixed-lag smoothing (Särkkä, 2013). Here the algorithm has a fixed working memory capacity a, and at each trial initially optimises the following quantity

$$VFE = E_{q(\mathbf{c}_{i-a+1:i},\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_{i-a+1:i},\mathbf{c}_{i-a+1:i},\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}{q(\mathbf{c}_{i-a+1:i},\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\right)\right] \tag{6.33}$$

As in the filtering model outlined above, the algorithm first forms a new cluster centred around $\mathbf{y}_i$ before iteratively evaluating the optimal solutions for approximate posteriors $q^*(\cdot)$ until convergence.

Similarly, the E step is carried out as described above for all the stimuli $\mathbf{y}_{i-a+1:i}$ in the cognitive window, evaluating responsibilities separately.

As for Cluster pruning, retrospective inference has the only effect of increasing to a the number of stimuli for which responsibilities need to be checked so that newly formed clusters can be readily pruned if not necessary.

Conversely, the update equations in the M step are adapted to multiple stimuli as follows:

$$\widehat{\alpha}_{i,n} = \tilde{\alpha}_{i,n} + \phi_{i,n} \tag{6.34}$$

$$\widehat{\beta}_{i,n} = \tilde{\beta}_{i,n} + \phi_{i,n} \tag{6.35}$$

$$\widehat{\boldsymbol{\mu}}_{i,n} = \frac{\tilde{\beta}_{i,n}\tilde{\boldsymbol{\mu}}_{i,n} + \phi_{i,n}\bar{\mathbf{y}}_{i,n}}{\widehat{\beta}_n} \tag{6.36}$$

$$\widehat{\mathbf{W}}_{i,n}^{-1} = \widetilde{\mathbf{W}}_{i,n}^{-1} + \frac{\tilde{\beta}_{i,n}\phi_{i,n}}{\tilde{\beta}_{i,n} + \phi_{i,n}}(\bar{\mathbf{y}}_{i,n} - \tilde{\boldsymbol{\mu}}_{i,n})(\bar{\mathbf{y}}_{i,n} - \tilde{\boldsymbol{\mu}}_{i,n})^T + \phi_{i,n}\mathbf{S}_{i,n} \tag{6.37}$$

$$\tilde{\upsilon}_n = \upsilon_n + \phi_{i,n} \tag{6.38}$$

with

$$\phi_{i,n} = \sum_{l=i-a+1}^{i} \widehat{r}_{l,n} \tag{6.39}$$

$$\bar{\mathbf{y}}_{i,n} = \frac{1}{\phi_{i,n}} \sum_{l=i-a+1}^{i} \widehat{r}_{l,n}\mathbf{y}_l \tag{6.40}$$

and

$$\mathbf{S}_{i,n} = \frac{1}{\phi_{i,n}} \sum_{l=i-a+1}^{i} \tilde{r}_{l,n}(\mathbf{y}_l - \bar{\mathbf{y}}_{i,n})(\mathbf{y}_l - \bar{\mathbf{y}}_{i,n})^T \tag{6.41}$$

Here $\widehat{\boldsymbol{\theta}}_i$ denote the sufficient statistics of the temporary variational posteriors evaluated by updating the prior parameters $\tilde{\boldsymbol{\theta}}_i$ with $\mathbf{y}_{i-m+1:i}$. Similarly $\widehat{\mathbf{r}}_{i-a+1:i}$ represent the temporary cluster responsibilities used to then estimate $\widehat{\boldsymbol{\theta}}_i$, which, in turn, will be used to estimate $\widehat{\mathbf{r}}_{i-a+1:i}$ in the subsequent iteration of the E step (note that as in the filtering model in the first iteration of the E step $\widehat{\boldsymbol{\theta}}_i$ are initialised as $\tilde{\boldsymbol{\theta}}_i$).

After convergence, the algorithm discards all the temporary parameters $\widehat{\boldsymbol{\theta}}_i$ and responsibilities $\widehat{\mathbf{r}}_{i-a+2:i}$, only keeping $\widehat{r}_{i-a+1}$, which is then used to optimise

$$VFE = E_{q(\boldsymbol{\pi},\mathbf{m},\boldsymbol{\Lambda})}\left[ln\left(\frac{p(\mathbf{y}_{i-a+1}, \boldsymbol{\pi}, \mathbf{m}, \boldsymbol{\Lambda} \mid \breve{\mathbf{r}}_{i-a+1})}{q(\boldsymbol{\pi}, \mathbf{m}, \boldsymbol{\Lambda})}\right)\right] \tag{6.42}$$

where it sets

$$\breve{\mathbf{r}}_{i-a+1} = \widehat{\mathbf{r}}_{i-a+1} \tag{6.43}$$

At this point, as responsibilities $\check{\mathbf{r}}_{i-a+1}$ have already been inferred optimally, the only thing left to do for the algorithm is to carry out the M step once, using $\mathbf{y}_{i-a+1}$ and $\check{\mathbf{r}}_{i-a+1}$ as described in the Filtering paragraph.

The model thus infers cluster parameters with only the oldest element in the cognitive window $\mathbf{y}_{i-a+1}$, which is going to slide out of it in the next trial. This means that to definitely estimate cluster responsibilities for a stimulus $\mathbf{y}_i$, the agent uses information coming from $\mathbf{y}_{1:i+a-1}$, making inference about cluster membership and cluster parameters estimation (which depend on estimated responsibilities) more accurate, as new stimuli can inform inference about old ones.

Note that both filtering and fixed-interval smoothing are special cases of fixed-lag smoothing, in which $a = 1$ and $a = I$ respectively.

6.3.2 Simulation experiment

6.3.2.1 Task and stimuli

We developed a variant of the rule-plus-exception task (Davis et al., 2012) with continuously-varying stimuli. Each trial the simulated agent was presented with a stimulus (a mushroom) and had to determine whether it was edible (good) or poisonous (bad). After the decision was made, the agent received feedback.

Unknown to the agent, there were two different species of good mushrooms and two different species of bad mushrooms, which varied only on one continuous dimension (size), sampled from a species-specific Gaussian distribution. The resulting probability distributions of good and bad mushrooms sizes thus were two mixtures of Gaussians (see Fig. 6.3 and Table 6.1). We chose our stimuli to vary on only one dimension to make visualisation clearer on one hand, and to maximise ambiguity of unimodal approximations of the stimuli distribution, on the other. These distributions were chosen so that a unimodal representation would result in very high overlap, hurting the agent's performance and punishing underfitting.

The agent carried out 30 trials, during which it was presented with 15 good mushrooms (10 from Species 1 and 5 from Species 2) and 15 bad mushrooms (10 from Species 3 and 5 from Species 4) in random order. The simulation was carried out 1000 times, and each time the stimuli were re-sampled from the true distributions.

Due to the nature of our task, our algorithm is organised in two steps: a *decision-making step* and a *learning step* (see Fig. 6.5). Both use VFE maximisation as described above to infer the value of model parameters, but only the latter step is used for belief

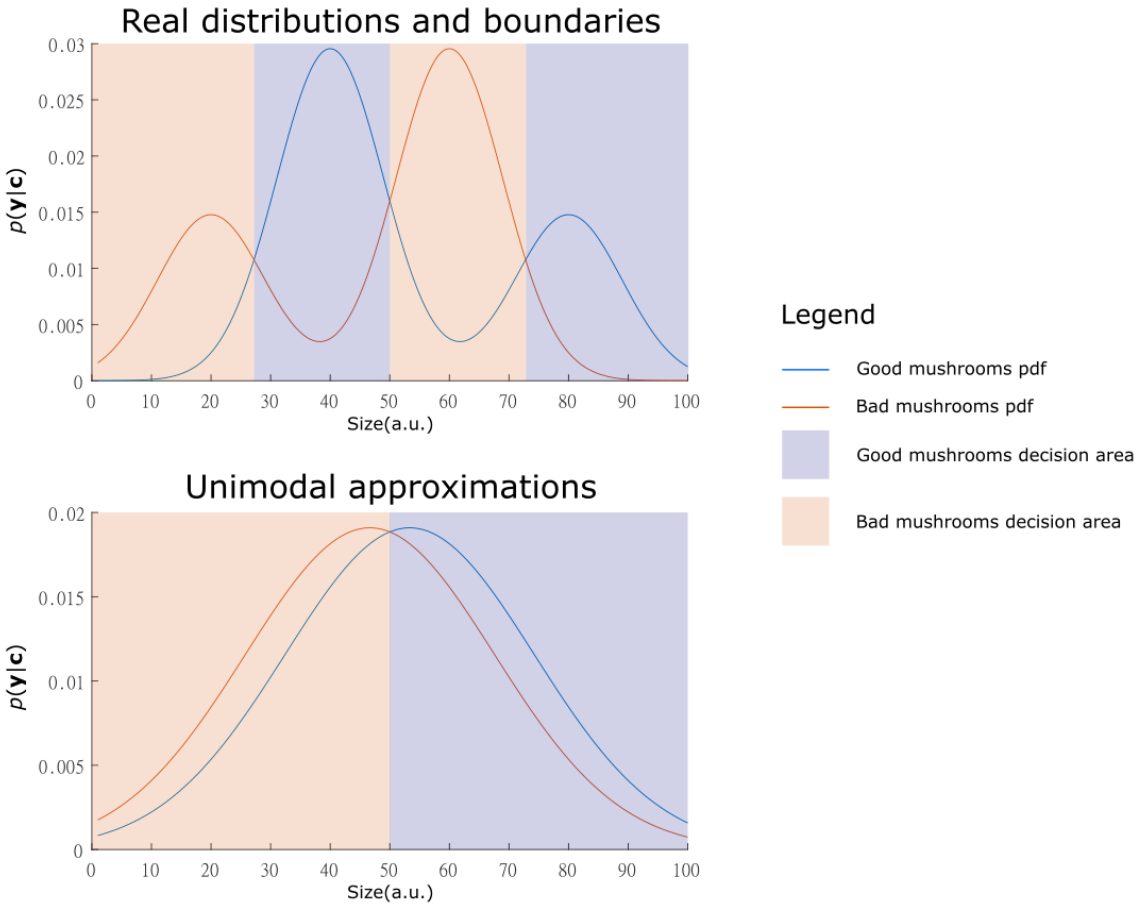

Fig. 6.3 Real probability distributions from which the stimuli were sampled (top) and unimodal approximations (bottom). These distributions were chosen to maximise the overlapping of unimodal approximation and punish underfitting agents.

update (as it can use the information provided by the feedback). In the next sections we outline these steps, incorporating retrospective inference.

6.3.2.2 Decision-making

As specified above, at each stimulus presentation the agent forms a new cluster centred around it. After this, the agent iteratively evaluates pre-feedback approximate posteriors through the EM loop. Importantly, at this stage there is no Cluster pruning.

