frontiers in
HUMAN NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 02 December 2010
doi: 10.3389/fnhum.2010.00215

# Attention, uncertainty, and free-energy

***Harriet Feldman and Karl J. Friston****

*The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK*

***Edited by:***
*Sven Bestmann, University College London, UK*

***Reviewed by:***
*William Milberg, Harvard Medical School, USA*
*Tamer Demiralp, Istanbul University, Turkey*
*Laurence T. Maloney, New York University, USA*

****Correspondence:***
*Karl J. Friston, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, Queen Square, London WC1N 3BG, UK.*
*e-mail: k.friston@fil.ion.ucl.ac.uk*

We suggested recently that attention can be understood as inferring the level of uncertainty or precision during hierarchical perception. In this paper, we try to substantiate this claim using neuronal simulations of directed spatial attention and biased competition. These simulations assume that neuronal activity encodes a probabilistic representation of the world that optimizes free-energy in a Bayesian fashion. Because free-energy bounds surprise or the (negative) log-evidence for internal models of the world, this optimization can be regarded as evidence accumulation or (generalized) predictive coding. Crucially, both predictions about the state of the world generating sensory data and the precision of those data have to be optimized. Here, we show that if the precision depends on the states, one can explain many aspects of attention. We illustrate this in the context of the Posner paradigm, using the simulations to generate both psychophysical and electrophysiological responses. These simulated responses are consistent with attentional bias or gating, competition for attentional resources, attentional capture and associated speed-accuracy trade-offs. Furthermore, if we present both attended and non-attended stimuli simultaneously, biased competition for neuronal representation emerges as a principled and straightforward property of Bayes-optimal perception.

**Keywords: attention, biased competition, precision, free-energy, perception, generative models, predictive coding**

## INTRODUCTION

Attention is a ubiquitous and important construct in cognitive neuroscience. Many accounts of attention fall back on Jamesian formulations, famously articulated as "the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought" (James, 1890). More recent and formal accounts appeal to information theory and computational principles (Duncan and Humphreys, 1989; Deco and Rolls, 2005; Jaramillo and Pearlmutter, 2007; Spratling, 2008; Bruce and Tsotsos, 2009; Reynolds and Heeger, 2009; Spratling, 2010), with an increasing emphasis on Bayesian formulations (Rao, 2005; Chikkerur et al., 2010; Itti and Baldi, 2009). We pursue these attempts to understand attention in computational terms. This means we will be using terms like uncertainty, surprise and precision in a rather formal way. Without exception, these terms refer to properties of probability distributions. Probability distributions are central to modern treatments of perception that cast perception as inference. Inference requires us to represent probability distributions (or densities) over possible causes or explanations for our sensations. These distributions have several important attributes: for example, a broad distribution encodes a high degree of *uncertainty* about a particular cause. This uncertainty is, mathematically, the average (expected) *surprise* over all possibilities. A key measure of uncertainty is the width or variance of the distribution, or its inverse, *precision*. (see Glossary of Terms). In what follows, we hope to show that attention is more concerned with optimizing the uncertainty or precision of probabilistic representations, rather than what is being represented. By describing perception in formal terms, one can see almost intuitively where attention fits into the larger picture and how it might be mediated neurobiologically. This is important because a formal framework allows one to link classical psychological constructs to current physiological perspectives on attention (e.g., communication through coherence; Fries, 2005). We hope to show that the perspectives afforded by cognitive psychology, neurophysiology and formal (theoretical) treatments are all remarkably consistent.

We have suggested recently that perception is the inference about causes of sensory inputs and attention is the inference about the uncertainty (precision) of those causes (Friston, 2009). This places attention in the larger context of perceptual inference under uncertainty (Rao, 2005; Spratling, 2008; Whiteley and Sahani, 2008; Chikkerur et al., 2010). In this work, we try to show that attention emerges naturally in a Bayes-optimal scheme used previously to address predictive coding (Friston and Kiebel, 2009), perceptual categorization (Kiebel et al., 2009), learning (Friston, 2008) and action (Friston et al., 2010a). In other words, we try to explain some simple attentional phenomena using an established framework that has explanatory power in domains beyond attention. Specifically, we show how attention can be construed as inferring the precision of sensory signals and their causes. The idea is illustrated using computational simulations of neuronal processing that try to establish face validity in terms of psychophysical and electrophysiological responses. We do this in the context of the Posner paradigm (Posner, 1980); a classical paradigm for studying directed spatial attention in vision, using cued targets. This paradigm also allows us to address, in a heuristic way, biased competition (Desimone, 1996) by presenting validly and invalidly cued targets simultaneously. Our hope was to connect psychophysical studies of attention with theories based upon detailed electrophysiological studies in monkeys.

The basic idea we pursue is that attention entails estimating uncertainty during hierarchical inference about the causes of sensory input. We develop this idea in the context of perception based on Bayesian principles, under the free-energy principle (Friston, 2009). Formally, this scheme can be regarded as a generalization of

predictive coding (Rao and Ballard, 1998) and involves recurrent message passing among hierarchical levels of cortical systems to optimize a probabilistic representation of the world (Mumford, 1992; Friston, 2009). In these generalized schemes, precision is encoded by the synaptic gain (post-synaptic responsiveness) of units reporting prediction errors (Friston, 2008). There are many metaphors for attention that relate closely to the idea we are trying to describe. Perhaps the simplest is that of statistical inference, which treats perception as hypothesis testing (Gregory, 1980): indeed, most modern theories of perception draw on Helmholtz's ideas about the brain as an inference machine (e.g., Gregory, 1968; Ballard et al., 1983; Dayan et al., 1995). These theories regard the brain as inferring how sensory data are generated using generative models (cf, hypotheses) in exactly the same way that we analyze scientific data. The simplest example of this is the Student's *t*-statistic, where a difference in group means is divided by its standard error to test for group differences. Under the null hypothesis, the observed difference is the *prediction error* and the standard error is an estimate of its *precision* (inverse variance). This means that one can regard the *t*-statistic as a precision-weighted prediction error. Crucially, both the prediction error and its precision have to be estimated, given empirical (sensory) data. The idea here is that attention rests on estimating precision and is therefore an integral part of perception. Things get more interesting if we consider that the precision of sensory signals depend on states of the world. This means that optimizing precision entails optimizing inferred states of the world that affect the precision or uncertainty about our sensations. It is this generalization of generative models we exploit in this paper. In brief, most generative models (including those used to simulate perception) ignore state-dependent noise or error variance; assuming that it is constant for any (sensory) data channel. In what follows, we relax this assumption and consider generative models in which the states of the world (for example the presence of attentional cues) can change the precision of sensory data. A simple example of this would be the direction (state) in which we pointed a searchlight. This determines which part of the sensorium contains precise information; namely visual information reflected by surfaces that are illuminated. Any coupling between the state of the world (content) and the precision of sensory samples (context) should be an inherent part of veridical generative models of sensory input. Under this perspective, searchlight (spotlight) metaphors for attention become a natural way to think about its functional role (Shulman et al., 1979; Crick, 1984; Cave and Bichot, 1999; Eckstein et al., 2002). Mechanistically, this role is to weight or bias selected sensory channels (Desimone and Duncan, 1995; Maunsell and Treue, 2006; Reynolds and Heeger, 2009; Stokes et al., 2009). In statistical terms, this is formally identical to weighted least squares that underlies all optimal (maximum *a posteriori*) estimates of model parameters. Put simply, this involves weighting data in proportion to their estimated precision.

In predictive coding schemes, sensory data are replaced by prediction error, because this is the only sensory information that has yet to be explained. Here, the weighting is implemented by synaptic gain. We therefore return to the central role of precision-weighted prediction errors in optimal inference. Neurobiologically, this is easy to relate to theories of attentional gain, where the post-synaptic responsiveness of sensory (prediction error) units is modulated by attentional mechanisms (Desimone, 1996; Maunsell and Treue, 2006). We will focus on two neurobiological candidates for modulating synaptic gain that have been linked to attention: synchronous gain (Chawla et al., 1999a) mediated by fast oscillatory or synchronized activity (Womelsdorf and Fries, 2006; Fries et al., 2008) and classical neuromodulatory (e.g., cholinergic) neurotransmission (Schroeder et al., 2001; Hirayama et al., 2004).

Electrophysiologically, desynchronization with increased gamma activity (between 30 and 100 Hz) is seen during attentional tasks in invasive (Steinmetz et al., 2000; Bichot et al., 2005; Fries et al., 2008), and non-invasive EEG and MEG studies (Gruber et al., 1999; Sokolov et al., 1999; Pavlova et al., 2006; Vidal et al., 2006). Gamma oscillations induced with subliminal flicker may improve attention-based performance (Bauer et al., 2009). Furthermore, increased gamma is associated with faster reaction times (Womelsdorf et al., 2006; Fründ et al., 2007). Gamma oscillations can control gain by affording synchronized neuronal discharges a greater influence on the firing rate of downstream neurons (Chawla et al., 1999a; Salinas and Sejnowski, 2001; Zeitler et al., 2008). Gamma activity has also been proposed as a solution to the "binding problem," which we discuss below in relation to attention (Treisman and Schmidt, 1982).

In terms of neurotransmitters, gamma oscillations are profoundly affected by acetylcholine, which is released into sensory cortex from nuclei in the basal forebrain. It acts through both fast ion channel (nicotinic) receptors and slow metabotropic (muscarinic) receptors (Wonnacott, 1997; Zilles et al., 2004; Hasselmo and Giocomo, 2006). Disruption of the cholinergic system by drugs or lesions can interfere with attentional processes, including the Posner paradigm (Voytko et al., 1994; Witte et al., 1997; Dalley et al., 2001; Herrero et al., 2008; Vossel et al., 2008). Acetylcholine appears to increase synaptic gain directly by, for example, reducing spike-frequency adaptation (McCormick and Prince, 1985, 1986). It may also facilitate the induction of gamma oscillations by reducing adaptation in pyramidal cells (Buhl et al., 1998; Börgers et al., 2005), decreasing activity of inhibitory interneurons (Buia and Tiesinga, 2006) or directly inactivating specific interneurons (Xiang et al., 1998). However, the time course of acetylcholine release can be quite protracted (Parikh et al., 2007). This suggests rapid (10–100 ms) attentional mechanisms may rest on an interaction of cholinergic mechanisms with fast activity-dependent modulation of synaptic gain. It is this activity (state) dependent optimization we pursue in this paper.

In summary, it may be the case that attention is the process of optimizing synaptic gain to represent the precision of sensory information (prediction error) during hierarchical inference. Furthermore, if we allow for state-dependent changes in precision, the neurobiology of attention must involve activity-dependent changes in synaptic gain; assuming that neuronal activity represents the states of the world and synaptic gain represents precision. Given this sort of architecture we can, in principle, simulate attentional processing with established (Bayes-optimal) inversion or recognition schemes, using models with state-dependent noise. What follows is an attempt to do this.

This paper comprises four sections. In the remainder of Section we provide a brief review of attention in psychological and neurobiological terms. This section focuses on directed spatial attention

and, in particular, the Posner (cueing) paradigm that emphasizes the importance of valid cues in establishing attentional set during target detection (Posner, 1980). To complement this psychophysical perspective, we consider biased competition models that are based on careful electrophysiological studies of evoked visual responses using intracranial recordings (Desimone, 1996). Biased competition is probably the most established and influential theory that accounts for unit responses in attentional paradigms framed at the level of receptive fields. We also review the concepts of attentional resources and other constructs associated with early and late attentional selection and the feature-integration theory of attention. In Section "Methods," we provide a more technical treatment of perception under the free-energy principle and consider the form of generative models that will be used in later sections. The emphasis here is on generalizing previous models to include state-dependent noise and what this means for their neurobiological optimization or inversion. The resulting inversion scheme corresponds to recognizing the causes of sensory data (that include both states of the world and their precision). We will see that precision is encoded by the synaptic gain of sensory or prediction error-units, which pass messages to units representing conditional expectations about the world. In this scheme, optimization of synaptic gain may correspond to attention. In Section "Results," we present simulations of the Posner paradigm using the recognition scheme of the previous section. This allows us to demonstrate some basic characteristics of attention-based inference; including attentional bias, attentional capture and the cue-validity effect. We supplement a direct interpretation of the probabilistic representations encoded by simulated neuronal activity with simulated psychophysical and electrophysiological data. These simulated responses make some clear predictions about speed-accuracy trade-offs and event-related electrophysiological responses, which we compare against the literature. In the final section, we use the same simulations but present both valid and invalidly cued targets together. This is a rough metaphor for paradigms used to study biased competition and allows us to see if biased competition emerges from the free-energy formulation. We examine this by looking at competition between cues via the effect of an attended cue on the responses evoked by an unattended cue. We conclude with a brief discussion and indicate how the scheme in this paper could be applied to empirical psychophysical and electroencephalographic observations. This is a rather long paper that tries to link a vast literature on the cognitive psychology of attention with a large body of theoretical work on perceptual inference, learning and action. Many readers, who are familiar with one or other of these areas, could easily skip the background material in this section or the next.