During the E step the agent estimates the pre-feedback cluster responsibilities $\hat{\mathbf{r}}_i$ as in equations 4.16-17 for the current stimulus $\mathbf{y}_i$. Crucially this is not the case for $\hat{\mathbf{r}}_{i-a+1:i-1}$, as feedback about $\mathbf{y}_{i-a+1:i-1}$ has already been provided in trials

Parameter	**Bad mushrooms**		**Good mushrooms**	
	Species 1	Species 2	Species 3	Species 4
π	0.1667	0.3333	0.1667	0.3333
m	20	60	40	80
σ^2	81	81	81	81

Table 6.1 True sufficient statistics of the Gaussians (clusters) from which the stimuli were sampled

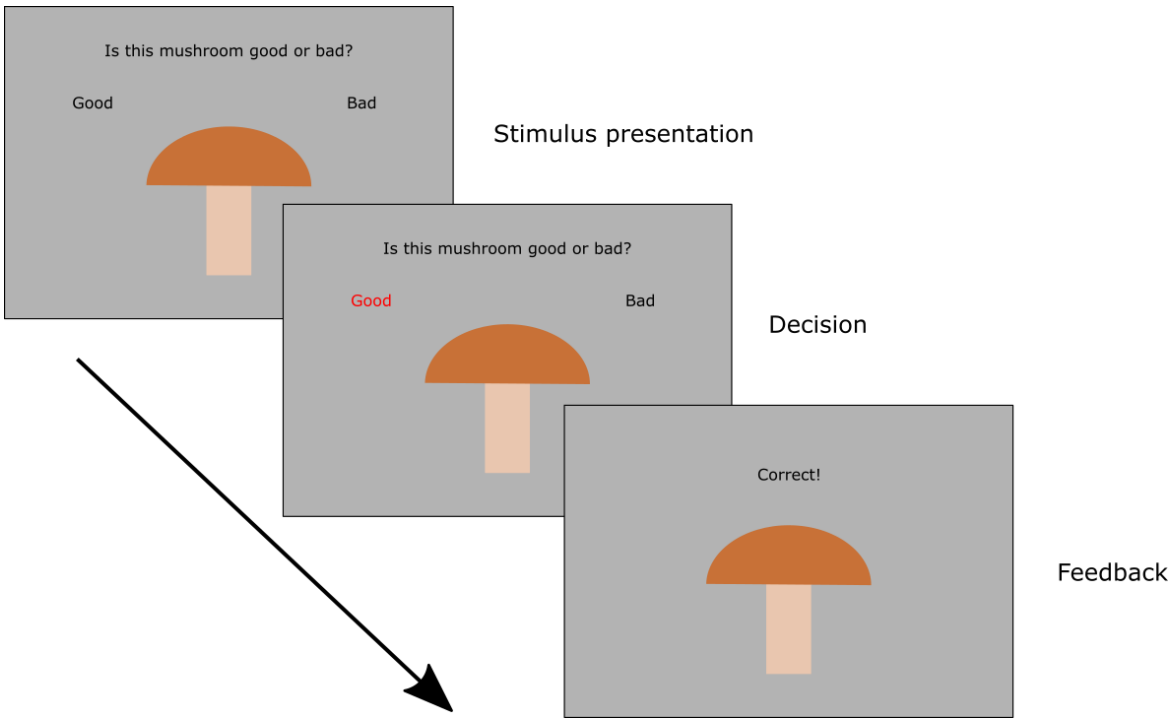

Fig. 6.4 Schematic representation of the task. First the agent is presented with a stimulus (a mushroom), and it has to infer whether it is edible (good) or poisonous (bad). After making a decision, it receives feedback and updates its beliefs.

$i-a+1:i-1$, and therefore the agent remembers whether those mushrooms were good or bad. To deal with this, we introduce the index vector of clusters previously labelled (see Learning paragraph) as "bad" $\tilde{\mathbf{b}}_i$ and that of clusters previously labelled as "good" $\tilde{\mathbf{g}}_i$ i.e. (mushroom species known to be bad or good at the beginning of trial i). Similarly, we introduce index vectors $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{k}}_i$, with the former including the indices of bad mushrooms and the latter those of good mushrooms in working memory at trial i. The current trial i is not included in either $\tilde{\mathbf{z}}_i$ or $\tilde{\mathbf{k}}_i$, the agent has not received feedback on it yet.

Then we have

$$\hat{r}_{l,n} = \begin{cases} \frac{\hat{\rho}_{l,n}}{\hat{\rho}_{l,N} + \sum_{j \in \mathbf{b}_i} \hat{\rho}_{l,j}} & \text{if} \quad l \in \mathbf{z} \quad \text{and} \quad \big(n \in \mathbf{b}_i \quad \text{or} \quad n = N\big) \\ \frac{\hat{\rho}_{l,n}}{\hat{\rho}_{l,N} + \sum_{j \in \mathbf{g}_i} \hat{\rho}_{l,j}} & \text{if} \quad l \in \mathbf{k} \quad \text{and} \quad \big(n \in \mathbf{g}_i \quad \text{or} \quad n = N\big) \\ 0 & \text{if} \quad \big(l \in \mathbf{z} \quad \text{and} \quad n \in \mathbf{g}_i\big) \quad \text{or} \quad \big(l \in \mathbf{k} \quad \text{and} \quad n \in \mathbf{b}_i\big) \end{cases} \tag{6.44}$$

Note that the new cluster $N+1$ is considered for both good and bad mushrooms, as it does not have a label yet. The M step is then carried out as described in equations 6.23-27.

Once the pre-feedback Variational Free Energy converges (rounded to 6 decimal places), the agent evaluates the probabilities of $\mathbf{y}_i$ being either bad or good as such:

$$p(\zeta_i = 0) = \frac{1}{2}\hat{r}_{i,N+1} + \sum_{j \in \mathbf{b}_i} \hat{r}_{i,j} \tag{6.45}$$

$$p(\zeta_i = 1) = \frac{1}{2}\hat{r}_{i,N+1} + \sum_{j \in \mathbf{g}_i} \hat{r}_{i,j} \tag{6.46}$$

where

$$\zeta_i = \begin{cases} 0 & \text{if} \quad \mathbf{y}_i \quad \text{is bad} \\ 1 & \text{if} \quad \mathbf{y}_i \quad \text{is good} \end{cases} \tag{6.47}$$

Again, as the new cluster $N+1$ has not jet been labelled, the agent remains agnostic on whether it is good or bad (assigning these possibilities a probability of 0.5 each), hence the first term of equations 6.45 and 6.46. Put it more simply, the probability of a mushroom being good is the sum of the probabilities of that mushroom belonging to the known edible species, plus the probability of it belonging to an unknown, edible species. Whilst the former is a straightforward sum (second term of equation 6.46), the latter is the joint probability of the mushroom belonging to a new species *and* the new species being edible (first term of the equation).

6.3.2.3 Learning

After feedback the true value of ζ_i becomes known, so the agent assigns a label to the new cluster and the new stimulus based on whether $\mathbf{y}_i$ was a bad or good mushroom.

Thus:

$$\breve{\mathbf{b}}_i = \begin{cases} \begin{bmatrix} \tilde{\mathbf{b}}_i \\ N+1 \end{bmatrix} & \text{if} \quad \zeta_i = 0 \\ \tilde{\mathbf{b}}_i & \text{if} \quad \zeta_i = 1 \end{cases} \tag{6.48}$$

$$\breve{\mathbf{z}}_i = \begin{cases} \begin{bmatrix} \tilde{\mathbf{z}}_i \\ i \end{bmatrix} & \text{if} \quad \zeta_i = 0 \\ \tilde{\mathbf{z}}_i & \text{if} \quad \zeta_i = 1 \end{cases} \tag{6.49}$$

$$\breve{\mathbf{g}}_i = \begin{cases} \begin{bmatrix} \tilde{\mathbf{g}}_i \\ N+1 \end{bmatrix} & \text{if} \quad \zeta_i = 1 \\ \tilde{\mathbf{g}}_i & \text{if} \quad \zeta_i = 0 \end{cases} \tag{6.50}$$

$$\breve{\mathbf{k}}_i = \begin{cases} \begin{bmatrix} \tilde{\mathbf{k}}_i \\ i \end{bmatrix} & \text{if} \quad \zeta_i = 1 \\ \tilde{\mathbf{k}}_i & \text{if} \quad \zeta_i = 0 \end{cases} \tag{6.51}$$

with

$$\tilde{\mathbf{b}}_{i+1} = \breve{\mathbf{b}}_i \tag{6.52}$$

$$\tilde{\mathbf{z}}_{i+1} = \breve{\mathbf{z}}_i \tag{6.53}$$

$$\tilde{\mathbf{g}}_{i+1} = \breve{\mathbf{g}}_i \tag{6.54}$$

$$\tilde{\mathbf{k}}_{i+1} = \breve{\mathbf{k}}_i \tag{6.55}$$

The agent then goes through the EM loop again as outlined above. Crucially, this time it uses $\breve{\mathbf{b}}_i$, $\breve{\mathbf{z}}_i$, $\breve{\mathbf{g}}_i$ and $\breve{\mathbf{k}}_i$ instead of $\tilde{\mathbf{b}}_i$, $\tilde{\mathbf{z}}_i$, $\tilde{\mathbf{g}}_i$ and $\tilde{\mathbf{k}}_i$ to estimate temporary cluster responsibilities $\widehat{r}_{i-a+1:i}$, taking advantage of the new information provided by the feedback.

Importantly, after feedback the EM loop includes Cluster pruning, so unnecessary clusters will be eliminated at this stage. For any pruned cluster n, N is updated:

$$N \longleftarrow N-1 \tag{6.56}$$

This is also true for $\breve{\mathbf{b}}_i$ and $\breve{\mathbf{g}}_i$, so that cluster n is eliminated (either from $\breve{\mathbf{b}}_i$ or $\breve{\mathbf{g}}_i$ depending on its label) and, for any element $\breve{b}_{w,i}$ and $\breve{g}_{f,i}$

$$\breve{b}_{w,i} \longleftarrow \breve{b}_{w,i} - 1 \quad \text{if} \quad \breve{b}_{w,i} > n \tag{6.57}$$

$$\breve{g}_{f,i} \longleftarrow \breve{g}_{f,i} - 1 \quad \text{if} \quad \breve{g}_{f,i} > n \tag{6.58}$$

This is done for index consistency.

After convergence, the agent sets $\breve{\mathbf{r}}_{i-a+1} = \widehat{\mathbf{r}}_{i-a+1}$, and performs the final M step to evaluate the the posteriors' parameters $\breve{\boldsymbol{\theta}}_i$, after which it finally updates its priors by setting $\tilde{\boldsymbol{\theta}}_{i+1}\breve{\boldsymbol{\theta}}_i$, as outline above.

If $i < a$, the agent does not update its model parameters, waiting for further information. Conversely, if $i = I$, the agent updates its priors with the parameters estimates $\hat{\boldsymbol{\theta}}_I$, as there are no more stimuli to learn from. This would not occur in a naturalistic setting (no mushroom is guaranteed to be your last) but it was necessary here for proper model comparison.

6.3.2.4 Parameter settings

In this simulation, we set

$$\alpha_0 = 0.5$$

$$\beta_0 = 1$$

$$\mathbf{W}_0 = 0.002\mathbf{I}$$

$$v_0 = D + 2$$

where D is the number of features of the stimuli (in our case $D = 1$) and $\mathbf{I}$ is a $D \times D$ identity matrix. This means that every cluster is formed with an initial precision of 0.002 (equivalent to a variance of 2000).

The only parameter we manipulated was a, which ranged between 1 and 7.

6.3.2.5 Metrics

As a measure of performance, we used

$$P = \sum_{i=1}^{I} log\gamma_i \tag{6.59}$$

where

$$\gamma_i = \begin{cases} p(\zeta_i = 0) & \text{if} \quad \zeta_i = 0 \\ p(\zeta_i = 1) & \text{if} \quad \zeta_i = 1 \end{cases} \tag{6.60}$$

Therefore, the more the agent was likely to get the right answers throughout the simulation, the better its performance.

We also took a measure of sensory surprise

$$\begin{aligned} s_i &= -lnp\left(\mathbf{y}_i \mid \widehat{\boldsymbol{\theta}}_{i-1}\right) \\ &= -ln\sum_{n=1}^{N}\left\{\frac{\widehat{\alpha}_{i-1,n}}{\sum_{j=1}^{N}\widehat{\alpha}_{i-1,j}}\mathcal{N}\left(\mathbf{y}_i \mid \widehat{\boldsymbol{\mu}}_{i-1,n}, \left((\widehat{v}_{i-1}-D-1)\widehat{\mathbf{W}}_{i-1,n}\right)^{-1}\right)\right\} \end{aligned} \tag{6.61}$$

at the very beginning of each trial, after stimulus presentation and before cluster formation. This is a measure of how surprising a stimulus is before any update takes place, and it can be seen as an inverse index of how well the generative model predicts new data. Note that here we use $\widehat{\boldsymbol{\theta}}_{i-1}$ rather than $\tilde{\boldsymbol{\theta}}_i$, so that the agent takes into account trials $i-a+1:i-1$ if $a>1$. Here we use the mode of each distribution to get a "best guess" of cluster parameters $\boldsymbol{\pi}$, $\mathbf{m}$ and $\boldsymbol{\Lambda}$. We did not take this measure at $i=1$, as the model had zero components at that stage. This measure is somewhat equivalent to the Recognition metric in (Davis et al., 2012), where the authors simply used the log-likelihood (instead of the negative log-likelihood) to measure how much a stimulus was expected given the model parameters.

As in Davis et al. (2012) we calculated the entropy of the decision (as an inverse measure of confidence) pre-feedback and cluster assignment post-feedback.

$$\mathcal{H}_{pre} = -p(\zeta_i=0)lnp(\zeta_i=0) - p(\zeta_i=1)lnp(\zeta_i=1) \tag{6.62}$$

$$\mathcal{H}_{post} = -\sum_{n=1}^{N}\widehat{r}_{i,n}ln\left(\widehat{r}_{i,n}\right) \tag{6.63}$$

Finally, we calculated the Kullback-Leibler divergence (KL) between prior and posterior distribution at the end of each trial

$$KL = \int p(\mathbf{y}_i \mid \tilde{\boldsymbol{\theta}}_i) ln\frac{p(\mathbf{y}_i \mid \tilde{\boldsymbol{\theta}}_i)}{p(\mathbf{y}_i \mid \breve{\boldsymbol{\theta}}_i)}d\mathbf{y}_t \tag{6.64}$$

to have a measure of how much the generative model was updated.

6.4 Results

6.4.1 Parametric vs non-parametric models

We first compared performance P between our non-parametric filtering model ($a=1$) with a parametric version of it, with random prior location of clusters and symmetrical priors about cluster assignment ($\alpha_n = \alpha_0 = 0.5 \quad \forall n \in N$).

As expected, the non-parametric model ($\mu = -23.26$, $\sigma^2 = 9.52$) performed significantly better than a parametric one with one cluster for good mushrooms and one for bad ones ($\mu = -29.85$, $\sigma^2 = 7.73$, $t(1998) = 50.14$, $p < .001$). Interestingly, the non-parametric model still outperformed the parametric one when the correct number of clusters was set (2 for good mushrooms and 2 for bad ones, $\mu = -27.60$, $\sigma^2 = 4.79$, $t(1998) = 36.31$, $p < .001$). Fig. 6.6 summarises these findings, with the orange bar representing the performance of the non-parametric model and green bars that of parametric ones.

6.4.2 Retrospective inference

We carried out a one-way ANOVA with working memory capacity as a 7-level factor (i.e. with value of a ranging from 1 to 7), which revealed a significant main effect ($F(6, 6993) = 40.49$, $p < 0.01$). We then further investigated this effect with post-hoc comparisons (Tukey test), which are summarised in Fig. 6.7. This bar plot represents the average performance (as indexed by P) of the 7 models, displaying a clear trend of performance improving with memory capacity, and therefore with amount of computational resources invested in the task, in take with our initial hypothesis. We also observe that the biggest performance improvement occurs in the switch from a cognitive window of length 1 to one of length 2, and that this improvement diminishes progressively with all subsequent increases, to the point of becoming statistically insignificant (at least in our simulation).

Interestingly, an unexpected pattern (see Fig. 6.8) emerged with regards to the number of clusters found by the agent. With $a = 1$, the agent tends to estimate the correct number of clusters, with increasing overfitting (i.e. finding more clusters than necessary) peaking at $a = 3$. As a further increases, a slow decrease in estimated number of clusters is observed. Furthermore, we can observe an effect of a not only on the average number of estimated clusters, but in the variance too. In fact, Fig. 6.8 shows how changes in the mean of the number of estimated clusters are accompanied by changes in its variance (with error bars representing standard deviations to highlight this effect), which starts low, sharply increases peaking at $a = 3$ and then gently decreases (just like the mean). We did not perform statistical tests on this, as it was not the object of our work and a proper investigation would have required further manipulations to verify its consistency with varying experimental conditions. Nevertheless, we believe it is worth mentioning and speculating about (which we briefly do in the Discussion section).

KL-s correlation

a	1	2	3	4	5	6	7
Spearman's r	0.62	0.19	0.18	0.15	0.15	0.13	0.11

Table 6.2 Correlations between the KL divergence (indexing belief update) and sensory surprise for cognitive window sizes from 1 to 7.

6.4.3 Surprise and update metrics

We calculated the average correlation between s and KL across the different simulations (Table 6.2). Despite being fairly strong in the filtering model ($r = 0.62$), the r value decreases sharply as retrospective inference is introduced. This means that initial sensory surprise and magnitude of belief update are differently associated with each other based on the presence of a working memory component.

6.5 Discussion

In this work we built a non-parametric clustering model incorporating a working memory component, and tested it against classical parametric and fully online approaches in a simulated task. Our results highlight the importance of structure learning, with algorithms with the ability to flexibly update the structure of their generative model clearly outperforming parametric models that simply updated cluster parameters (i.e. that did not perform structure learning). Strikingly, in this task our non-parametric agent outperformed a parametric one even when the latter's generative model had the correct structure and random priors about cluster means, showing how a completely agnostic model is better than one with correct structure and random (and thus often incorrect) cluster centroids initialisation.

We further probed the advantages of FRI by adding a working memory component to our algorithm based on the work described in FitzGerald et al. (2020). Other models of decision-making involving a working memory component have been put out in the past (Collins and Frank, 2012; Viejo et al., 2015), but the deterministic associations between stimuli, actions and rewards made it unnecessary to perform retrospective inference on stored items. They did however include progressive memory decay (as opposed to a fixed cognitive window as the one we used), which could be an interesting addition to our model in future work.

FRI proved to significantly improve performance, showing how updating one's beliefs about past events is instrumental for developing a better understanding of current events, and making more adaptive choices as a consequence. Our results show that the bigger the cognitive window (i.e. working memory capacity), the better the performance. From a psychological perspective, cognitive window size can be thought of as amount of cognitive resources invested in the task, as keeping in working memory and updating beliefs about several (past) stimuli is clearly more cognitive demanding than doing the same with just the present one. In an experimental context this could be indirectly manipulated with reward size, with higher potential rewards motivating human participants to invest more cognitive resource on the task and thus perform retrospective inference on a bigger cognitive window. Alternatively, manipulating the distributions complexity and amount of overlap would change the difficulty of the task, and thus the amount of cognitive resources necessary to achieve a satisfactory performance. Using retrospective inference models to study cognitive resources allocation would add to the broader field of computational rationality (Gershman et al., 2015a), which deals with the trade-off between rewards and cognitive costs.

Applying these models to empirical studies would also allow to test the somewhat counterintuitive prediction of our model about the number of identified clusters. In fact, our results show that intermediate cognitive window sizes on average lead to the formation of more complex models of the environment, with the average number of identified clusters peaking at $a = 3$ and then slowly decreasing as working memory capacity increases. This can be explained by old stimuli in working memory being sometimes (partially) assigned to unnecessary newly formed clusters, which would as a consequence have a slightly higher chance of surviving pruning. In this case a big cognitive window would give the algorithm more time to correct this "mistake", as the unnecessary cluster's mixing component π_n would keep decreasing as all other $\alpha_{j \neq n}$ increase with time, bringing $r_{l,n} < max(\mathbf{r}_l)\ \forall l \in \{i-a+1, ..., i\}$, which would in turn cause the cluster to be pruned. Encouraging different levels of cognitive resources investment through experimental manipulations could allow to test our model's somewhat odd prediction, and provide an insight on how and when generative model augmentation happens in the brain.

In addition to this, our model provides update metrics that can be used for neuroimaging investigations. In fact, it allows to calculate separate quantities for sensory surprise and belief update (i.e. KL divergence between priors and posteriors), which in our simulation were only partially correlated. In addition, in our task they occur at different times (the former at stimulus presentation and the latter after feedback)

and could thus be associated with dissociable EEG or fMRI responses. Our results also show that the correlation between the two becomes much weaker as retrospective inference is introduced. This is unsurprising, as in this case sensory surprise would be in response to the current stimulus $\mathbf{y}_i$ and update would be performed using the oldest stimulus in the cognitive window $\mathbf{y}_{i-a+1}$. Therefore it should be possible to separately locate brain signal associated with these two metrics, investigating the relationship between information-theoretic surprise (i.e. negative log probability) and "Bayesian" surprise (i.e. KL divergence between priors and posteriors, see Baldi and Itti (2010); Nour et al. (2018)).

In addition our entropy measures ($\mathcal{H}_{pre}$ and $\mathcal{H}_{post}$) can be instrumental for investigating decision making under uncertainty (Davis et al., 2012).

In conclusion, we believe our work, in addition to being informative in itself, can provide a solid theoretical foundation for empirical investigations on structure learning in human participants, which we plan to carry out in the future.