## ATTENTION, BIASED COMPETITION AND THE POSNER PARADIGM

In this section, we review some of the key paradigms and theories that have dominated attention research over the past decades. This review can be regarded as a primer for readers who do not have a cognitive neuroscience background (and should be omitted by readers who do). Our focus will be the Posner paradigm, which we simulate in later sections, and biased competition, which is one of the most prevalent electrophysiologically grounded theories of attention. We will also cover some key distinctions, such as the difference between early and late selection and exogenous versus endogenous cueing.

Early cognitive models of attention, although inherently limited by lack of knowledge about the underlying neural processes, elucidated the important difference between early and late selection. Broadbent (1958), working in the auditory domain, suggested that attention operated by selecting stimuli at an early stage of processing, when only basic physical attributes had been encoded. The selected stimulus was then processed by an "identification system," which could handle only one stimulus at a time; to explain why semantic information about unattended stimuli is unavailable to recall (Broadbent, 1952a,b). However, there are stimuli which violate this principle: Moray (1959) demonstrated that a subject's name, which is salient only after semantic processing, could shift attention to a previously unattended auditory stream. The competing theory, that all stimuli are processed semantically before selection for consciousness recall, was posited by Deutsch and Deutsch (1963), whereas Treisman (1964) suggested that unattended stimuli are attenuated so that attention can be diverted to them, if they become behaviorally salient. Lavie (1995) attempted to reconcile these models by demonstrating that perceptual load plays an important role in attentional selection: intuitively, early selection occurs with higher attentional load and late selection with lower load. This differential selection rests on the notion of limited capacity. Many of these ideas can be understood in the framework of biased competition theory, which tries to explain some of the phenomena described above using neurobiological mechanisms.

### *Biased competition*

Biased competition (Desimone and Duncan, 1995) is a model of attention based on electrophysiological studies and earlier behavioral models. Its main contribution was to make the notion of limited capacity or resources more concrete, by suggesting small lower level receptive fields (RFs) compete to drive larger RFs at higher hierarchical levels. Biased competition says that attention is an emergent property of competition between stimuli for attentional resources, which is influenced by the properties of the stimuli and task requirements. Its premise is that, in a crowded visual field, objects must compete for neural representation at some point along the visual processing stream. This can be deduced from the large size of classical RFs in higher visual areas, such as monkey area TE, which can cover up to 25° of visual angle (Gross et al., 1972; Desimone and Gross, 1979). Clearly, many objects can fall into such a visual field but the neuron can only represent (report) one thing with its firing. If an object is represented by these higher-order visual neurons, they are unavailable to represent other objects. Thus the object has consumed some finite "attentional resource."

This premise leads to a key prediction: if two stimuli are presented within a cell's receptive field, the response to both will be smaller than the sum of the response to the stimuli presented separately (Reynolds et al., 1999). Single-cell electrophysiological studies have confirmed that stimuli interact in this mutually suppressive manner in areas V2 and V4 (Reynolds et al., 1999), IT (Rolls and Tovee, 1995) and MT (Recanzone et al., 1997), but not V1, where RFs are so small it is difficult to present competing stimuli (Moran and Desimone, 1985). The average responses of visual cortical areas in fMRI studies in humans mirror the results from electrophysiological studies in animals (Kastner et al., 1998; Beck and Kastner, 2005). An important result is that the maximum spatial

separation between stimuli, which induces suppressive interactions, increases at higher levels of visual processing, which is consistent with increasing receptive field size (Kastner et al., 2001).

Large RFs thus cause stimuli to compete. The probability with which stimuli are represented by cells is thought to be influenced by a number of top-down and bottom-up biases. Bottom-up biases result from the properties of the stimulus itself, such as visual or emotional salience and novelty. Abrupt-onset stimuli, which have high temporal contrast, and thus salience, can attract attention even if they are task-irrelevant (Yantis and Jonides, 1984). In the visual search paradigm, used to address feature-integration and binding (Treisman and Gelade, 1980; Treisman and Schmidt, 1982; Treisman, 1998), subjects are required to find a unique object in a display cluttered with distracters. If the unique object is particularly salient, for example if it is brighter than the distracters or has a unique color, the search time remains constant regardless of the number of distracters. This phenomenon is called "pop-out." Salience does not have to be a function of simple visual attributes: distractor faces exhibiting negative emotions slow search times more than neutral faces (Pessoa et al., 2002). Novelty preference, the well-documented tendency for neurons to respond more strongly to a new stimulus than to a familiar one, can also be considered as a bottom-up bias (Desimone, 1996).

Top-down biases reflect the cognitive requirements of the task rather than the stimuli. Top-down biases have been most studied via spatially-directed attention experiments. Electrophysiologically, if attention is directed toward one of two competing stimuli in a receptive field, the mutually suppressive effect disappears and the response of the cell emulates the response to the attended stimulus alone (Moran and Desimone, 1985; Chelazzi et al., 1993; Treue and Maunsell, 1996; Desimone, 1998). Even in the absence of visual stimulation, baseline increases in firing rate of 30–40% may be seen, if attention is directed to a location within a cell's receptive field (Luck et al., 1997). Indeed, in fMRI studies, responses are increased in visual areas after attentional cuing but before stimulus onset (Chawla et al., 1999b; Kastner et al., 1999; O'Connor et al., 2002; Stokes et al., 2009). In addition, cells respond more strongly to attended than unattended stimuli (Luck et al., 1997). Thus, top-down bias has both additive (baseline shift) and multiplicative (attentional gain) components that may depend on each other (Chawla et al., 1999b). In summary, biased competition is a mechanistic framework, which provides a plausible neurobiological account of attention. Later, we will see how biased competition emerges naturally in predictive coding formations of Bayes-optimal perception.

### *The Posner paradigm*

In later sections we will simulate optimal perception under the Posner task, a covert attention task. Attending to an object usually involves looking at it; that is placing its image at the fovea (the central area of the retina with highest acuity). However, attention can be directed independently of eye movement (Posner et al., 1978). Under the Posner paradigm, subjects are required to foveate a central spot and respond as quickly as possible to the appearance of a peripheral target. The target is cued with either a central arrow indicating the side it will appear on, or a peripheral box around the target's eventual location. The cue is correct (valid) most (usually 80%) of the time. Posner found that reaction times to validly cued targets were significantly shorter than to invalidly cued targets, which appeared on the opposite side. This demonstrated that attention could be moved to salient locations in the absence of gaze shift. The cuing seen in the Posner paradigm seems to be separable from the phenomenon of "alerting," in which a non-directional signal indicates the imminent onset of a target (Fernandez-Duque and Posner, 1997; Posner, 2008). Subjects are quicker to respond to a target if the cue indicates the location of the target than when it only indicates the timing (Davidson and Marrocco, 2000). In addition, a pharmacological double dissociation exists such that inhibitors of the cholinergic system selectively reduce the benefits of spatial cues, while noradrenergic inhibitors selectively reduce the benefits of alerting cues (Marrocco et al., 1994). Furthermore, dopamine and noradrenalin antagonists can reduce the reaction time cost of invalidly cued targets, while preserving the validly effect (Clark et al., 1989). However, this effect may be due to the role of noradrenalin in task switching (Sara, 1998; Yu and Dayan, 2005).

The two types of cues used in the Posner paradigm – central and peripheral – show the same facilitation effect. However, they may operate by different mechanisms. Peripheral stimuli are labeled as "exogenous," because the change in attention is triggered by an external event. It is well established that abrupt-onset peripheral stimuli can attract attention via bottom-up mechanisms (Yantis and Jonides, 1984), even when task-irrelevant (Theeuwes, 1991). Central stimuli are "endogenous" because they do not in themselves indicate target location; attention must be directed to the correct location according to information conveyed by the cue. The most common central cues are inherently directional: an arrow pointing to where the target will appear, or an asterisk just to one side of fixation. Although cues such as this may automatically "push" attention, even when the subject has been told the cue is invalid (Hommel et al., 2001).

Exogenous and endogenous cuing fit well with biased competition theory: exogenous cuing can be thought of as a bottom-up bias, based on the prior expectation that salient events recur in the same part of the visual field. On the other hand the effect of endogenous cues must be mediated by top-down bias. However, these top-down effects do not necessarily call on semantic or explicit processing: for example, Decaix et al. (2002) examined performance on the Posner paradigm when subjects were not informed about the cue-target relationship but subjects still learnt cue-target relationships within 90 trials, and performance was independent of whether the learnt relationship was accessible to verbal report. Bartolomeo et al. (2007) compared performance of informed and non-informed subjects and found no effect of explicit knowledge on reaction time. Finally, Risko and Stolz (2010) demonstrated that knowledge of the proportion of valid trials did not affect reaction time. In short, the basic phenomena disclosed in the Posner paradigm may not depend on high-level cognitive processing. This suggests that a low-level simulation of perceptual processing should be able to account for cue-validity effects. This is what we attempt to show and demonstrate that cue-validity effects are Bayes-optimal. In the next section, we review the principles that lie behind Bayes-optimal perception and apply these principles to the Posner paradigm in the subsequent section.

## METHODS

### FREE-ENERGY AND BAYES-OPTIMAL PERCEPTION

This section reviews the theoretical principles used later to explain perception and attention. This treatment is a bit technical but serves as a standalone summary of the mathematical principles behind the simulations of subsequent sections. More mathematical details can be found in (Friston et al., 2010b). Readers who are familiar with generalized predictive coding should skip directly to "Perception and Attention." We start with the objective of the free-energy formulation; namely to suppress surprise. We end with a set of ordinary differential equations describing changes in synaptic activity, gain and efficacy. These dynamics correspond to perceptual inference, attention and learning respectively. Basically, the resulting scheme can be regarded as a form of evidence accumulation (Mazurek et al., 2003) that is formally equivalent to generalized predictive coding. Free-energy is a bound on surprise and is therefore a bound on the log-evidence for the brain's generative model of its world. The second half of this section considers particular forms of these generative models, with a focus on state-dependent noise and the implications for the neurobiology of perception. The amount of this noise is measured in terms of its variance, which reflects the degree of randomness in the processes generating sensory data. Inverse variance is called precision; therefore precision increases with certainty about states of the world. We will see that precision is encoded by the post-synaptic gain of sensory or prediction error-units. This means that state-dependent changes in precision may be modeled in the brain by activity-dependent modulation of the synaptic gain of principal cells originating forward connections. This is the optimization we associate with attention.

#### *Recognition from basic principles*

Our objective, given a model (brain), $m$, is to minimize the average uncertainty (entropy) about some generalized sensory states, $\tilde{s} = s \oplus s' \oplus s'' \oplus \ldots \in S$ it experiences ($\oplus$ means concatenation). Generalized states comprise the state itself, its velocity, acceleration, jerk, *etc*. This average uncertainty is

$$H(S \mid m) = -\int p(\tilde{s} \mid m) \ln p(\tilde{s} \mid m) d\tilde{s} \tag{1}$$

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

$$H(S \mid m) \propto -\int_0^T dt \ln p(\tilde{s}(t) \mid m) \tag{2}$$

Minimizing sensory entropy therefore corresponds to maximizing the accumulated log-evidence for a model of the world. Although sensory entropy cannot be minimized directly, we can create an upper bound $\mathcal{S}(\tilde{s}, q) \geq H(S)$. This bound is induced with a recognition density $q(t) := q(\vartheta)$ on the causes (i.e., environmental states and parameters) of sensory signals. We will see later that these causes comprise time-varying states $u(t) \subset \vartheta$ and slowly varying parameters $\varphi(t) \subset \vartheta$. The recognition density is sometimes called a proposal density and becomes the conditional density over causes, when it minimizes the bound. The bound itself is the path integral of free-energy $\varphi(t)$, which is created simply by adding a non-negative function of the recognition density to surprise:

$$\begin{aligned} \mathcal{S} &= \int dt \mathcal{F}(t) \\ \mathcal{F}(t) &= -\ln p(\tilde{s}(t) \mid m) + D_{KL} \\ D_{KL} &= \left\langle \ln q(\vartheta) - \ln p(\vartheta \mid \tilde{s}, m) \right\rangle_q \end{aligned} \tag{3}$$

This function is a Kullback–Leibler divergence $D_{KL}$ and is greater than zero, with equality when $q(\vartheta) = p(\vartheta \mid \tilde{s}, m)$ is the true conditional density. This means that minimizing free-energy, by changing $q(\vartheta)$, makes the recognition density an approximate conditional density on sensory causes. This is Bayes-optimal recognition. The free-energy can be evaluated easily because it is a function of the recognition density and a generative model $\mathcal{L}(t)$ entailed by $m$

$$\begin{aligned} \mathcal{F}(t) &= \left\langle \mathcal{L}(t) \right\rangle_q - \mathcal{H}(t) \\ \mathcal{L}(t) &= -\ln p(\tilde{s}(t), \vartheta \mid m) \\ \mathcal{H}(t) &= -\left\langle \ln q(t) \right\rangle_q \end{aligned} \tag{4}$$

The free-energy has been expressed here in terms of $\mathcal{H}(t)$, the negentropy of $q(t)$ and an energy $\mathcal{L}(t)$ expected under $q(t)$. The energy $\mathcal{L}(t)$ reports the surprise about sensations and their causes under a generative model. If we assume that recognition density $q(\vartheta) = \mathcal{N}(\mu, C)$ is Gaussian (known as the Laplace assumption), we can express free-energy in terms of the mean and covariance of the recognition density

$$\mathcal{F} = \mathcal{L}(\mu) + \tfrac{1}{2} tr(C\mathcal{L}_{\mu\mu}) - \tfrac{1}{2} \ln |C| - \tfrac{n}{2} \ln 2 \quad e \tag{5}$$

Where $n = \dim(\mu)$ and subscripts denote derivatives. We can now minimize free-energy with respect to the conditional precision (inverse covariance). The free-energy is minimized when $\mathcal{P} = \mathcal{C}^{-1} = \mathcal{L}_{\mu\mu} \Rightarrow \mathcal{F}_C = 0 \Rightarrow \delta_C \mathcal{S} = 0$ and allows us to eliminate $\mathcal{C}$ from Eq. 5

$$\mathcal{F} = \mathcal{L}(\mu) + \tfrac{1}{2} \ln |\mathcal{L}_{\mu\mu}| - \tfrac{n}{2} \ln 2 \tag{6}$$

Crucially, this means the free-energy is only a function of the conditional mean or expectation. The expectations that minimize free-energy are the solutions to the following differential equations. For the generalized states $\tilde{u}(t) \subset \vartheta$

$$\begin{aligned} \dot{\tilde{\mu}}^{(u)} &= \mathcal{D}\tilde{\mu}^{(u)} - \mathcal{F}_{\tilde{u}} \\ &\Leftrightarrow \\ \dot{\mu}^{(u)} &= \mu'^{(u)} - \mathcal{F}_u \\ \dot{\mu}'^{(u)} &= \mu''^{(u)} - \mathcal{F}_{u'} \\ \dot{\mu}''^{(u)} &= \mu'''^{(u)} - \mathcal{F}_{u''} \\ &\vdots \end{aligned} \tag{7}$$

Where $\mathcal{D}$ is a derivative matrix operator with identity matrices above the leading diagonal, such that $\mathcal{D}\tilde{u} = u' \oplus u'' \oplus \ldots$. Here and throughout, we assume all gradients are evaluated at the mean; here $\tilde{u} = \tilde{\mu}^{(u)}$. The stationary solution of Eq. 7 minimizes free-energy and its path integral: $\dot{\tilde{\mu}}^{(u)} - \mathcal{D}\tilde{\mu}^{(u)} = 0 \Rightarrow \mathcal{F}_{\tilde{u}} = 0 \Rightarrow \delta_{\tilde{u}} \mathcal{S} = 0$. This ensures that when free-energy is minimized the mean of the motion is the motion of the mean; that is $\mathcal{F}_{\tilde{u}} = 0 \Rightarrow \dot{\tilde{\mu}}^{(u)} = \mathcal{D}\tilde{\mu}^{(u)}$.

For slowly varying parameters $\varphi(t) \subset \vartheta$ this motion disappears and we can use the following scheme

$$\begin{aligned}\dot{\mu}^{(\varphi)} &= \mu'^{(\varphi)} \\ \dot{\mu}'^{(\varphi)} &= -\mathcal{F}_{\varphi} - \kappa\mu'^{(\varphi)}\end{aligned} \tag{8}$$

Here, the solution $\dot{\tilde{\mu}}^{(\varphi)} = 0$ minimizes free-energy, under constraint that the motion of the expected parameters is small: $\dot{\mu}^{(\varphi)} = \dot{\mu}'^{(\varphi)} = 0 \Rightarrow \mathcal{F}_{\varphi} = 0 \Rightarrow \delta_{\varphi}\mathcal{S} = 0$. The last equality $\delta_{\varphi}\mathcal{S} = 0$ just means that variations in the parameters do change the path integral of free-energy (cf, keeping to the floor of a valley to minimize the average height of ones path).

Equations 7 and 8 prescribe recognition dynamics for the expected states and parameters of the world respectively. The dynamics for states can be thought of as a gradient descent in a frame of reference that moves with the expected motion of the world (cf, surfing a wave). Conversely, the dynamics for the parameters can be thought of as a gradient descent that resists transient fluctuations with a damping term $(-\kappa\mu'^{(\varphi)})$, which instantiates our prior belief that the fluctuations in the parameters are small. We use $\kappa = N$, where $N$ is the number of sensory samples.

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

### *Hierarchical dynamic models*

We next introduce a very general model based on the hierarchal dynamic model discussed in Friston (2008). We will assume that any sensory data can be modeled with a special case of this model

$$\begin{aligned}s &= f^{(v)}(x, v, \theta) + z^{(v)} : z^{(v)} \sim \mathcal{N}\left(0, \Sigma^{(v)}(x, v, \gamma)\right) \\ \dot{x} &= f^{(x)}(x, v, \theta) + z^{(x)} : z^{(x)} \sim \mathcal{N}\left(0, \Sigma^{(x)}(x, v, \gamma)\right)\end{aligned} \tag{9}$$

The non-linear functions $f^{(u)} : u \in v, x$ represent a sensory mapping and equations of motion respectively and are parameterized by $\theta \subset \varphi$. The variables $v \subset u$ are referred to as hidden causes, while hidden states $x \subset u$ meditate the influence of the causes on sensory data and endow the system with memory. We assume the random fluctuations $z^{(u)}$ are analytic, such that the covariance of $\tilde{z}^{(u)}$ is well defined. Unlike our previous treatments (Friston, 2008), this model allows for state-dependent changes in the amplitude of random fluctuations. It is this generalization that furnishes a model of attention and introduces the key distinction between the effect of states on first- and second-order sensory dynamics. These effects are meditated by the vector and matrix functions $f^{(u)} \in \Re^{dim(u)}$ and $\Sigma^{(u)} \in \Re^{\dim(u) \times \dim(u)}$ respectively, which are parameterized by first- and second-order parameters $\{\theta, \gamma\} \subset \varphi$.

Under local linearity assumptions, the generalized motion of the sensory response and hidden states can be expressed compactly as

$$\begin{aligned}\tilde{s} &= \tilde{f}^{(v)} + \tilde{z}^{(v)} \\ \mathcal{D}\tilde{x} &= \tilde{f}^{(x)} + \tilde{z}^{(x)}\end{aligned} \tag{10}$$

Where the generalized predictions are

$$\tilde{f}^{(u)} = \begin{bmatrix} f^{(u)} = f^{(u)} \\ f'^{(u)} = f_x^{(u)} x' + f_v^{(u)} v' \\ f''^{(u)} = f_x^{(u)} x'' + f_v^{(u)} v'' \\ \vdots \end{bmatrix} \tag{11}$$

Equation 10 means that Gaussian assumptions about the random fluctuations specify a generative model in terms of a likelihood and empirical priors on the motion of hidden states

$$\begin{aligned}p(\tilde{s} \,|\, \tilde{x}, \tilde{v}, \theta, m) &= \mathcal{N}\left(\tilde{f}^{(v)}, \tilde{\Sigma}^{(v)}\right) \\ p(D\tilde{x} \,|\, x, \tilde{v}, \theta, m) &= \mathcal{N}\left(\tilde{f}^{(x)}, \tilde{\Sigma}^{(x)}\right)\end{aligned} \tag{12}$$

These probability densities are encoded by their covariances $\tilde{\Sigma}^{(u)}$ or precisions $\tilde{\Pi}^{(u)} := \tilde{\Pi}(x, v, \gamma^{(u)})$ with precision parameters $\gamma \subset \varphi$ that control the amplitude and smoothness of the random fluctuations. Generally, the covariances factorize; $\tilde{\Sigma}^{(u)} = V^{(u)} \otimes \Sigma^{(u)}$ into a covariance proper and a matrix of correlations $V^{(u)}$ among generalized fluctuations that encodes their smoothness.

Given this generative model we can now write down the energy as a function of the conditional means, which has a simple quadratic form (ignoring constants)

$$\begin{aligned}\mathcal{L} &= \mathcal{L}^{(v)} + \mathcal{L}^{(x)} + \mathcal{L}^{(\varphi)} \\ \mathcal{L}^{(v)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(v)T}\tilde{\Pi}^{(v)}\tilde{\varepsilon}^{(v)} - \tfrac{1}{2}\ln|\tilde{\Pi}^{(v)}| \\ \mathcal{L}^{(x)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(x)T}\tilde{\Pi}^{(x)}\tilde{\varepsilon}^{(x)} - \tfrac{1}{2}\ln|\tilde{\Pi}^{(x)}| \\ \mathcal{L}^{(\varphi)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \tfrac{1}{2}\ln|\tilde{\Pi}^{(\varphi)}| \\ \tilde{\varepsilon}^{(v)} &= \tilde{s} - \tilde{f}^{(v)} \\ \tilde{\varepsilon}^{(x)} &= \mathcal{D}\tilde{\mu}^{(x)} - \tilde{f}^{(x)} \\ \tilde{\varepsilon}^{(\varphi)} &= \tilde{\mu}^{(\varphi)} - \tilde{\eta}^{(\varphi)}\end{aligned} \tag{13}$$

Here, the auxiliary variables $\tilde{\varepsilon}^{(j)} : j \in v, x, \varphi$ are prediction errors for sensory data, the motion of hidden states and parameters respectively. The predictions for the states are $\tilde{f}^{(u)}(\mu) : u \in v, x$ and the predictions for the parameters are the prior expectations $\tilde{\eta}^{(\varphi)}$. Equation 13 assumes flat priors on the states and that priors $p(\varphi \,|\, m) = \mathcal{N}(\tilde{\eta}^{(\varphi)}, \tilde{\Sigma}^{(\varphi)})$ on the parameters are Gaussian. We next consider hierarchical forms of this model. These are just special cases of Eq. 9, in which we make certain conditional independencies explicit. Although they may look more complicated, they are simpler than the general form above. They are useful because they provide an empirical Bayesian perspective on inference and learning that may be exploited by the brain. Hierarchical dynamic models have the following form

$$\begin{aligned}s &= f^{(v)}\left(x^{(1)}, v^{(1)}, \theta\right) + z^{(1,v)} \\ \dot{x}^{(1)} &= f^{(x)}\left(x^{(1)}, v^{(1)}, \theta\right) + z^{(1,x)} \\ &\vdots \\ v^{(i-1)} &= f^{(v)}\left(x^{(i)}, v^{(i)}, \theta\right) + z^{(i,v)} \\ \dot{x}^{(i)} &= f^{(x)}\left(x^{(i)}, v^{(i)}, \theta\right) + z^{(i,x)} \\ &\vdots \\ v^{(h-1)} &= \eta^{(v)} + z^{(h,v)}\end{aligned} \tag{14}$$