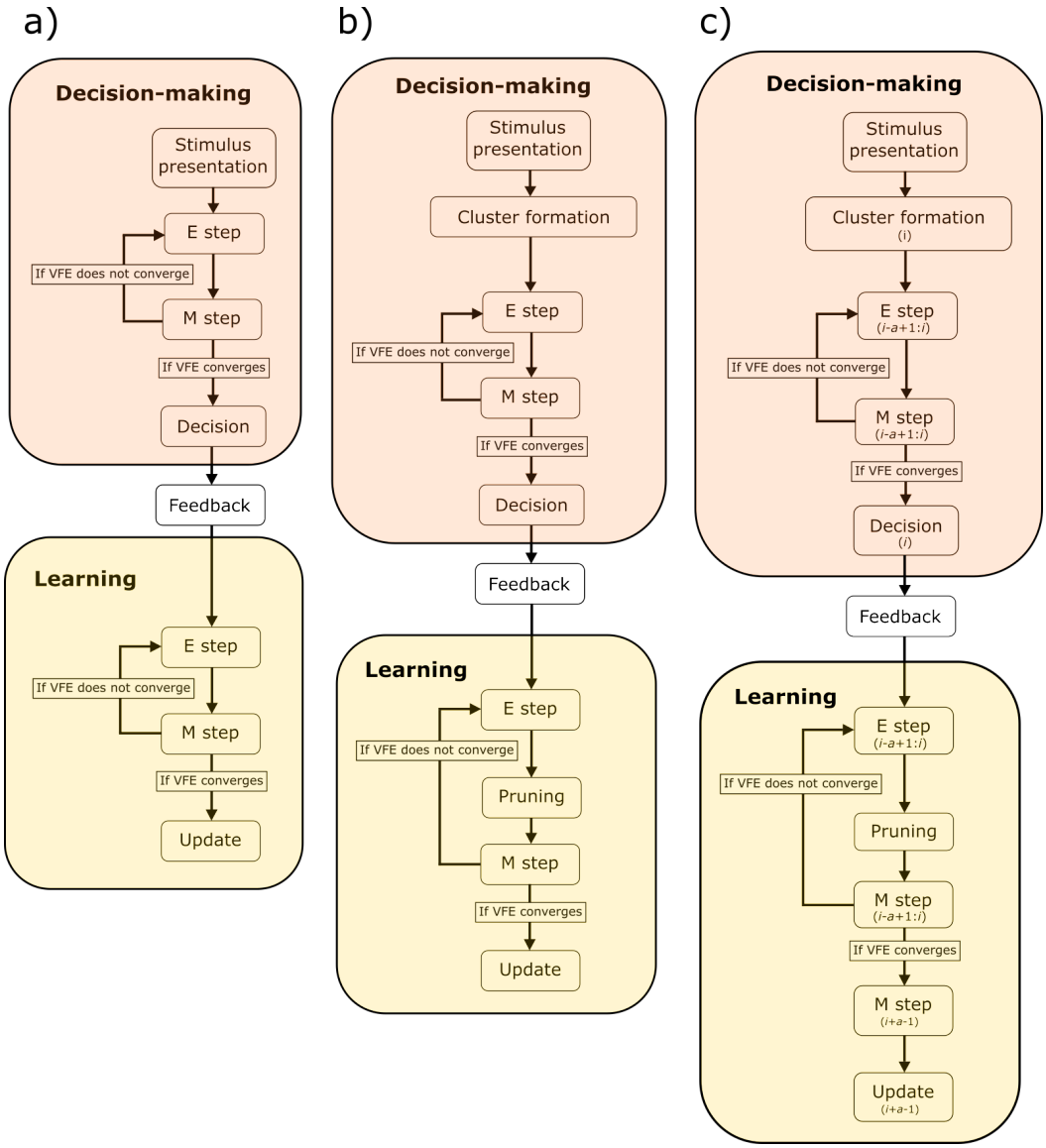

Fig. 6.5 Graphical representation of the algorithm pipelines. In the parametric filtering model (a) the agent evaluates the probabilities of the current stimulus being good or bad by inverting a GMM with fixed number of clusters, recursively evaluating responsibilities (E step) and cluster parameters (M step). After it receives feedback, it carries out the EM loop again to update cluster parameters in light of the new information. The non-parametric filtering model (b) is similar, but it involves two additional steps: cluster formation at the beginning of the trial, to take into account the possibility of having just encountered a new species of mushroom; and a pruning function embedded in the post-feedback EM loop, eliminating unnecessary clusters and keeping the generative model as simple as it can be. Finally, the non-parametric fixed-lag smoothing model (c) carries out all the steps described in (b), but it evaluates several stimuli at the same time, based on all trials in working memory. After VFE convergence in the post-feedback EM loop, it discards all the inferred parameters except for the cluster responsibilities of the oldest element in its cognitive window, which are used for a last M step (i.e. cluster parameters estimation) before belief update.

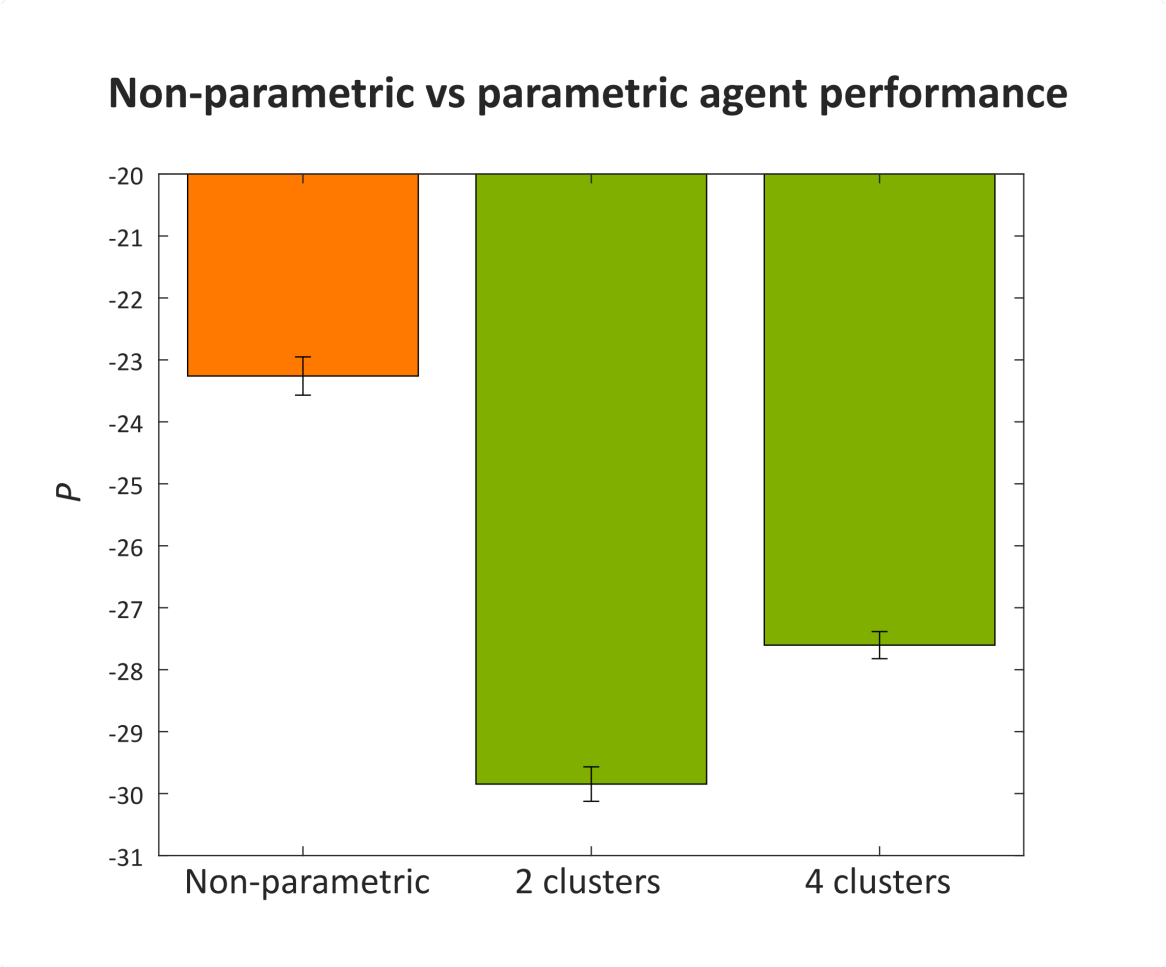

Fig. 6.6 Performance comparison between the non-parametric model and parametric versions with 2 (1 good and 1 bad) and 4 (2 good and 2 bad) clusters. Error bars represent standard errors.

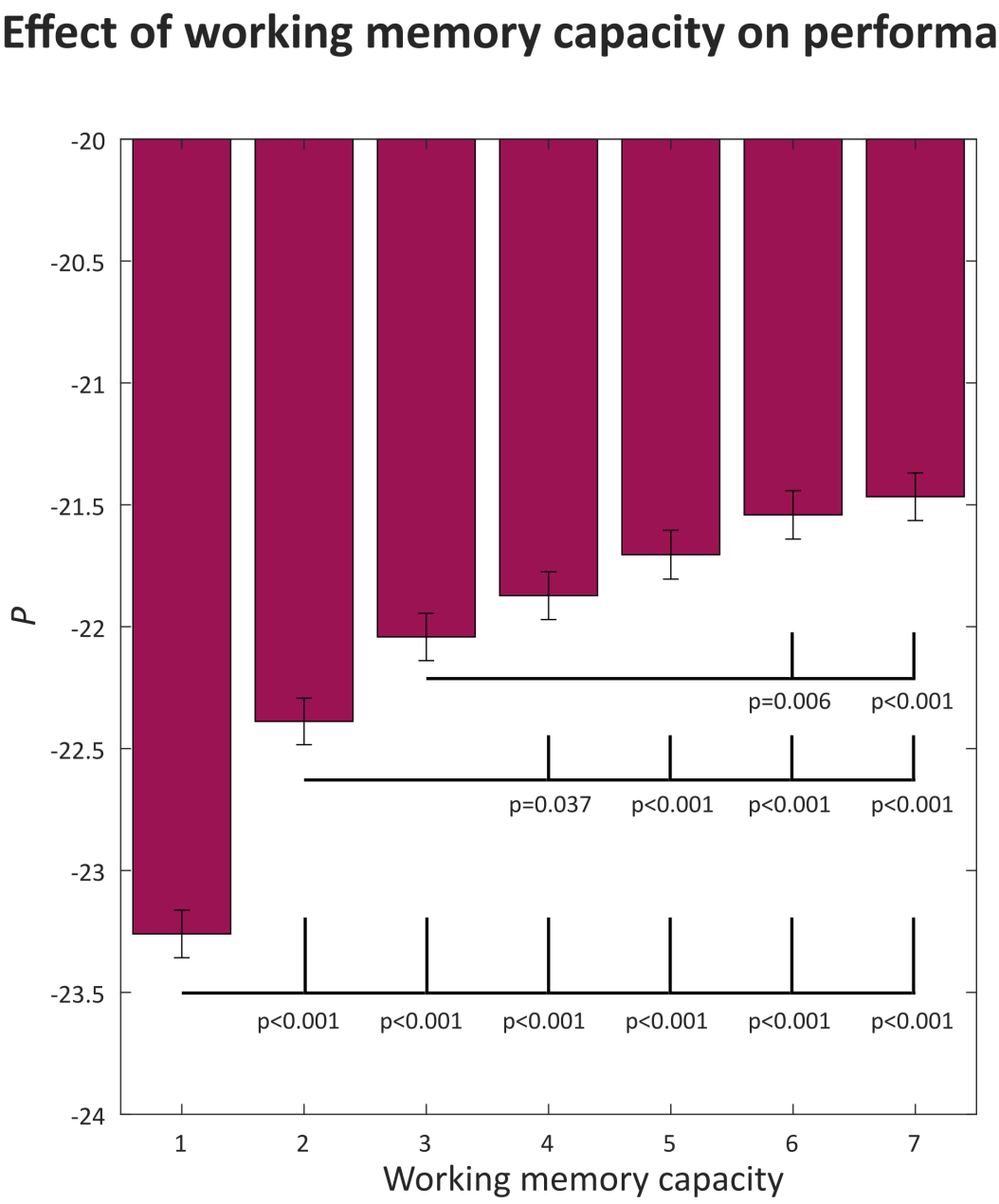

Fig. 6.7 Performance of the retrospective inference non-parametric model with different working memory capacities (1 to 7). Error bars represent standard errors.

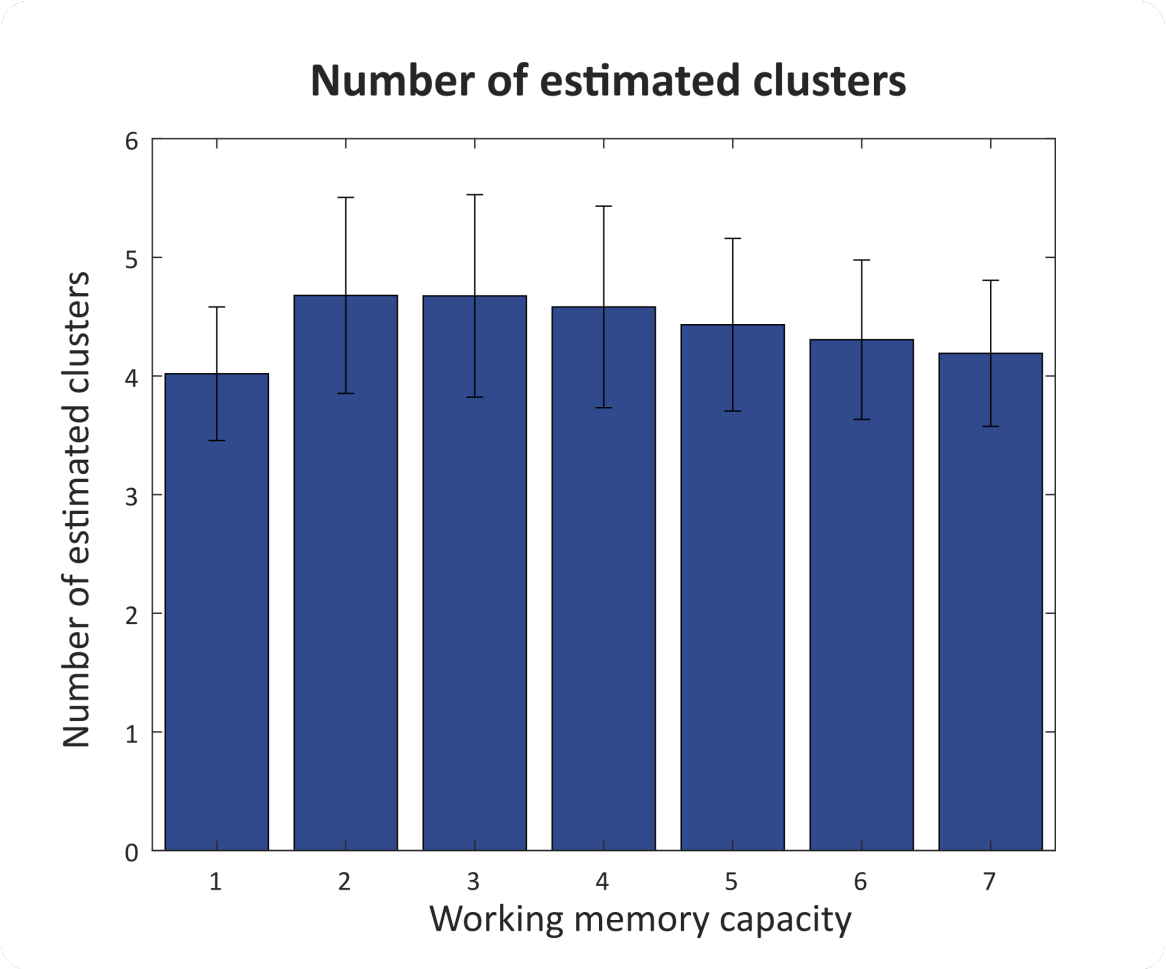

Fig. 6.8 Number of clusters estimated as a function of working memory capacity. Note that here error bars represent standard deviations, not standard errors. This was done to highlight the increase of variability, not only average, of estimated number of clusters for models with $a \sim 3$.

Chapter 7

General Discussion

In Chapters 4, 5 and 6 we presented three separate pieces of work, two of which are experiments with human participants and one is simulation-based. In all three we strongly drew from the Bayesian brain hypothesis (Knill and Pouget, 2004), framing inference and learning as machine learning problems and studying them with machine learning techniques (i.e. variational inference, clustering, Bayesian non-parametric methods). We considered situations mostly neglected by the literature, namely precision estimation and multimodal distributions, and showed how one of the most popular probabilistic accounts of brain function (i.e. predictive coding, Friston and Kiebel (2009)) fails to account for these. We then developed and described models that overcome these limitations.

In this Chapter we briefly review the findings presented in Chapters 4, 5 and 6, discussing their implications for the broader field. We also analyse the limitations of our studies, and suggest how could these could be overcome in future work. We finally discuss how our work can be used as a starting point for future research.

7.1 Summary of findings

7.1.1 Chapter 4

In Chapter 4 we tackled the problem of dynamic precision estimation. As discussed in Chapter 1, precision (i.e. inverse variance) is a fundamental quantity in predictive coding, as well as any probabilistic account of cognition involving Gaussian distributions. We built a predictive coding model that could accommodate trial-by-trial estimation of precision using a Gaussian prior over the log precision k, thus allowing to use the volatility parameter $\eta^{(k)}$ (the precision of a Gaussian random walk) to account for

possible objective variations in precision over time. Our model was thus able to account for situations in which the precision was non-stationary.

In fact, our experiment (a variant of the classic auditory oddball task, as described in Garrido et al. (2013)) involved presenting participants with a series of auditory stimuli whose pitch was sampled from a Gaussian distribution with non-stationary precision. We recorded our participant's pupil diameter while they were performing the task, using pupil dilation as an index of surprise to track learning (see Chapter 2). Our model-free analysis revealed that pupil responses to the same deviant stimulus differed according to the precision of the probability distribution it came from (deviant tones elicited greater surprise if the distribution was narrow), replicating the findings of Garrido et al. (2013). These results suggest participants kept track of that distribution's precision. Furthermore, we fitted pupillometry data to a range of competing models, and found that those including dynamic precision estimation vastly outperformed the others. There was little difference between the model which included both (log) precision and mean estimation and the one that included (log) precision estimation only, but this is easily explained by the fact that the mean of the distributions was kept constant, and thus mean estimation was largely unnecessary for the task (as one would likely correctly estimate it in a few trials). Therefore, our findings clearly show that human participants (even if this had no relevance to the task they were asked to perform) dynamically estimated precision over time.

Furthermore, our model-based analysis allowed to fit individual pupil response functions to pupil data, capture individual differences in pupil response. The combination of an individually-tailored convolutional kernel and the auto-regressive component resulted in our model gaining considerable explanatory power. Furthermore, this approach allowed to directly fit pupil data to cognitive models, contrary to the more common approaches of fitting them to behavioural data instead (Daw et al., 2011; O'Doherty et al., 2007; Schwartenbeck et al., 2015; Smittenaar et al., 2013), and use parameters estimates as regressors for some other physiological data analysis (Collins and Frank, 2016; Diaconescu et al., 2017).

7.1.2 Chapter 5

In Chapter 5 we tested a somewhat counter-intuitive prediction of predictive coding, namely that all sensory observations are assigned a unimodal Gaussian prior (i.e. variables are encoded as Gaussian distributions). We thus performed two experiments with identical task and stimuli, one with pupillometry and one with EEG. These were very similar to that described in Chapter 4, except for the fact that we kept

the probability distribution from which the stimuli were sampled constant. Crucially, this distribution was bimodal (a mixture of two Gaussians), meaning a prediction error-minimising agent (i.e. a predictive coding agent) would fail to identify the two modes and misrepresent the distribution as a unimodal Gaussian (for a mathematical demonstration of this see Chapter 1).

As in Chapter 4 we introduced probe tones, allowing for more straightforward model-free analyses at the expense of slightly distorting the distribution. In the two experiments described in Chapter 5 this didn't pay off as much as it did in the one described in Chapter 4. The model free analysis of both EEG and pupillometry data did suggest participants represented the stimuli distribution as bimodal, but the effects were considerably weaker than those reported in Chapter 4. This was true for both experiments, and especially for the EEG one, where only three out of the four contrasts of interest resulted significant, revealing a mismatch negativity effect (MMN, see Chapter 2). Overall, the results of our model-free analyses confirmed our hypothesis that participants would learn bimodal distribution, violating classical predictive coding. However, they did not so as convincingly as we would have hoped.

Our model-based analyses, on the other hand, left considerably less doubts about the ability of participants to form multimodal priors. In fact, model comparison favoured a Gaussian mixture model with 4 components over classical predictive coding (here represented as a unimodal Gaussian).

7.1.3 Chapter 6

Finally, in Chapter 6 we discussed structure learning and presented a simulation experiment. We introduced a novel structure learning task and had a series of simulated agents perform it. We found that an agent performing structure learning (in this case clustering with a Chinese restaurant process prior, see Chapter 3) vastly outperformed all agents with a fixed generative model, even if that model had the right structure (but its parameters were initialised at random). We thus showed that, in our task, an agnostic agent who grows its internal model of the environment as required by the data learns better than one who is already given the number of clusters observations can come from (i.e. who already knows the structure of the distribution generating the data), but has (more often than not) wrong priors about its parameters. This is important, as it shows that the best solution is also the most parsimonious, since the structure learning agent starts simple and increases the complexity of its model only if necessary.

Furthermore, we investigated the effect of finite retrospective inference (i.e. performing inferences about past observations, see FitzGerald et al. (2020)) on structure learning. We presented an online learning model incorporating a working memory component in the form of a "sliding cognitive window" of fixed length. At each trial, the agent not only performed inference on the current stimulus, but revisited past events in its cognitive window, reinterpreting them (i.e. re-performing cluster assignment) in light of the new information it gathered. This led its performance to drastically improve, growing with working memory capacity (i.e. the size of the cognitive window).

Finally, we presented a few useful metrics that can be extracted from our model, which we believe will be useful for empirical investigations. Our task naturally separates in time sensory surprise (at stimulus presentation) and belief update formalised as KL divergence between prior and posterior (at feedback), which is convenient for neuroimaging investigations looking for the brain correlates of both. Furthermore, in retrospective inference models these two metrics are only weakly correlated, as surprise is caused by the current stimulus and update is performed with a past one. Furthermore, we found that this correlation decreases as the cognitive window grows in size. This was most likely due to the fact that beliefs actually started to be updated later on for large cognitive windows, and the first updates are naturally the biggest in terms of KL divergence. Therefore, a simple filtering model would have its biggest updates in the first few trials, while a retrospective inference model with memory capacity a would have its biggest updates from trial a. This entails a strong prediction for studies trying to identify brain signals associated with surprise and belief update, namely that as participants invest more cognitive resources in the task (i.e. keep in working memory more stimuli) the two will be less and less correlated. We further showed how two uncertainty metrics can be derived from our model, which take the form of entropies over discrete distributions in line with previous work (Davis et al., 2012). Our last and perhaps most surprising finding was that the number of clusters estimated by the model peaked for cognitive windows of length 3.

7.2 Implications for the broader field

7.2.1 Predictive coding and the Bayesian brain

In all the work presented in this thesis we approached cognition from a probabilistic perspective, in line with the Bayesian brain hypothesis (see Chapter 1). In particular, Chapters 4 and 5 are closely related to a more specific framework, predictive coding.

We highlighted some of the limitations of this framework, and used it as a starting point to develop augmented models addressing them.

In Chapter 4 we introduce a predictive coding model with dynamic precision estimation, showing how this feature improved the fit to pupil data. As discussed at various points throughout this thesis, precision is a key quantity in predictive coding, regulating the relative influence of different information streams (Crucianelli et al., 2019) and of priors against sensory evidence (Friston, 2008) during inference, as well as how quickly to update one's beliefs (Behrens et al., 2007; Mathys et al., 2011). Precision is also at the core of many theories of psychopathology (Adams et al., 2013; Fletcher and Frith, 2009; Lawson et al., 2014), with hyper-precise priors thought to play a role in autism (Van de Cruys et al., 2014), depression (Kube et al., 2020) and anxiety (Paulus et al., 2019). What lacks in these theories is a formal Bayesian account of how precision is estimated and updated. There are models that include precision estimation (Bogacz, 2017; Friston, 2005), but, as discussed in Chapter 2, these don't give precision a fully probabilistic treatment, estimating only its most likely value and not its full posterior distribution. This might seem unimportant, but uncertainty associated with parameter estimation (i.e. precision hyperparameters, which we denoted as τ in Chapters 4) is crucial to understand plasticity (Auksztulewicz and Friston, 2016; Van de Cruys et al., 2014), and thus learning. To have such quantity in a model one must include priors over model parameters, which, in a predictive coding context, should be Gaussian (see Chapter 1).