Again, $f^{(i,u)} := f^{(u)}(x^{(i)}, v^{(i)}, \theta) : u \in v, x$ are continuous non-linear functions and $\eta^{(v)}(t)$ is a prior mean on the hidden causes at the highest level. The random terms $z^{(i,u)} \sim \mathcal{N}(0, \Sigma(x^{(i)}, v^{(i)}, \gamma^{(i,u)})$ are conditionally independent and enter each level of the hierarchy. They play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causes $v = v^{(1)} \oplus v^{(2)} \oplus \ldots$ link levels, whereas the hidden states $x = x^{(1)} \oplus x^{(2)} \oplus \ldots$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. This input can enter non-linearly to produce quite complicated generalized convolutions with deep (hierarchical) structure. This structure appears in the energy as empirical priors $\mathcal{L}^{(i,u)} : u \in x, v$ where, ignoring constants

$$\begin{aligned}
\mathcal{L} &= \sum\nolimits_i \mathcal{L}^{(i,v)} + \sum\nolimits_i \mathcal{L}^{(i,x)} + \mathcal{L}^{(\varphi)} \\
\mathcal{L}^{(i,v)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(i,v)T}\tilde{\Pi}^{(i,v)}\tilde{\varepsilon}^{(i,v)} - \tfrac{1}{2}\ln|\tilde{\Pi}^{(i,v)}| \\
\mathcal{L}^{(i,x)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(i,x)T}\tilde{\Pi}^{(i,x)}\tilde{\varepsilon}^{(i,x)} - \tfrac{1}{2}\ln|\tilde{\Pi}^{(i,x)}| \\
\tilde{\varepsilon}^{(i,v)} &= \tilde{v}^{(i-1)} - \tilde{f}^{(i,v)} \\
\tilde{\varepsilon}^{(i,x)} &= D\tilde{x}^{(i)} - \tilde{f}^{(i,x)}
\end{aligned} \tag{15}$$

Note that the data enter the prediction errors at the lowest level; $\tilde{\varepsilon}^{(1,v)} = \tilde{s} - f^{(1,v)}$. At intermediate levels the prediction errors mediate empirical priors on the causes.

In summary, these models are as general as one could imagine; they comprise hidden causes and states, whose dynamics can be coupled with arbitrary (analytic) non-linear functions. Furthermore, these states can be subject to random fluctuations with state-dependent changes in amplitude and arbitrary (analytic) autocorrelation functions. A key aspect is their hierarchical form, which induces empirical priors on the causes. In the next section, we look at the recognition dynamics entailed by this form of generative model, with a particular focus on how recognition might be implemented in the brain. We consider perception first and then attention. For completeness, we also mention learning; but will only pursue this in subsequent papers on learning and related phenomena (e.g., inhibition of return; Posner and Cohen, 1984; Rafal et al., 1989).

### *Perception and attention*

If we now write down the recognition dynamics (Eq. 7) using precision-weighted prediction errors $\xi^{(i,u)} = \tilde{\Pi}^{(i,u)}\tilde{\varepsilon}^{(i,u)}$ from Eq. 15, one can see the hierarchical message passing this scheme entails (ignoring the derivatives of the energy curvature);

$$\begin{aligned}
\dot{\tilde{\mu}}^{(i,v)} &= \mathcal{D}\tilde{\mu}^{(i,v)} + \chi_{\tilde{v}}^{(i,v)}\xi^{(i,v)} + \chi_{\tilde{v}}^{(i,x)}\xi^{(i,x)} + \lambda_{\tilde{v}}^{(i,v)} + \lambda_{\tilde{v}}^{(i,x)} - \xi^{(i+1,v)} \\
\dot{\tilde{\mu}}^{(i,x)} &= \mathcal{D}\tilde{\mu}^{(i,x)} + \chi_{\tilde{x}}^{(i,v)}\xi^{(i,v)} + \chi_{\tilde{x}}^{(i,x)}\xi^{(i,x)} + \lambda_{\tilde{x}}^{(i,v)} + \lambda_{\tilde{x}}^{(i,x)} - \mathcal{D}^T\xi^{(i,x)} \\
\xi^{(i,v)} &= \tilde{\Pi}^{(i,v)}\tilde{\varepsilon}^{(i,v)} = \tilde{\Pi}^{(i,v)}\left(\tilde{\mu}^{(i-1,v)} - \tilde{f}^{(i,v)}\right) \\
\xi^{(i,x)} &= \tilde{\Pi}^{(i,x)}\tilde{\varepsilon}^{(i,x)} = \tilde{\Pi}^{(i,x)}\left(\mathcal{D}\tilde{\mu}^{(i,x)} - \tilde{f}^{(i,x)}\right) \\
\chi_w^{(i,v)} &= \tilde{f}_w^{(i,v)T} - \tfrac{1}{2}\tilde{\varepsilon}^{(i,v)T}\tilde{\Omega}_w^{(i,v)} \\
\chi_w^{(i,x)} &= \tilde{f}_w^{(i,x)T} - \tfrac{1}{2}\tilde{\varepsilon}^{(i,x)T}\tilde{\Omega}_w^{(i,x)}
\end{aligned} \tag{16}$$

Here, we have assumed the amplitude of random fluctuations is parameterized in terms of log-precisions, where

$$\begin{aligned}
\tilde{\Pi}^{(i,u)} &= R^{(i,u)} \otimes diag\left(\exp\left(\pi^{(i,u)}\right)\right) \\
\tilde{\Omega}^{(i,u)} &= I^{(i,u)} \otimes diag\left(\pi^{(i,u)}\right)
\end{aligned} \tag{17}$$

The vector function $\pi^{(i,u)} := \pi(x, v, \gamma^{(i,u)})$ returns state-dependent log-precisions and $R^{(i,u)}$ is the inverse smoothness matrix $V^{(i,u)}$. In what follows we will quantify the amplitude (variance) of random fluctuations in terms of log-precisions, such that the associated variance is $\exp(-\pi^{(i,u)})$. With this particular form for the precisions, the terms $\tilde{\Omega}_w^{(i,u)}$ and $\lambda_w^{(i,u)} = tr\left(\tilde{\Omega}_w^{(i,u)}\right)$ are constant for states $w \in \tilde{v}, \tilde{x}$ that affect the log-precisions linearly and zero if they have no effect.

It is difficult to overstate the generality and importance of Eq. 16: it grandfathers nearly every known statistical estimation scheme, under parametric assumptions about either additive or multiplicative noise. These range from ordinary least squares to advanced variational deconvolution schemes (see Friston, 2008). For example, the schemes used to invert stochastic dynamic causal models of imaging time series (e.g., Daunizeau et al., 2009) use Eq. 16. This is generalized predictive coding.

In neural network terms, Eq. 16 says that error-units receive messages from the states in the same level and the level above. Conversely, state-units are driven by error-units in the same level and the level below, were $\chi_w^{(i,u)} : u \in v, x$ are the forward connection strengths to the state-unit representing $w \in \tilde{v}, \tilde{x}$. Crucially, recognition requires only the prediction error from the lower level $\xi^{(i,v)}$ and the level in question, $\xi^{(i,x)}$ and $\xi^{(i+1,v)}$ (see **Figure 1**). These constitute bottom-up and lateral messages that drive conditional means $\tilde{\mu}^{(i,u)}$ toward a better prediction, which reduces the prediction error in the level below. These top-down and lateral predictions correspond to $\tilde{f}^{(i,u)}$. This is the essence of recurrent message passing between hierarchical levels to optimize free-energy or suppress prediction error; i.e., perceptual inference (see Friston, 2008 for a more detailed discussion).

In the present context, the key thing about this scheme is that the precisions $\tilde{\Pi}^{(i,u)} := \tilde{\Pi}(v^{(i)}, x^{(i)}, \gamma^{(i,u)})$ depend on the expected hidden causes and states. It is this dependency that we propose mediates attentional processing. Equation 16 tells us that the state-dependent precisions modulate the responses of the error-units to their pre-synaptic inputs. This modulation depends on the conditional expectations about the states and suggests something intuitive; attention is mediated by activity-dependent modulation of the synaptic gain of principal cells that convey sensory information (prediction error) from one cortical level to the next. These are generally thought to be the superficial pyramidal cells responsible for generating EEG signals. More specifically, precision sets the synaptic gain of error-units to their top-down and lateral inputs. In hierarchical models, the gain modulation of error-unit activity $\xi^{(i,u)}$ depends on $\tilde{\Pi}\left(v^{(i)}, x^{(i)}, \gamma^{(i,u)}\right)$ and therefore depends on the conditional expectations of $x^{(i)}$ in the current level and $v^{(i)}$ in the level above. This translates into a top-down control of synaptic gain in principal (superficial pyramidal) cells elaborating prediction errors and fits comfortably with the modulatory effects of top-down connections in cortical hierarchies that have been associated with attention. Note that the precisions or synaptic gain $\tilde{\Pi}^{(i,u)}$ also depends on the slowly varying parameters $\gamma \subset \varphi$ responsible for learning. It is these parameters we associate with the slower dynamics of classical neuromodulation (e.g., cholinergic neurotransmission; Friston, 2008).

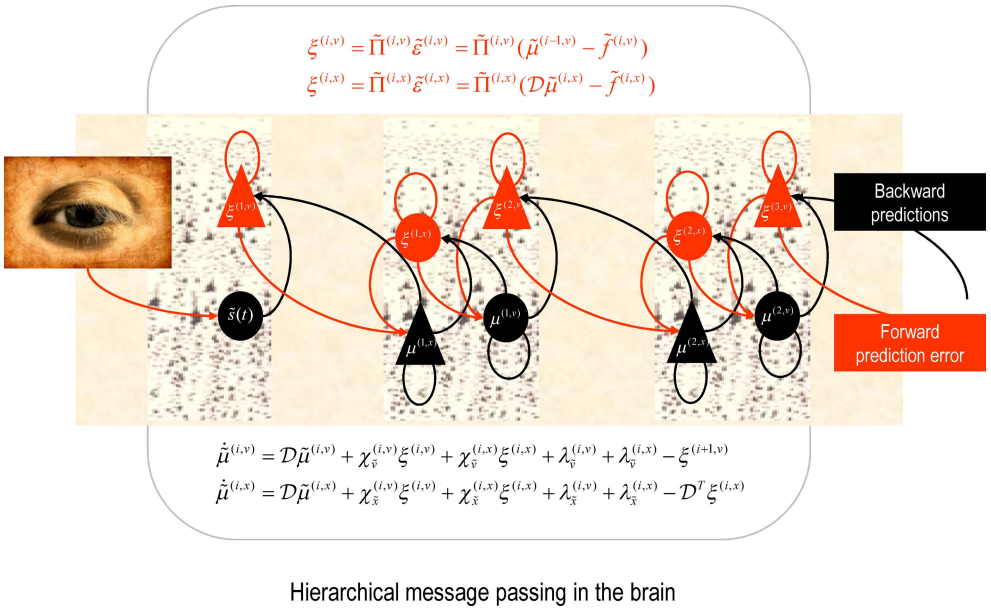



**FIGURE 1 | Schematic detailing the neuronal architecture that might implement the generalized predictive coding described in Eq. 16**. This shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that construct predictions (Mumford, 1992; Friston, 2008). These predictions try to explain away prediction error in lower levels. In this scheme, the sources of forward and backward connections are superficial and deep pyramidal cells respectively. The equations represent a gradient descent on free-energy under a hierarchical dynamic model (see Eq. 16). State-units are in black and error-units in red. Here, neuronal populations are deployed hierarchically within three cortical areas (or macro-columns). Subscripts denote derivatives.