As the conjugate prior of the precision of a Gaussian is a Gamma distribution (Bishop, 2006), we had to introduce a further approximation in our model, placing a Gaussian prior over the log precision k. This achieved two things. First, it allowed for k to be non-stationary, as we could use a Gaussian random walk to capture the parameter's volatility. Second, it allowed to formalise precision estimation as a trade-off between prediction error minimisation at different hierarchical levels, a key feature of predictive coding (see Chapter 1). This might not seem evident from equation 4.31, but if we were to re-write it as

$$\begin{aligned}\frac{\partial q^*(k_i)}{\partial k_i} &= \frac{1}{2} - \frac{e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2}{2} - \tilde{\tau}_i^{(k)}(k_i - \tilde{\mu}_i^{(k)}) \\ &= -\frac{1}{2}\left(e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2 - 1\right) - \tilde{\tau}_i^{(k)}(k_i - \tilde{\mu}_i^{(k)})\end{aligned} \tag{7.1}$$

where the first term is the difference between actual $(e^{k_i}(y_i - \breve{\mu}_i^{(m)})^2)$ and expected (1) product between estimated precision and squared error (i.e. deviation from the

prior's mean). This product is expected to be because the variance of any distribution corresponds to the expected squared error, which in our case means

$$e^{-k_i} = E\left[(y_i - \breve{\mu}_i^{(m)})^2\right] \tag{7.2}$$

As the second term of equation 7.1 is itself a precision weighted prediction error, it is clear how bringing the value of this derivative to zero (i.e. finding the maximum of $q^*(k_i)$) involves the aforementioned trade-off between prediction error, with the first having a fixed weight of 0.5 and the second having a weight determined by the precision of the prior on k.

In general, the importance of precision can be extended to any probabilistic theory of cognition, but it's of particular relevance for predictive coding for its strong Gaussian assumptions (see Chapter 1).

Even more so than Chapter 4, Chapter 5 deals with a problem which is specific of predictive coding, namely its prediction about the inability of the brain to represent non-Gaussian distributions. We offered empirical evidence that this is not the case, highlighting the need to improve the flexibility of predictive coding models, which can be done by adopting a Gaussian mixture model. This has profound implications for neurobiology, as the simple prediction error minimisation models described in Chapter 2 (Bogacz, 2017; Friston, 2005) cannot account for our finding and would need to be somewhat modified.

Finally, one last aspect worth mentioning is that, with bimodal distributions, precision ceases to be a suitable measure of uncertainty. We suggest future studies investigating uncertainty should focus on distribution entropy instead, as it is a more flexible measure, applicable to any distribution (both continuous and discrete).

7.2.2 Structure learning

In Chapter 6 we did not present a lab experiment, but we provided an experimental framework which we believe to be promising for studying structure learning (and, specifically, clustering) in humans. We devised a novel structure learning task, requiring participants (be them human or artificial agents) to cluster a set of mushrooms into species. We designed the real distributions to punish underfitting (i.e. not creating enough clusters), which unsurprisingly let agents unable to grow their generative models to perform poorly. The task is particularly attractive because it solves a fundamental problem of online unsupervised learning, namely the need to get behavioural responses at every trial without providing feedback about cluster membership. We get around this

by requiring our participants to cluster within categories (i.e. good and bad mushrooms) and to make decisions based on category, and not cluster membership. We can thus obtain trial-by-trial responses that are informative of the cluster assignment of each observation without explicitly asking participants about what species of mushrooms they just saw. This resembles the task used by (Davis et al., 2012), but we make use of a probabilistic setting (i.e. our stimuli are samples from a probability distribution) with continuous stimulus features, providing a task more similar to a real-world scenario. As presented in Chapter 6, the stimuli have only one feature. This choice was made to maximise the ambiguity about species membership, as the reduced feature space forces more overlapping between distributions. However, one could easily deploy the task with different (and possibly multivariate) stimuli distributions.

We believe this groundwork will allow us (and other research groups) to grow the scientific literature about online investigations of structure learning, which is at the moment quite limited in size (although see Collins and Frank (2013, 2016)).

In addition to designing a task, we developed a set of online learning models to simulate an agent carrying it out. We did this firstly to illustrate the importance of structure learning for this type of task, which a biological agent is likely to encounter in a naturalistic environment (e.g. learning which plants or mushrooms are edible, which animals are dangerous, etc.). Structure learning agents (i.e. non-parametric models) significantly outperformed agents with a fixed-structure generative model (i.e. parametric models). As one would expect this was very evident when the fixed model had the wrong number of components (i.e. 2 instead of 4), but, strikingly, it was the case also when the fixed model had the right structure, but its parameters were initialised at random. In other words, an agent growing its generative model from scratch outperformed one that only had to tune its model parameters. The latter had random, but very relaxed priors, so that they would not have a great weight against sensory evidence during model update. Both models in the long run are bound to find a nearly optimal solution, tuning both structure (for non-parametric models) and parameters (for all models) to the stimuli distributions. We could thus see the non-parametric models outperforming a parametric one with the right structure as building a structure from scratch being more efficient (i.e. faster) than adjusting model parameters. This finding, however, could be specific to the particular distributions we used in our simulation. Further investigations are required to verify how well this generalises to different structure learning problems.

The second objective of our simulation was to investigate the effect of retrospective inference on structure learning. We developed non-parametric models incorporating a

working memory component, and verified that performance increased with working memory capacity. This is particularly relevant for the field of computational rationality (Gershman et al., 2015a), as retrospective inference does improve performance, but a computational cost, namely storing observations in working memory and processing several (as many as there are in working memory) stimuli at once. We observed that the biggest improvement in performance when increasing the cognitive window size a from 1 (simple filtering model) to 2, with improvements decreasing in magnitude as a further increased. A biological agent would therefore have to evaluate the optimal compromise between computational efficiency and accuracy, avoiding overloading its working memory for little returns in terms of performance. Intelligent use of cognitive resources in a clustering task has already been investigated in a simulation study (Dasgupta and Griffiths, 2021), but this was limited to the concentration parameter θ (which we called α in Chapter 3) regulating the agent's propensity to form new clusters. With the work described in Chapter 6 we add a layer of complexity to the problem, paving the way for future empirical investigations to study the relationship between θ and a.

Finally, somewhat unexpectedly, we found a curious pattern in the relationship between working memory capacity and number of estimated clusters, with the maximum number reached at $a = 3$. In our case, this caused the model to overfit at a 3, but if we were to decrease the value of θ me might very well have observed the increase of estimated clusters to bring this number closer to the correct one. It is not clear how this effect would change with different distributions and values of θ, so we do not discuss this effect further in this section (we briefly do in Chapter 6). It is nevertheless worth investigating this further in future simulation work.

7.3 Limitations and future directions

Our studies had of course some limitations, which we will discuss in this section. In the experiments described in Chapters 4 and 5 participants were presented with a series of auditory stimuli varying in pitch, and we aimed at investigating how participants learned their probability distributions. For the sake of more understandable analysis we slightly distorted the distributions with probe tones as in Garrido et al. (2013), preventing our model to fully capture these distorted distributions. This was not a significant issue for the experiment presented in Chapter 4, where the probe tone model-free analysis resulted in a clean replication of the results of Garrido et al. (2013). On the other end, the model-free results of the two experiments described in Chapter

5 are not nearly as clean, and this might very well be for the distortion caused by the probe tones, which effectively increased the probability of stimuli that would otherwise have been in a low probability area (Probes 2 and 4, see Figures 5.1 and 5.2). We still found evidence supporting our hypothesis, but it would be worth repeating the experiment excluding the probe tones and relying solely on a model-based analysis.

The work presented in Chapter 6 has several limitations as well. First, it is a simulation experiment, and therefore how much it can tell us about human cognition is debatable. Unfortunately the pandemic prevented us from investigating structure learning in human subjects, but we are planning to remedy this as soon as we have the chance. Furthermore, our simulation experiment is not sufficient to investigate in depth the two unexpected effects we observed, namely the overfitting for windows size 3 and the non-parametric models outperforming a parametric one with the right structure and random parameter initialisation. Exploring these would involve several simulations, manipulating the values of θ and a, as well as the stimuli distributions. This would require a simulation study in itself, so we did not explore it in Chapter 6, but it could be the object of future simulation work.

Further limitations concern the computational modelling. In fact, contrary to Chapter 4, in Chapter 5 and 6 we did not focus on keeping our models biologically plausible or to align them with specific theories of brain function. Furthermore, in Chapter 4 we used univariate distributions, which made the problem of precision estimation easier to address. This becomes harder when non-diagonal covariance matrices are involved, which is likely to happen with multivariate naturalistic stimuli. In Chapter 6, despite using univariate stimuli, we built our model to be able to deal with multi-dimensional observations. Following Bishop (2006) we placed a Gaussian-Wishart prior on the joint distribution over the mean vector $\mathbf{m}$ and the precision matrix (i.e. the inverse of the variance matrix) $\mathbf{\Lambda}$, which resulted in very complicated and inelegant update equations. This worked nicely for testing our hypotheses, but we cannot make claims about the specific computations the brain would undergo when performing such a task, as there is no straightforward way to picture a neural implementation of the algorithm. Unfortunately, where Gamma priors over precision can be approximated to Gaussian priors over log precision (see Chapter 4), there is no such an easy solution for covariances that we are aware of, even when applying the mean-field approximation and factorising the distributions over means $\mathbf{m}$ (which results in a Gaussian prior) and over the precision matrix $\mathbf{\Lambda}$ (which results in a Wishart prior). There are models of online covariance estimation (Bogacz, 2017), but, as mentioned above and in Chapter 2, these do not give a fully probabilistic treatment to means and covariances, estimating

only their *maximum a posteriori* value. A possible avenue to explore in this sense would be that of abandoning variational inference altogether and focusing on sampling approaches (Aitchison and Lengyel, 2016; Bishop, 2006; Gershman and Beck, 2017).

In past sections we have already discussed some future research that could be done based on our work. Specifically, based on our findings decribed in Chapter 5, we suggested that future neuroimaging studies aimed investigating neural markers of uncertainty should focus on entropy rather than precision (or, equivalently, variance). This distinction is irrelevant for Gaussian distributions, as entropy and variance are monotonically related, but becomes crucial when the stimuli distributions are radically non-Gaussian, as it is the case for the bimodal one we made use of. Furthermore, in Chapter 6 we presented a series of metrics that can be derived from our model and used as predictors in neuroimaging studies. As discussed, these would allow to look for brain correlates of sensory surprise and belief update separately, allowing to disentangle them.

We identified two further possible directions our research could take using the work presented in this thesis as a starting point. The first consists in applying our models and tasks (and developing new ones) to understand psychopathology. This would fall into the emerging field of *computational psychiatry* (Huys et al., 2016). As discussed in Chapter 1 and in previous sections, precision is at the centre of many probabilistic accounts of psychopathology (Kube et al., 2020; Lawson et al., 2014; Paulus et al., 2019; Van de Cruys et al., 2014), which makes the model we developed for Chapter 4 a valuable tool to investigate when and how individual suffering from different psychiatric conditions might perform aberrant precision estimation, causing them to form non-adaptive models of the environment.

Second, as anticipated above, the retrospective inference clustering model presented in Chapter 6 is particularly suitable for investigating rational cognitive resources allocation (i.e. *computational rationality*, Gershman et al. (2015a)). It would be interesting to probe how a and θ can adapt to task complexity and potential reward in human participants, investigating how they tune cognitive resources investment to the contingencies of the task. Furthermore, one could explore alternatives to finite retrospective inference (or fixed lag smoothing). We have already developed a variant of our clustering algorithm which, instead of performing retrospective inference on a sliding cognitive window of fixed length, can choose whether to retain a stimulus in working memory based on the uncertainty (formalised as entropy) associated with its cluster assignment. Going back to the mushroom example, if the algorithm is not very sure what species a certain mushroom belongs to, it will just retain it in working

memory until further evidence makes cluster assignment less ambiguous. On the other hand, if a certain mushroom clearly belongs to a certain species, the algorithm will avoid wasting cognitive resources to hold it in working memory and revisit it in the future. Such an algorithm would be considerably more parsimonious than the one described in Chapter 6, as it would perform retrospective inference only on ambiguous stimuli. We have not tested this systematically yet, but this augmented version of our algorithm should maximise the benefits of retrospective inference while keeping its costs as contained as possible. This would open yet another avenue to investigate how human participants manage their cognitive resources.

Finally, it is worth pointing out that, although we discussed computational psychiatry and computational rationality separately, but this does not need to be the case for empirical investigations. One could, for example, investigate rumination (Nolen-Hoeksema et al., 2008) in depressed individuals as a non-adaptive version of retrospective inference. Also, it has been observed that individuals with depression tend to rely on cognitive inexpensive cognitive strategies (Huys et al., 2012), often leading to over-generalisation (Huys et al., 2015), which can be seen as a failure in structure learning (not generating enough clusters, on more generally not forming a complex enough model). All this is worth exploring, and our models can be a powerful tool to do so.

7.4 Conclusions

The work discussed in this thesis tackled some important topics within the wider framework of the Bayesian brain hypothesis. Some studies (i.e. Chapters 4 and 5) were more grounded on a specific theoretical framework (predictive coding), and explored its limitations, providing augmented versions of its traditional models and testing them on empirical data. Chapter 6 was more theoretically agnostic, but it expanded the findings of Chapter 5 (i.e. humans can learn multimodal distributions), increasing the complexity of the computational models deployed (i.e. adopting a non-parametric approach and introducing retrospective inference) to study structure learning. Overall, we believe this work represents a solid contribution to the the specific fields of predictive coding and structure learning, as well as to the Bayesian brain framework as a whole.

References

Adams, R. A., Brown, H. R., and Friston, K. J. (2014). Bayesian inference, predictive coding and delusions. *AVANT. J. Philos. Int. Vanguard*, 5:51–88.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47.

Aitchison, L. and Lengyel, M. (2016). The hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, 12(12):e1005186.

Alamia, A., VanRullen, R., Pasqualotto, E., Mouraux, A., and Zenon, A. (2019). Pupil-linked arousal responds to unconscious surprisal. *Journal of Neuroscience*, 39(27):5369–5376.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8):2960–2966.

Allman, J., Miezin, F., and McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual review of neuroscience*, 8(1):407–430.

Auksztulewicz, R. and Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *cortex*, 80:125–140.

Austerweil, J. L. and Griffiths, T. L. (2008). Analyzing human feature learning as nonparametric bayesian inference. In *NIPS*, pages 97–104.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10.

Baldi, P. and Itti, L. (2010). Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

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

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom).

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–39.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–1221.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bodatsch, M., Ruhrmann, S., Wagner, M., Müller, R., Schultze-Lutter, F., Frommann, I., Brinkmeyer, J., Gaebel, W., Maier, W., Klosterkötter, J., et al. (2011). Prediction of psychosis by mismatch negativity. *Biological psychiatry*, 69(10):959–966.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211.

Boly, M., Garrido, M. I., Gosseries, O., Bruno, M.-A., Boveroux, P., Schnakers, C., Massimini, M., Litvak, V., Laureys, S., and Friston, K. (2011). Preserved feedforward but impaired top-down processes in the vegetative state. *Science*, 332(6031):858–862.

Brainard, D. H. and Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10(4):433–436.

Braun, D. A., Mehring, C., and Wolpert, D. M. (2010). Structure learning in action. *Behavioural brain research*, 206(2):157–165.

Browning, M., Behrens, T. E., Jocham, G., O'reilly, J. X., and Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature neuroscience*, 18(4):590–596.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., and Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143(4):1476.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., and Sanborn, A. (2020). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, 29(5):506–512.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Collins, A. G. and Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035.

Collins, A. G. and Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190.

Collins, A. G. E. and Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, 152:160–169.

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

Crucianelli, L., Paloyelis, Y., Ricciardi, L., Jenkinson, P. M., and Fotopoulou, A. (2019). Embodied precision: intranasal oxytocin modulates multisensory integration. *Journal of cognitive neuroscience*, 31(4):592–606.

Damsma, A. and van Rijn, H. (2017). Pupillary response indexes the metrical hierarchy of unattended rhythmic violations. *Brain and cognition*, 111:95–103.

Dasgupta, I. and Griffiths, T. L. (2021). Clustering and the efficient use of cognitive resources.

Dasgupta, I., Schulz, E., and Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive psychology*, 96:1–25.

Daunizeau, J. (2017). The variational laplace approach to approximate bayesian inference. *arXiv preprint arXiv:1703.02089*.

Davis, T., Love, B. C., and Preston, A. R. (2012). Learning the exception to the rule: Model-based fmri reveals specialized representations for surprising category members. *Cerebral Cortex*, 22(2):260–273.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Dayan, P. and Yu, A. J. (2006). Phasic norepinephrine: a neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4):335–350.

De Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., and Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature communications*, 7(1):1–11.

de Gee, J. W., Knapen, T., and Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, 111(5):E618–E625.

Denison, R. N., Parker, J. A., and Carrasco, M. (2020). Modeling pupil responses to rapid sequential events. *Behavior research methods*, 52(5):1991–2007.

Denison, R. N., Piazza, E. A., and Silver, M. A. (2011). Predictive context influences perceptual selection during binocular rivalry. *Frontiers in human neuroscience*, 5:166.

Diaconescu, A. O., Litvak, V., Mathys, C., Kasper, L., Friston, K. J., and Stephan, K. E. (2017). A computational hierarchy in human cortex. *arXiv preprint arXiv:1709.02323*.

Ebitz, R. B. and Moore, T. (2019). Both a gauge and a filter: Cognitive modulations of pupil size. *Frontiers in neurology*, 9:1190.

Erdogan, G. and Jacobs, R. A. (2017). Visual shape perception as bayesian inference of 3d object-centered shape representations. *Psychological review*, 124(6):740.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Feldman, H. and Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130.

FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S.-C., and Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS computational biology*, 13(5):e1005418.

FitzGerald, T. H., Hammerer, D., Sambrook, T. D., and Penny, W. D. (2019). Bayesian inference over model-spaces increases the accuracy of model comparison and allows formal testing of hypotheses about model distributions in experimental populations. *arXiv preprint arXiv:1901.01916*.

FitzGerald, T. H., Penny, W. D., Bonnici, H. M., and Adams, R. A. (2020). Retrospective inference as a form of bounded rationality, and its beneficial influence on learning. *Frontiers in artificial intelligence*, 3:2.

Fletcher, P. C. and Frith, C. D. (2009). Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58.

Ford, J. M. and Hillyard, S. A. (1981). Event-related potentials (erps) to interruptions of a steady rhythm. *Psychophysiology*, 18(3):322–330.

Franklin, N. T. and Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS computational biology*, 14(4):e1006116.

Friedman, D., Hakerem, G., Sutton, S., and Fleiss, J. L. (1973). Effect of stimulus uncertainty on the pupillary dilation response and the vertex evoked potential. *Electroencephalography and clinical neurophysiology*, 34(5):475–484.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Friston, K. (2008). Hierarchical models in the brain. *PLoS computational biology*, 4(11):e1000211.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, 29(1):1–49.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–234.

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., and Tenenbaum, J. (2021). World model learning and inference. *Neural Networks*, 144:573–590.

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS one*, 4(7):e6421.

Friston, K. J., Stephan, K. E., Montague, R., and Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2):148–158.

Garrido, M. I., Friston, K. J., Kiebel, S. J., Stephan, K. E., Baldeweg, T., and Kilner, J. M. (2008). The functional anatomy of the mmn: a dcm study of the roving paradigm. *Neuroimage*, 42(2):936–944.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., and Friston, K. J. (2009a). Dynamic causal modeling of the response to frequency deviants. *Journal of Neurophysiology*, 101(5):2620–2631.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009b). The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*, 120(3):453–463.

Garrido, M. I., Sahani, M., and Dolan, R. J. (2013). Outlier responses reflect sensitivity to statistical structure in the human brain. *PLoS Computational Biology*, 9(3):e1002999.

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, 24(2):557–565.

Gershman, S. J. and Beck, J. M. (2017). Complex probabilistic inference. *Computational models of brain and behavior*, 453.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Gershman, S. J., Blei, D. M., and Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, 117(1):197.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015a). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Gershman, S. J. and Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & behavior*, 40(3):255–268.

Gershman, S. J., Norman, K. A., and Niv, Y. (2015b). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50.

Gershman, S. J., Pouncy, H. T., and Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, 41:545–575.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural computation*, 24(1):1–24.

Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384.

Hämmerer, D., Schwartenbeck, P., Gallagher, M., FitzGerald, T. H. B., Düzel, E., and Dolan, R. J. (2019). Older adults fail to form stable task representations during model-based reversal inference. *Neurobiology of aging*, 74:90–100.

Hoeks, B. and Levelt, W. J. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, 25(1):16–26.

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3):687–701.

Hong, L., Walz, J. M., and Sajda, P. (2014). Your eyes give you away: prestimulus changes in pupil diameter correlate with poststimulus task-related eeg dynamics. *PLoS One*, 9(3):e91321.

Hou, R., Freeman, C., Langley, R., Szabadi, E., and Bradshaw, C. (2005). Does modafinil activate the locus coeruleus in man? comparison of modafinil and clonidine on arousal and autonomic functions in human volunteers. *Psychopharmacology*, 181(3):537–549.

Hovsepyan, S., Olasagasti, I., and Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature communications*, 11(1):1–12.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3):e1002410.

Huys, Q. J., Guitart-Masip, M., Dolan, R. J., and Dayan, P. (2015). Decision-theoretic psychiatry. *Clinical Psychological Science*, 3(3):400–421.

Huys, Q. J., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, 19(3):404–413.

Jääskeläinen, I. P., Ahveninen, J., Bonmassar, G., Dale, A. M., Ilmoniemi, R. J., Levänen, S., Lin, F.-H., May, P., Melcher, J., Stufflebeam, S., et al. (2004). Human posterior auditory cortex gates novel sounds to consciousness. *Proceedings of the National Academy of Sciences*, 101(17):6809–6814.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision research*, 39(21):3621–3629.

Javitt, D. C., Grochowski, S., Shelley, A.-M., and Ritter, W. (1998). Impaired mismatch negativity (mmn) generation in schizophrenia as a function of stimulus deviance, probability, and interstimulus/interdeviant interval. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 108(2):143–153.

Kang, O. E., Huffer, K. E., and Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PloS one*, 9(8):e102463.

Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, 34(7):1185–1243.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, 10(3):307–321.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692.

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology*, 4(11):e1000209.

Kim, C. S. (2021). Bayesian mechanics of perceptual inference and motor control in the brain. *Biological Cybernetics*, 115(1):87–102.

Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2):B35–B42.

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A., Bonneh, Y. S., and Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience*, 41(8):1068–1078.

Knapen, T., de Gee, J. W., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., and Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equiluminance. *PloS one*, 11(5):e0155574.

Knill, D. C. (1998). Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision research*, 38(11):1683–1711.

Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Koelsch, S., Busch, T., Jentschke, S., and Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical mmn reflects the magnitude of transitional probabilities in auditory sequences. *Scientific reports*, 6(1):1–11.

Kok, P., Jehee, J. F., and De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

Korn, C. W. and Bach, D. R. (2016). A solid frame for the window on cognition: Modeling event-related pupil responses. *Journal of vision*, 16(3):28–28.