### *Perceptual learning*

Perceptual learning corresponds to optimizing the first-order parameters $\theta \subset \varphi$ according to Eq. 8. This describes a process that is remarkably similar to models of associative plasticity based on correlated pre- and post-synaptic activity. This can be seen most easily by assuming an explicit form for the generating functions; for example (for a single parameter and ignoring high-order derivatives)

$$\begin{aligned}\dot{x}_q^{(i)} &= f_q^{(i,x)} = \theta x_p^{(i)} \Rightarrow \\ \dot{\mu}^{(\theta)} &= \mu'^{(\theta)} \\ \dot{\mu}'^{(\theta)} &= -\tilde{\mu}_p^{(i,x)T}\xi_q^{(i,x)} - \Pi^{(\theta)}\mu^{(\theta)} - \kappa\mu'^{(\theta)}\end{aligned} \tag{18}$$

Here $\mu^{(\theta)}$ is the connection strength mediating the influence of the $p$-th hidden state on the motion of the $q$-th, at hierarchical level $i \in 1, 2, \ldots$ This strength changes in proportion to a "synaptic tag" $\mu'^{(\theta)}$ that accumulates in proportion to the product of the $p$-th pre-synaptic input $\tilde{\mu}_p^{(i,x)}$ and post-synaptic response $\xi_q^{(i,x)}$ of the $q$-th error-unit (first term of Eq. 18). The tag is auto-regulated by the synaptic strength and decays with first-order kinetics (second and third terms respectively).

We conclude by considering the equivalent dynamics for the second-order (precision) parameters $\gamma \subset \varphi$. These precision parameters govern lateral and top-down state-dependent gain control and are learned according to Eq. 8 (for a single parameter)

$$\begin{aligned}\dot{\mu}^{(\gamma)} &= \mu'^{(\gamma)} \\ \dot{\mu}'^{(\gamma)} &= -\tfrac{1}{2}\tilde{\varepsilon}^{(i,u)T}\tilde{\Omega}_\gamma^{(i,u)}\tilde{\Pi}^{(i,u)}\tilde{\varepsilon}^{(i,u)} + \tfrac{1}{2}tr\left(\tilde{\Omega}_\gamma^{(i,u)}\right) - \Pi^{(\gamma)}\mu^{(\gamma)} - \kappa\mu'^{(\gamma)}\end{aligned} \tag{19}$$

As with perceptual learning, the precision parameters change in proportion to a synaptic tag that decays in proportion to the precision *per se* and the amount of tag. This tag accumulates sum of squared predications errors, weighted by $\tilde{\Omega}_\gamma^{(i,u)}\tilde{\Pi}^{(i,u)}$ to select those errors whose precision is encoded. In this paper, we will focus on perceptual inference and return to learning in a later paper. The numerics of the integration scheme used to simulate inference (Eq. 16) and learning (Eq. 8) are provided in Appendix "Integrating the Recognition Dynamics (Generalized Filtering)."

### *Summary*

In this section, we have applied the general form of recognition dynamics prescribed by the free-energy treatment to a generic hierarchical dynamic model with state-dependent noise. When formulated as a neuronal message-passing scheme something quite important emerges; namely, a lateral and top-down modulation of synaptic gain in principal cells that convey sensory information (prediction error) from one cortical level to the next. It is this necessary and integral component of perpetual inference that we associate with attention.

## RESULTS

### SIMULATING THE POSNER PARADIGM

In this section, we use the hierarchical dynamic model of the previous section as a generative model of stimuli used in the Posner paradigm. Inversion of this model, using generalized predictive coding (Eq. 16) will be used to simulate neuronal responses. This allows us to explore some of the inferential and empirical aspects of perception the Posner paradigm was designed to study. We first

describe the particular model and stimuli used. We then present simulated responses to valid and invalid targets to highlight their differences, in terms of implicit speed-accuracy trade-offs and their electrophysiological correlates.

### *The Posner model*

We deliberately tried to keep the generative model as simple as possible so that its basic behavior can be seen clearly. To this end, we used a model with two levels, the first representing visual input and the second representing the causes of that input. The model has the following form, which we unpack below.

$$
\begin{aligned}
s &= \begin{bmatrix} s_L \\ s_C \\ s_R \end{bmatrix} = v^{(1)} + z^{(1,v)} \\
\dot{x}^{(1)} &= \begin{bmatrix} \dot{x}_L^{(1)} \\ \dot{x}_R^{(1)} \end{bmatrix} = \underbrace{\tfrac{1}{4}\begin{bmatrix} 1 \\ -1 \end{bmatrix} v_L^{(1)} + \tfrac{1}{4}\begin{bmatrix} -1 \\ 1 \end{bmatrix} v_R^{(1)}}_{\text{exogenous}} + \underbrace{\tfrac{1}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix} v_C^{(1)}}_{\text{endogenous}} - \tfrac{1}{32} x^{(1)} + z^{(1,x)} \\
v^{(1)} &= \begin{bmatrix} v_L^{(1)} \\ v_C^{(1)} \\ v_R^{(1)} \end{bmatrix} = z^{(2,v)} \\
z^{(1,v)} &\sim \mathcal{N}\left(0, diag\left(\exp\left(-\pi^{(1,v)}\right)\right)\right) \\
z^{(1,x)} &\sim \mathcal{N}\left(0, e^{-8} I\right) \qquad \pi^{(1,v)} = 2 + \gamma \begin{bmatrix} \bar{x}_L^{(1)} \\ 2 \\ \bar{x}_R^{(1)} \end{bmatrix} \\
z^{(2,v)} &\sim \mathcal{N}\left(0, I\right)
\end{aligned}
\tag{20}
$$

This minimal model has all the ingredients needed to demonstrate some complicated but intuitive phenomena. It helps to bear in mind that this is a generative model of how sensory data are caused that is used by the (synthetic) brain; we actually generated sensory data by simply presenting visual cues in various positions. Because this is a model the prior assumptions about the causes of visual input are that they are just random fluctuations about a mean of zero; i.e., $v^{(1)} = z^{(2,v)}$. Perception (model inversion) uses this model to explain sensory input in terms of conditional expectations about what caused that input.

We first describe the model in terms of the way that it explains sensory data; in other words, how it maps from causes to consequences. We then reprise the description in terms of its inversion; namely, mapping from consequences (sensory data) to causes (percepts). As a generative model, Eq. 20 describes how hidden causes generate sensory input. There are three causes, which are just random fluctuations with a mean of zero and a precision of one. Two causes generate targets in the right and left visual fields $v_{L,R}^{(1)}$ respectively and a third $v_C^{(1)}$ generates a cue. This cue establishes the probabilistic context in which the first two causes are expressed. This context is determined by hidden states $x_{L,R}^{(1)}$, which modulate the log-precision (inverse amplitude) of random fluctuations that are added to the hidden causes to create sensory data. Here, $\bar{x}_{L,R}^{(1)}$ are mean centered versions and $\gamma \subset \varphi$ is a constant that controls the potency of hidden states. Unless stated otherwise we used $\gamma = 2$. Crucially, the hidden causes induce sensory signals directly but also drive increases or decreases in the hidden states (second equality in Eq. 20). The two hidden states represent a high precision on the left and a low precision on the right and *vice versa*. In other words, they induce a redistribution of precision to the left and right in a complementary way. The first cause $v_L^{(1)}$ generates a stimulus $s_L$ in the left hemi-field and drives its corresponding hidden state $x_L^{(1)}$ to increase precision on the left; similarly for the right cause. This means that hidden causes not only cause sensory signals but also augment their precision. In other words, they cause precise visual information with spatial specificity.

Note how the log-precision $\pi^{(1,v)}(x^{(1)}, \gamma)$ of sensory noise $z^{(1,v)} \sim \mathcal{N}(0, diag(\exp(-\pi^{(1,v)})))$ depends on the hidden states. The motivation for this dependency is simple: high levels of signal are generally associated with lower levels of noise (i.e., high signal to noise). More formally, this represents a prior expectation that sensory input conforms to Weber's law (Formankiewicz and Mollon, 2009): for stimulus intensities with a fixed precision (of sensory noise), under Weber's law (after log-transform) the log-precision scales with the magnitude of the signal. See Appendix "State-Dependent Noise and Weber's Law."

The ensuing increase in local precision can be regarded as analogous to exogenous cuing in the Posner paradigm, in the sense that it co-localizes in space and time with its sensory expression. Endogenous effects on precision that do not co-localize correspond to the probabilistic context established by $v_C^{(1)}$ that enables endogenous cuing. This hidden cause drives hidden states to increase precision on the right. One can think of $s_C$ as the corresponding endogenous cue in the center of the field of view. Note that the hidden states decay slowly. This represents a formal prior that once a cause has been expressed in any part of the visual field, subsequent causes will be expressed in the same vicinity with a high sensory precision. The time constants for the accumulation of hidden causes by hidden states (4 and 2) and their decay (32) are somewhat arbitrary, because we can assign any units of time to the dynamics. The important thing is that the decay is slower than the accumulation (by factors of 8 and 16 here).

Some readers may wonder why we have used two hidden states that are placed in (redundant) opposition to each other. The reason for this is that we will use this model for more realistic simulations in the future, where hidden states encode a high precision in their circumscribed part of the visual field: this involves generating data in multiple sensory channels, with a hidden state for each channel or location. The vectors of ones and minus ones in Eq. 20 then become (radial) basis functions. Furthermore, one can easily add further hierarchical levels to make the sensory dynamics more realistic (i.e., the causes at the sensory level could excite hidden states in a lower level to produce spatiotemporally structured or moving stimuli; cf, Nobre et al., 2007). However, the basic behavior we want to illustrate here does not change. Finally, note that there is no hand-crafted gain modulation of sensory signals in the generative model. Attentional boosting of sensory signals is an emergent property of model inversion, which we now consider:

From the perspective of model inversion (mapping from sensory signals to causes) the predictive coding scheme of the previous section implies the following sort of behavior. When a cue $s_C$ is presented, it induces high-precision prediction errors, which excite the representation of the hidden cause $v_C^{(1)}$ at the higher level. This then drives up the hidden states biasing precision to the valid (right) hemi-field, which remain activated after the cue disappears. If a

subsequent (valid) target is presented, it will induce high-precision prediction errors and a consequent representation of its associated cause at the second level $v_R^{(1)}$, with a reasonably high degree of conditional confidence. Conversely, if an invalid target is presented, it faces two challenges. First, the prediction errors it elicits have low precision and will therefore exert less drive on its associated cause $v_L^{(1)}$. Furthermore, this cause has to activate its associated hidden or contextual state $x_L^{(1)}$ from much lower (negative) levels. This means that the invalid target may never actually be perceived or, if it is inferred, then it will take considerably longer before the prediction error increases its own potency (by changing the hidden causes and states). In short, invalid targets will be perceived later and with a lower degree of conditional certainty (cf. Vibell et al., 2007).

**Figure 2** shows an example of these dynamics. In this simulation, both cue and target stimuli were generated with Gaussian functions presented one-quarter and two-thirds of the way during the trial (each trial comprised sixty-four 10 ms time bins; i.e., 640 ms). When generating stimuli we suppressed all random

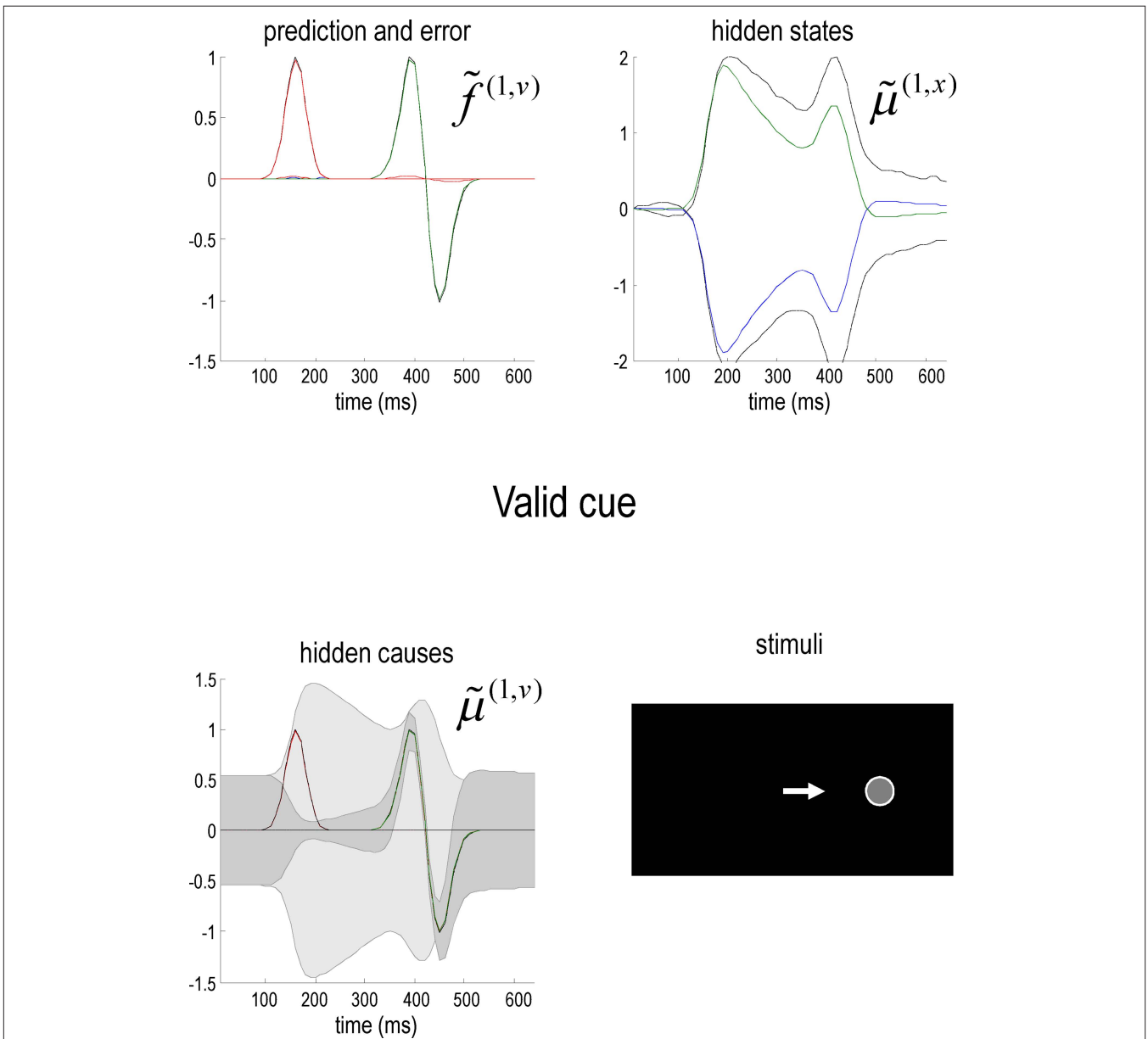



**FIGURE 2 | Simulation of the Posner task (validly cued target).** Upper left panel: the time-dependent expression of the cue and target stimuli are shown as broken gray lines, while the respective predictions are in red $s_C$ and green $s_L$ respectively. The dotted red lines show the prediction error and reflect the small amount of noise we used in these simulations. Lower left panel: the ensuing conditional expectations of the underlying hidden causes $v_R^{(1)}, v_C^{(1)}, v_L^{(1)}$ are shown below. The gray areas correspond to 90% conditional confidence tubes; this confidence reflects the estimated precision of the sensory data, which is encoded by the expectations of the hidden states in the upper right panel. The green line corresponds to a precision or attentional bias to the right $x_R^{(1)}$ and the blue line to the left $x_L^{(1)}$. They gray lines are the true precisions. Lower right panel: this insert indicates the sort of stimuli that would be generated by these hidden causes.

fluctuations, using a log-precision of eight. The cue was a simple bump function with a duration (standard deviation) of about 45 ms. The target was a (biphasic) time derivative of a Gaussian bump function with a duration of about 90 ms. The cue and target stimuli are shown as broken gray lines in **Figure 2**. These are nearly underneath the respective predictions in red $s_C$ and green $s_L$ respectively. The dotted red lines show the prediction error and reflect the small amount of noise we used in the simulations (a log-precision of eight; see Eq. 20). The ensuing conditional expectations of the underlying causal states $v_R^{(1)}, v_C^{(1)}, v_L^{(1)}$ are shown below (lower left). The gray areas around the expectations correspond to 90% conditional confidence regions (referred to as tubes). Note that the conditional tube for the cued target (green line $v_R^{(1)}$) is relatively tight because the precision of the prediction errors associated with this location is high. Conversely, the tube for the non-target $v_L^{(1)}$ is somewhat wider but correctly centered on an expectation of zero. The precisions are determined by the hidden states shown on the upper right. The green line corresponds to a precision or attentional bias to the right $x_R^{(1)}$ and the blue line to the left $x_L^{(1)}$. It can be seen that by the time the target arrives, the log-precision is about four (see Eq. 20). This is substantially greater than the prior precision on the hidden causes (we set this to a log-precision of zero). Therefore, the representation of the hidden cause (target) is driven primarily by sensory input. The insert on the lower level provides a schematic indicating the sort of stimuli that would be generated by these hidden causes. Now, compare these results with the equivalent responses to an identical stimulus but presented in the other hemi-field.

**Figure 3** uses the same format as **Figure 2** to show the responses to an invalid target (blue lines) presented on the right. It can be seen here that the predictions on this sensory channel are substantially less than the true value (compare the blue and dotted gray lines) with a consequent and marked expression of prediction error (dotted red line). As anticipated, the conditional confidence regions for the conditional expectation of this invalid target (lower left panel) are now much larger; with the 90% confidence tube always containing the value zero. The reason for this is that this invalid cue has failed to reverse the attentional context and is still operating under low levels of precision. This is reflected by the hidden states. In comparison with the previous figure, the attentional bias (difference between the right and left hidden states) has been subverted by the invalid cue but has not been reversed (the dotted gray lines show the true values of these hidden or contextual states).

The result of this asymmetry between valid and invalid cueing means that responses to valid targets are of higher amplitude and have much tighter confidence tubes, in relation to invalid targets. This is shown on the lower right panel of **Figure 3**, where one can compare the conditional estimates of the valid (green) and the invalid (blue) cause. Note that these profoundly different responses were elicited using exactly the same stimulus amplitude, after the cue had disappeared. This means that the difference is attributable only to the context (hidden states) that is instantiated by the endogenous cue. This is the basic phenomenon that we wanted to demonstrate, namely attentional bias in the ability of stimuli to capture attentional resources, where these resources correspond to the precision of sensory samples encoded by inferred hidden states or context. The reason that precision behaves like a resource is that the generative model contains prior beliefs that log-precision is redistributed over sensory channels in a context-sensitive fashion but is conserved over all channels.

### *The psychophysics of the Posner paradigm*

The difference in the confidence tubes between valid and invalidly cued targets (**Figure 3**; lower right) can be usefully interpreted in relation to behavior (cf. Gordon, 1967). At each point in peristimulus time, the conditional density implicit in the conditional mean and precision can be used to compute the conditional probability that the target intensity is present. This provides the posterior probability $p(v_i^{(1)} > 0 \,|\, \tilde{s}, m) : i \in R, L$ of the presence of a target as a function of peristimulus time shown in **Figure 4** (left panel). These results can be interpreted in terms of a speed-accuracy trade-off. For example, one can identify the amount of peristimulus time required to accumulate sufficient evidence for a fixed level of accuracy, as determined by the posterior conditional confidence. Note how the conditional probability of the target being present shrinks toward chance (50%) levels, under invalid cueing. In this example, 80% conditional confidence for valid targets (solid line) is attained at about 20 ms before the same accuracy for invalid targets (broken line). This translates into a reaction time advantage for valid targets of about 20 ms.

**Figure 4** (right panel) shows the time taken to reach 80% conditional confidence after the onset of invalid, neutral and valid cues (we simulated these reaction times with $\gamma = 0.8$). Neutral cues are modeled by reducing $\gamma = 0.2$ and removing any spatial bias afforded by the hidden states (by only using valid targets). This produces a temporal facilitation (temporal alerting effect) but without spatial specificity. The reaction time advantage with valid cues and the cost with invalid cues can be seen clearly. The reaction time to neutrally cued stimuli lies between these values. Note the asymmetry between the reaction time benefit of a valid cue and the cost of an invalid cue; this asymmetry is evident in behavioral data and is an emergent property of the non-linearities inherent in this Bayes-optimal scheme.

Recall that the time course of the Posner effect depends on the slowly-decaying hidden states encoding precision (with a time constant of 32 in Eq. 20). This reflects a formal prior that changes in precision show a temporal persistence at any location in visual space. This sort of prior means that attentional biasing will persist but decay monotonically following a cue. This effect manifests in reaction times as a slow decay of benefits and costs with valid and invalid cures respectively. **Figure 5** (left panel) shows the difference in reaction times following the three types of cue for various asynchronies between cue and stimulus onset (the "foreperiod"). The small benefit seen for neutral cues is due to a temporal alerting effect and reflects an increase in precision with no spatial bias (i.e., a small increase in precision at both locations). Note that cue-dependent effects emerge over 200 ms, during which time conditional expectations accumulate evidence (see **Figure 2**; upper right panel). The ensuing profiles of reaction times are pleasing similar to empirical observations. The right panel of **Figure 5** shows the corresponding behavioral results reported in Posner et al. (1978). Note again that the asymmetry in costs and benefits, over different foreperiods, is an emergent property of the scheme used in the simulations.

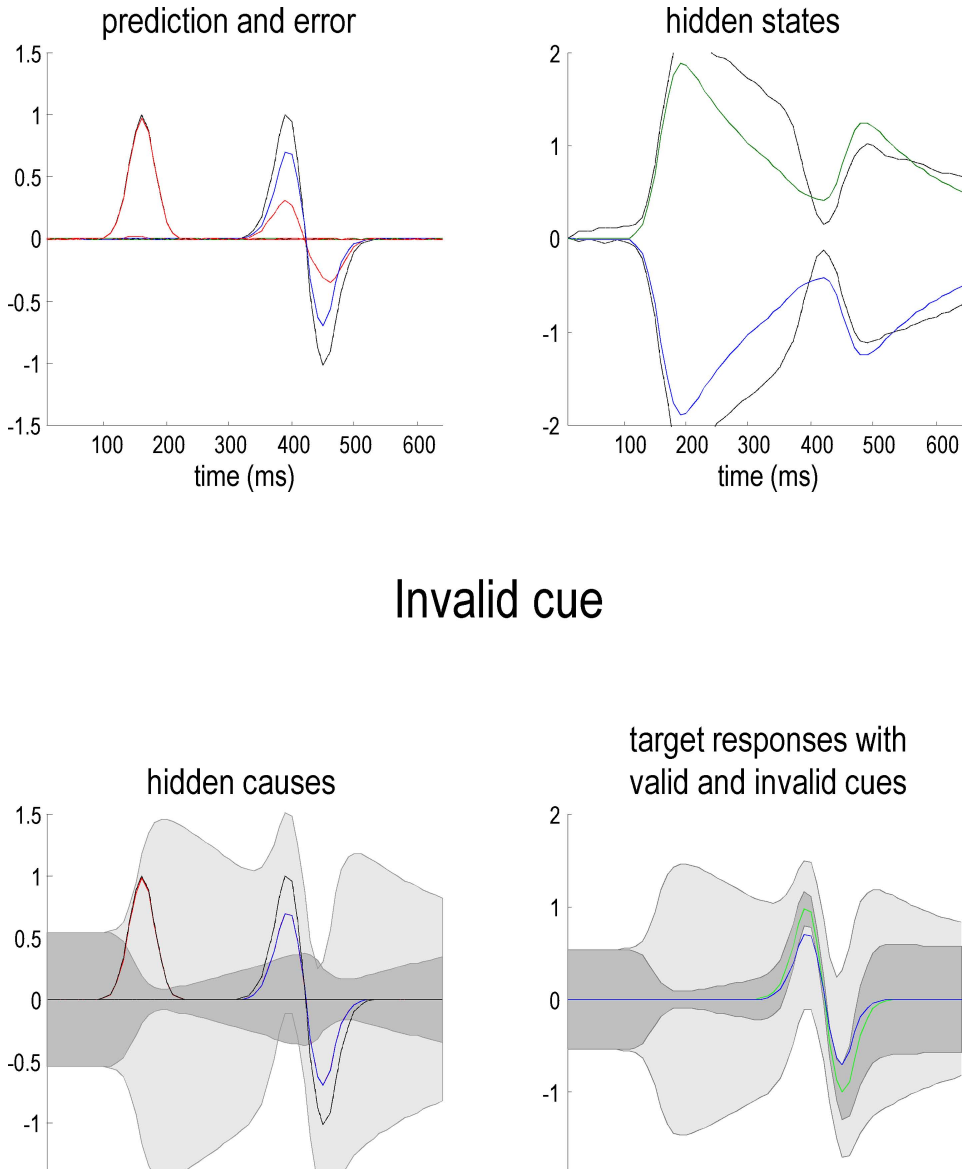
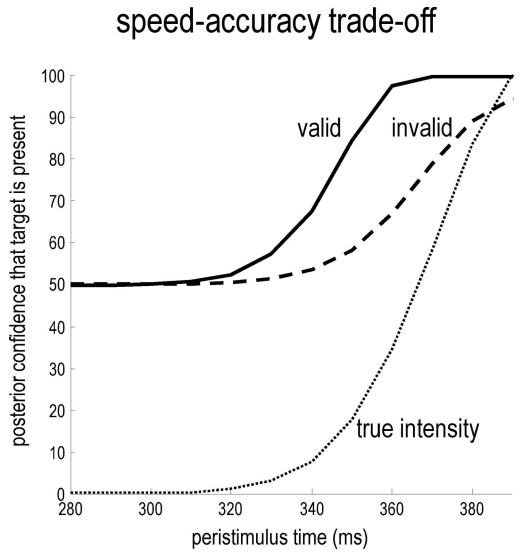



**FIGURE 3 |This figure uses the same format as Figure 2 but shows responses to an invalid target (blue line) presented on the right.** The predictions of this sensory channel are substantially less than the true value (compare the blue and dotted gray lines) with a consequent expression of prediction error (dotted red line). The conditional confidence regions for the conditional expectation of this invalid target (lower left panel) are now much larger than in the previous figure. This is shown in the lower right panel, where one can compare the conditional estimates of the valid (green; see **Figure 2**) and the invalid (blue) hidden cause, with their respective conditional confidences (gray). Note that these responses were elicited using exactly the same stimulus amplitude.