Koster-Hale, J. and Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5):836–848.

Krystal, J. H., Murray, J. D., Chekroud, A. M., Corlett, P. R., Yang, G., Wang, X.-J., and Anticevic, A. (2017). Computational psychiatry and the challenge of schizophrenia.

Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A., and Rief, W. (2020). Distorted cognitive processes in major depression: a predictive processing perspective. *Biological psychiatry*, 87(5):388–398.

Kuchinke, L., Võ, M. L.-H., Hofmann, M., and Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, 65(2):132–140.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Laughlin, S. B. (1992). Retinal information capacity and the function of the pupil. *Ophthalmic and Physiological Optics*, 12(2):161–164.

Lavín, C., San Martín, R., and Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in behavioral neuroscience*, 7:218.

Lawson, R. P., Rees, G., and Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in human neuroscience*, 8:302.

Lee, M., Sehatpour, P., Hoptman, M. J., Lakatos, P., Dias, E. C., Kantrowitz, J. T., Martinez, A. M., and Javitt, D. C. (2017). Neural mechanisms of mismatch negativity dysfunction in schizophrenia. *Molecular psychiatry*, 22(11):1585–1593.

Liao, H.-I., Yoneya, M., Kidani, S., Kashino, M., and Furukawa, S. (2016). Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention. *Frontiers in Neuroscience*, 10:43.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1993). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In *Neurophysiology of consciousness*, pages 249–268. Springer.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS computational biology*, 9(2):e1002911.

Lucas, C. G. and Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, 34(1):113–147.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., and Adam, R. (2016). The curious incident of attention in multisensory integration: bottom-up vs. top-down. *Multisensory Research*, 29(6-7):557–583.

Mackay, D. J. C. (1998). Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

Mark, S., Moran, R., Parr, T., Kennerley, S. W., and Behrens, T. E. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature communications*, 11(1):1–12.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., and Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, 28(47):12539–12545.

Mathôt, S. (2018). Pupillometry: psychology, physiology, and function. *Journal of Cognition*, 1(1).

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, 5:39.

Menghi, N., Kacar, K., and Penny, W. (2021). Multitask learning over shared subspaces. *PLoS computational biology*, 17(7):e1009092.

Minka, T. P. (2013). Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*.

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, 16(1):72–80.

Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., and Friston, K. J. (2013). Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience*, 33(19):8227–8236.

Murphy, P. R., O'connell, R. G., O'sullivan, M., Robertson, I. H., and Balsters, J. H. (2014). Pupil diameter covaries with bold activity in human locus coeruleus. *Human brain mapping*, 35(8):4140–4154.

Murphy, P. R., Robertson, I. H., Balsters, J. H., and O'connell, R. G. (2011). Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, 48(11):1532–1543.

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (mmn): towards the optimal paradigm. *Clinical neurophysiology*, 115(1):140–144.

Näätänen, R. and Picton, T. (1987). The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., and Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, 15(7):1040–1046.

Nolen-Hoeksema, S., Wisco, B. E., and Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on psychological science*, 3(5):400–424.

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H., Coello, C., Wall, M. B., Dolan, R. J., and Howes, O. D. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proceedings of the National Academy of Sciences*, 115(43):E10167–E10176.

Näätänen, R., Paavilainen, P., and Reinikainen, K. (1989). Do event-related potentials to infrequent decrements in duration of auditory stimuli demonstrate a memory trace in man? *Neuroscience Letters*, 107(1):347–352.

O'Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of sciences*, 1104(1):35–53.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011.

Orbán, G., Fiser, J., Aslin, R. N., and Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105(7):2745–2750.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., and Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 110(38):E3660–E3669.

Otten, M., Seth, A. K., and Pinto, Y. (2017). A social bayesian brain: How social knowledge can shape visual perception. *Brain and cognition*, 112:69–77.

Parise, C. V. and Ernst, M. O. (2017). Noise, multisensory integration, and previous response in perceptual disambiguation. *PLoS computational biology*, 13(7):e1005546.

Paulus, M. P., Feinstein, J. S., and Khalsa, S. S. (2019). An active inference approach to interoceptive psychopathology. *Annual review of clinical psychology*, 15:97–122.

Pause, B. M. and Krauel, K. (2000). Chemosensory event-related potentials (cserp) as a key to the psychology of odors. *International Journal of Psychophysiology*, 36(2):105–122.

Pazo-Alvarez, P., Cadaveira, F., and Amenedo, E. (2003). Mmn in the visual modality: a review. *Biological psychology*, 63(3):199–236.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian inference for fmri time series. *NeuroImage*, 19(3):727–741.

Penny, W. D. (2012). Comparing dynamic causal models using aic, bic and free energy. *Neuroimage*, 59(1):319–330.

Phillips, M., Szabadi, E., and Bradshaw, C. (2000). Comparison of the effects of clonidine and yohimbine on spontaneous pupillary fluctuations in healthy human volunteers. *Psychopharmacology*, 150(1):85–89.

Preuschoff, K., t Hart, B. M., and Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in neuroscience*, 5:115.

Qiyuan, J., Richer, F., Wagoner, B. L., and Beatty, J. (1985). The pupil and stimulus probability. *Psychophysiology*, 22(5):530–534.

Raisig, S., Welke, T., Hagendorf, H., and van der Meer, E. (2010). I spy with my little eye: Detection of temporal violations in event sequences and the pupillary response. *International Journal of Psychophysiology*, 76(1):1–8.

Rajkowski, J. (1993). Correlations between locus coeruleus (lc) neural activity, pupil diameter and behavior in monkey support a role of lc in attention. *Soc. Neurosc., Abstract, Washington, DC, 1993*.

Rajkowski, J., Kubiak, P., and Aston-Jones, G. (1994). Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Brain research bulletin*, 35(5-6):607–616.

Rajkowski, J., Majczynski, H., Clayton, E., and Aston-Jones, G. (2004). Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *Journal of neurophysiology*, 92(1):361–371.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., and Tolias, A. S. (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, 84(2):355–362.

Reinhard, G. and Lachnit, H. (2002). The effect of stimulus probability on pupillary response as an indicator of cognitive processing in human learning and categorization. *Biological Psychology*, 60(2-3):199–215.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, 84:971–985.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52.

Sanborn, A., Griffiths, T., and Navarro, D. (2006). A more rational model of categorization.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893.

Särkkä, S. (2013). *Bayesian filtering and smoothing*. Number 3. Cambridge University Press.

Schwartenbeck, P., FitzGerald, T. H., and Dolan, R. (2016). Neural signals encoding shifts in beliefs. *Neuroimage*, 125:578–586.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., and Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral cortex*, 25(10):3434–3445.

Schwartz, S., Shinn-Cunningham, B., and Tager-Flusberg, H. (2018). Meta-analysis and systematic review of the literature characterizing auditory mismatch negativity in individuals with autism. *Neuroscience & Biobehavioral Reviews*, 87:106–117.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573.

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2:395.

Shannon, C. (1948). Claude shannon. *Information Theory*, 3:224.

Shinozaki, N., Yabe, H., Sutoh, T., Hiruma, T., and Kaneko, S. (1998). Somatosensory automatic responses to deviant stimuli. *Cognitive Brain Research*, 7(2):165–171.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in psychology*, 7:1792.

Silvestrin, F., Penny, W. D., and FitzGerald, T. H. (2019). Pupil dilation indexes statistical learning about the uncertainty of stimulus distributions. In *Conference paper at Cognitive Computational Neuroscience (CCN) conference*.

Silvestrin, F., Penny, W. D., and FitzGerald, T. H. (2021). Pupil dilation indexes automatic and dynamic inference about the precision of stimulus distributions. *Journal of Mathematical Psychology*, 101:102503.

Silvetti, M., Seurinck, R., van Bochove, M., and Verguts, T. (2013). The influence of the noradrenergic system on optimal control of neural plasticity. *Frontiers in behavioral neuroscience*, 7:160.

Smith, R., Schwartenbeck, P., Parr, T., and Friston, K. J. (2020). An active inference approach to modeling structure learning: concept learning as an example case. *Frontiers in computational neuroscience*, 14:41.

Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D., and Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, 80(4):914–919.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area v1. *Journal of neuroscience*, 30(9):3531–3543.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and cognition*, 112:92–97.

Steinhauer, S. and Zubin, J. (1982). Vulnerability to schizophrenia: Information processing in the pupil and event-related potential. In *Biological markers in psychiatry and neurology*, pages 371–385. Elsevier.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017.

Steyvers, M., Griffiths, T. L., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in cognitive sciences*, 10(7):327–334.

Sussman, E. and Winkler, I. (2001). Dynamic sensory updating in the auditory system. *Cognitive Brain Research*, 12(3):431–439.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Structure learning in human causal induction. *Advances in neural information processing systems*, pages 59–65.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature neuroscience*, 7(9):907–915.

Tomov, M. S., Dorfman, H. M., and Gershman, S. J. (2018). Neural computations underlying causal structure learning. *Journal of Neuroscience*, 38(32):7143–7157.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., and Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, 30(33):11177–11187.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de Wit, L., and Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological review*, 121(4):649.

van der Wel, P. and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, 25(6):2005–2015.

van Pelt, S., Heil, L., Kwisthout, J., Ondobaka, S., van Rooij, I., and Bekkering, H. (2016). Beta-and gamma-band activity reflect predictive coding in the processing of causal events. *Social cognitive and affective neuroscience*, 11(6):973–980.

Viejo, G., Khamassi, M., Brovelli, A., and Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in behavioral neuroscience*, 9:225.

Vincent, P., Parr, T., Benrimoh, D., and Friston, K. J. (2019). With an eye on uncertainty: Modelling pupillary responses to environmental volatility. *PLoS computational biology*, 15(7):e1007126.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637.

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience*, 32(11):3665–3678.

Wang, L., Conner, J. M., Rickert, J., and Tuszynski, M. H. (2011). Structural plasticity within highly specific neuronal populations identifies a unique parcellation of motor learning in the adult brain. *Proceedings of the National Academy of Sciences*, 108(6):2545–2550.

Ward, E. J. (2008). A review and comparison of four commonly used bayesian and maximum likelihood model selection tools. *Ecological Modelling*, 211(1-2):1–10.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., and Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:345.

Weber, L. A., Diaconescu, A. O., Mathys, C., Schmidt, A., Kometer, M., Vollenweider, F., and Stephan, K. E. (2020). Ketamine affects prediction errors about statistical regularities: a computational single-trial analysis of the mismatch negativity. *Journal of Neuroscience*, 40(29):5658–5668.

Weilnhammer, V., Stuke, H., Hesselmann, G., Sterzer, P., and Schmack, K. (2017). A predictive coding account of bistable perception-a model-based fmri study. *PLoS computational biology*, 13(5):e1005536.

Wetzel, N., Buttelmann, D., Schieler, A., and Widmann, A. (2016). Infant and adult pupil dilation in response to unexpected sounds. *Developmental psychobiology*, 58(3):382–392.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.

Wierda, S. M., van Rijn, H., Taatgen, N. A., and Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences*, 109(22):8456–8460.

Winkler, I., Karmos, G., and Näätänen, R. (1996). Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. *Brain research*, 742(1-2):239–252.

Wu, S., Élteto, N., Dasgupta, I., and Schulz, E. (2021). Learning structure from the ground up: Hierarchical representation learning by chunking. In *Tenth International Conference on Learning Representations (ICLR 2022)*.

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308.

Zénon, A. (2017). Time-domain analysis for extracting fast-paced pupil responses. *Scientific reports*, 7(1):1–10.

Zénon, A. (2019). Eye pupil signals information gain. *Proceedings of the Royal Society B*, 286(1911):20191593.