The speed-accuracy trade-off is a useful psychophysical function, which can also be interpreted in terms of relative accuracies at a fixed reaction time. In this example, at 360 ms after the cue (about 50 ms after the onset of the target), the posterior confidence about the presence of valid targets is about 98%, whereas it is only about 70% for invalid targets (**Figure 4**). The relative position and divergence of the speed-accuracy curves may provide a useful and quantitative link to empirical psychophysical data. In a subsequent paper, we will use the stimuli generated by Eq. 20 to elicit speed-accuracy performances from real subjects and use this performance to optimize the model and its parameters.

### *The electrophysiology of the Posner paradigm*

In what follows, we attempt to explain the well characterized electrophysiological correlates of the Posner paradigm using simulated event-related activity evoked by target stimuli. Spatial cueing effects are expressed in the modulation of event-related potentials (ERPs) to valid and invalid cues (Mangun and Hillyard, 1991; Eimer, 1993; Perchet et al., 2001). Generally, one sees an increase in P1 and N1 and a decrease in posterior P3 components in validly cued trials with respect to invalid ones. In other words, there is usually a validity-related enhancement of early components and an invalidity-related enhancement of late components. The P1



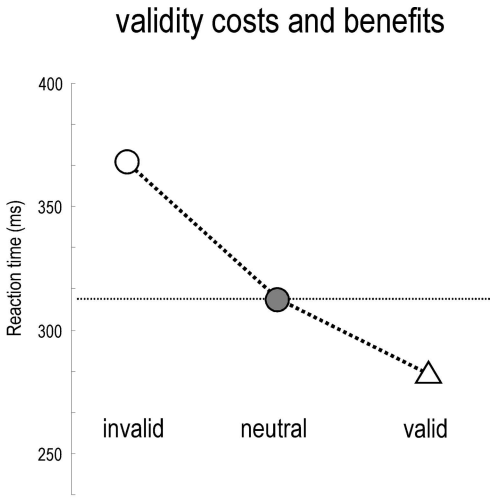
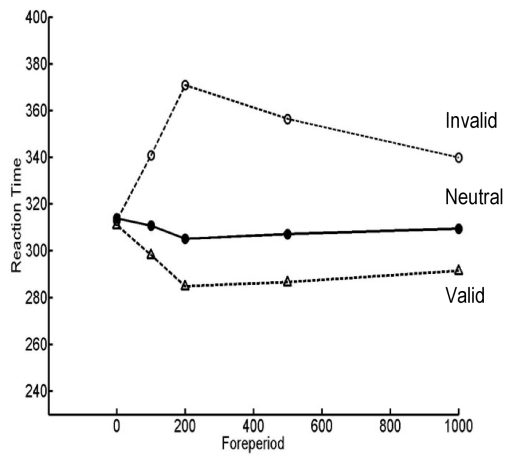




**FIGURE 4 | Left panel: the posterior probability of a target being present as a function of peristimulus time, which can be interpreted in terms of a speed-accuracy trade-off.** A reaction time can be derived from this data, as the post-stimulus time taken to achieve a fixed level of accuracy, as determined by the posterior or conditional confidence. In this example, 80% conditional confidence is attained at about 340 ms for valid targets (solid line). However, for invalid targets (broken line) the same accuracy is only attained after about 360 ms. This translates into a reaction time advantage for valid targets of about 20 ms. Right panel: this shows the reaction times for invalid, neural and valid cues, where neutral cues caused a small reduction in precision but with no spatial bias. The reaction times here are shown to within an additive constant, to better reflect empirical data (see **Figure 5**).



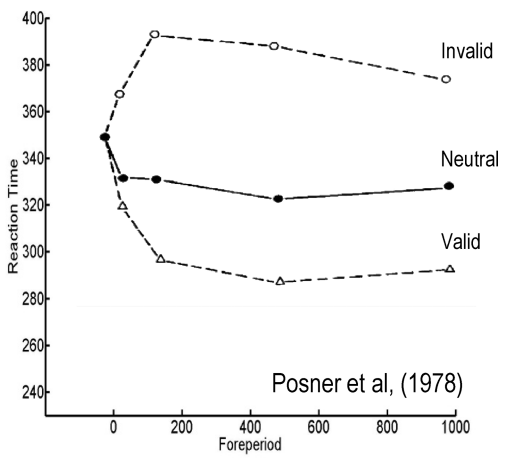




**FIGURE 5 | Left panel: simulated reaction times showing the time course of the Posner effect over different delays (foreperiod) between the onset of the cue and the target increases.** Right panel: empirical reaction time data, redrawn from Posner et al. (1978). In both the simulated and empirical data, reaction time benefit and cost increase swiftly to a maximum and then decay slowly. This reflects the quick rise and slow decay of the inferred hidden states seen in **Figures 2 and 3** (upper right panels). There is a slight reaction time benefit for neutral cues due to a temporal alerting effect. This was modeled by allowing neutral cues to induce a small rise in both the inferred hidden states. The simulated reaction times were taken as the time at which there was 80% confidence that the target was present. The simulated reaction times are shown to within an arbitrary constant (to accommodated unmodeled motor responses). The asymmetric difference between the cost for an invalid cue and the benefit for a valid cue is an emergent property of the simulations.

component is the earliest component showing attentional modulation and is considered to reflect attentional gain or the cost of attending to the wrong location (Luck et al., 1990; Mangun and Hillyard, 1991; Coull, 1998). It is well known that the amplitude of the later P3 component is inversely related to the probability of stimuli (Donchin and Coles, 1988). The anterior P3a is generally evoked by stimuli that deviate from expectations. Indeed, novel stimuli generate a higher-amplitude P3a component than deviant but repeated stimuli (Friedman et al., 2001). The P3b is a late positive component with a parietal (posterior) distribution seen in

oddball paradigms and is thought to represent a context-updating operation (Donchin and Coles, 1988; Polich, 2007). Increased P3 amplitudes during invalid trials, relative to valid trials, suggest that invalidly cued targets produce a novelty-like effect (P3a) and change the representation of probabilistic contingencies (P3b) or context (Vossel et al., 2006; Gómez et al., 2008). These hypotheses sit very comfortably with the formal scheme in this paper; in that sensory signals (prediction errors) evoked by valid targets will enjoy a selective gain, leading to enhanced early (P1 and N1) responses. Conversely, initial responses to invalid targets are suppressed until they revise the probabilistic context encoded by inferred hidden states. The prediction errors on the hidden states reflect (and drive) this revision and may contribute the later (P3) ERP components. The prediction errors on the hidden causes and states representing the content and context respectively are shown in **Figure 6**.

**Figure 6** shows synthetic ERPs based on the simulations in **Figures 2 and 3**. Here, we have made the simplifying assumption that electrophysiological signals represent the activity of superficial pyramidal cells (which we presume encode prediction error; Friston, 2008). This means we can focus on the prediction error as a proxy for electrophysiological responses. The results in the top panels of **Figure 6** show the prediction errors on the sensory signals ($\tilde{\varepsilon}^{(1,v)}$ – left panel) and hidden states ($\tilde{\varepsilon}^{(1,x)}$ – right panel). The prediction errors for valid trials are shown as dotted lines and invalid trials as solid lines. These simulations show an early suppression of prediction error for an invalidly cued target, as its low precision fails to drive its representation to its veridical level. This violation of predictions causes prediction errors on the hidden states encoding context that are expressed later in peristimulus time and drive the hidden states to revise their conditional expectations (shown in **Figures 2 and 3**). This double dissociation between validity effects in early and late peristimulus time is exactly the same as that observed by Mangun and Hillyard (1991). The empirical results of their ERP study are shown in the lower panel of **Figure 6** and are very similar to the simulations.

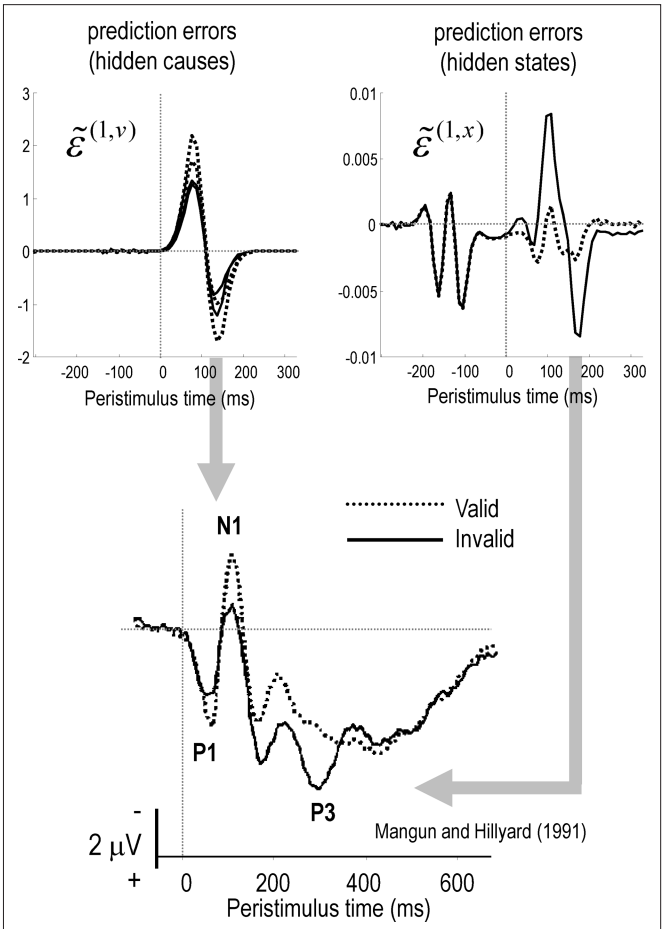



**FIGURE 6 | Simulated EEG data from our simulations (upper panels) and empirical EEG data (lower panel) from Mangun and Hillyard (1991).** The EEG traces were created from the prediction errors on the hidden causes (left) and states (right). The empirical data were recorded via EEG from the occipital cortex contralateral to the target (i.e., the cortex processing the target). The simulated data exhibits two important features of empirical studies: early in peristimulus time, stimulus-driven responses are greater for valid cues (upper left panel) relative to invalid cues. This is often attributed to a validity enhancement of early (e.g., N1) components. Conversely, later in peristimulus time, invalid responses are greater in amplitude. This can be related to novelty (and salience) responses usually associated with late waveform components (e.g., P3). In the simulations, this invalidity effect is explained simply by greater prediction errors on inferred hidden states encoding precision (upper right panel). It is these prediction errors that report a surprising or novel context, following the failure to predict invalidly cued stimuli in an optimal fashion.

### *Summary*

In summary, this section has applied the Bayes-optimal scheme established in the previous section to a minimal model of the Posner paradigm. This model provides a mechanistic if somewhat simplified explanation for some of the key psychophysical and electrophysiological aspects of the Posner effect, namely, validity effects on reaction times and the time course of these effects as stimulus onset asynchrony increases. Furthermore, the model exhibits an asymmetry in costs and benefits for invalid and valid trials respectively. Electrophysiologically, it suggests early attentional P1 enhancement can be attributed to a boosting or biasing of sensory signals (prediction errors) evoked by a target, while later P3 invalidity (cf, novelty) effects are mediated by prediction errors about the context in which targets appear.

## SIMULATING BIASED COMPETITION

In this final section, we revisit the simulations above but from the point of view of biased competition. Although the Posner paradigm considers a much greater spatial and temporal scale than the paradigms normally employed in a monkey electrophysiology, we can emulate similar phenomena by presenting both cued and non-cued targets simultaneously using the Posner model. We hoped to see a competitive interaction between stimuli that favored the cued target. Furthermore, we hoped to see responses to the unattended (invalid) target changed in the presence of an attended target. This is one of the hallmarks of biased competition and is usually attributed to lateral interactions among competing representations for stimuli, within a cell's receptive field (see Attention, Biased Competition and the Posner Paradigm). Although our model is too simple to distinguish between stimuli presented inside and outside the classical receptive field (because we do not model the spatial support of sensory channels in this paper), we can assume that targets fall

within the extraclassical receptive of field of units representing hidden causes. This is because the response to one target depends on the presence of the other, as we will see next.

**Figure 7** shows the results of presenting both stimuli simultaneously. Again the cue is in red, the valid target in green and the invalid (unattended) target in blue. It is immediately obvious that biased competition between the targets is profound, such that the response to the unattended target is about half of the response to the attended target. Furthermore, the conditional confidence about the unattended target is substantially less than that for the attended target (light and dark confidence tubes in the lower left panel). The lower right panel of **Figure 7** compares the conditional expectations and confidence intervals associated with the unattended (invalid) target presented with and without the attended (valid) target. The latter response is exactly the same as the data presented in the lower left of **Figure 3** simulating invalid cue responses. One can see that when the same stimulus is presented in conjunction with an attended target, its conditional expectation is attenuated by about 20% and the conditional confidence tubes are much wider (light with an attended distractor and dark without). In other words, the attended target has competed for attentional resources to subvert conditional confidence about the unattended target. This is despite the fact that both unattended targets were identical; they were just presented in a different context.

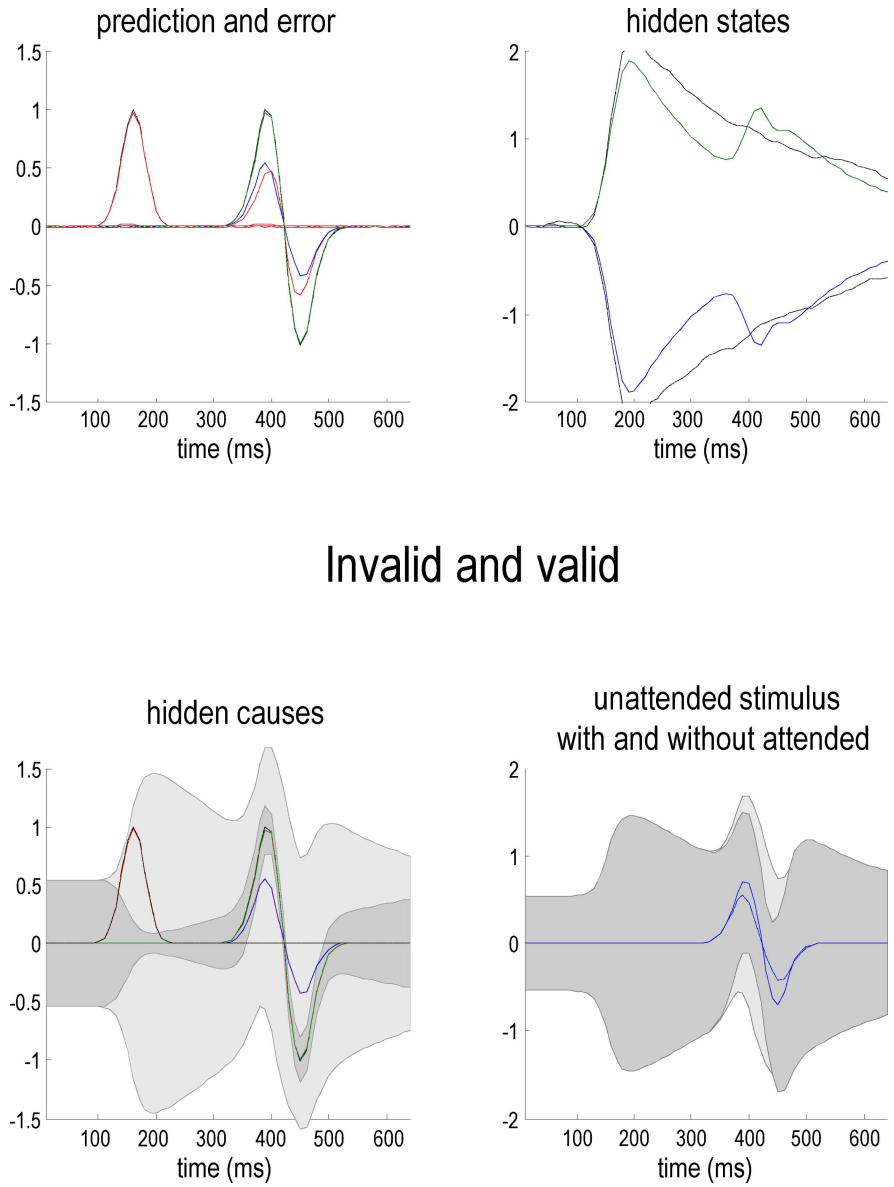



**FIGURE 7 |This figure uses the same format as Figures 2 and 3 but reports the results when both targets are presented simultaneously.** The ensuing conditional responses can be compared with the responses in **Figure 3**, when the invalidly cued target was presented alone: when the valid target is also presented, it prevents the invalid target from reversing the precision bias established by the cue; i.e., it fails to capture attention resources. The lower right panel shows the conditional expectation and confidence regions for the invalid target, with and without the valid target, to show how the responses evoked are suppressed; i.e., biased competition.

This context is encoded by the expected hidden states and explains the biased competition for resources: in contrast with the hidden states inferred with the invalid target alone (see the equivalent panel in **Figure 3**) the partial reversal of contextual representations has been precluded by the presence of the valid target. This means that the invalid cue can no longer capture precision and consequently is never able to fully express itself, through precise prediction errors, on the conditional representation of its cause. It is this effect, and only this effect, that is needed to explain biased competition. Note that we have not needed to model lateral interactions or explicit competition among representations; competition emerges naturally in a Bayes-optimal fashion through the non-linear effects of precision encoded by the units representing context, where the influence of these units is mediated by top-down or lateral projections.

The results in **Figure 7** are strikingly similar to data obtained from electrophysiological studies. **Figure 8** (upper panel) shows the conditional expectations about valid (solid line) and invalid (dashed line) targets from **Figure 7**. The lower panel shows peristimulus histograms reported in Luck et al. (1997) following simultaneous presentation of two (effective and ineffective) stimuli averaged over V4 neurons that showed a significant attention effect. The solid line reports trials in which attention was directed to the effective stimulus (cf, responses to a valid target) and the dashed line when attention was directed to the ineffective stimulus (cf, responses to an invalid target). The quantitative agreement between these simulated and empirical responses is evident and speaks quantitatively to biased competition among stimuli.

### *Summary*

Biased competition emerges naturally in Bayes-optimal schemes as a simple consequence of the fact that only one context can exist at a time. This unique aspect of context is encoded in the way that the representation of hidden states (context) modulates or distributes precision over sensory channels. Optimizing this representation leads to competition among stimuli to make the inferred context more consistent with their existence. This highlights the simplicity and usefulness of appealing to formal (Bayes-optimal) schemes, when trying to understand perception.

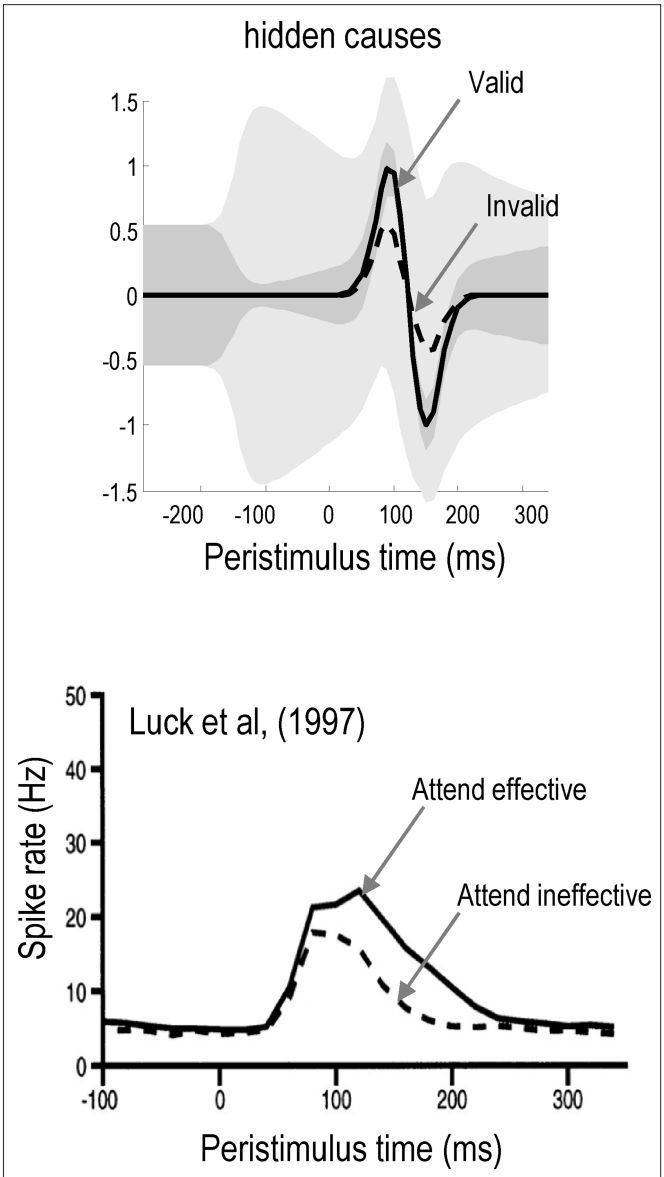



**FIGURE 8 | This figure demonstrates how generalized predictive coding reproduces some quantitative aspects of biased competition.** The simulation (upper panel) reproduces the conditional expectations in the previous figure about valid (solid line) and invalid (dashed line) targets, when presented simultaneously. These two responses resemble those reported in Luck et al. (1997). Lower panel: peristimulus histograms (over 20 ms bins) redrawn from Luck et al. (1997), following simultaneous presentation of two (effective and ineffective) stimuli averaged over 29 V4 neurons that showed a significant attention effect. The solid line reports trials in which attention was directed to the effective stimulus (cf, responses to a valid target) and the dashed line when attention was directed to the ineffective stimulus (cf, responses to an invalid target). Note that the empirical data are non-negative spike counts, whereas the simulated activity represent firing rate deviations around baseline levels.

## DISCUSSION

Our treatment of attention is one of many accounts that emphasize the role of probabilistic inference in sensory processing; including sensorimotor integration (Wolpert et al., 1995; Körding and Wolpert, 2004), sensory integration (Jacobs, 1999; Ernst and Banks, 2002; Knill and Saunders, 2003; Alais and Burr, 2004), salience and value estimation (Trommershauser et al., 2003b; Seydell et al., 2008; Whiteley and Sahani, 2008) and perception (Langer and Bulthoff, 2001; Adams et al., 2004). There have been some notable Bayesian accounts of attention using formal models (Rao, 2005; Spratling, 2008, 2010). Others have tried to define statistical measures of saliency, i.e., that which draws our attention (Duncan and Humphreys, 1989; Bruce and Tsotsos, 2009; Itti and Baldi, 2009). We now discuss these developments in the light of the more general free-energy formulation used in this paper.

The free-energy formulation is a generalization of information theoretic treatments that subsumes Bayesian schemes by assuming the brain is trying to optimize the evidence for its model of the world. This optimization involves changing the model to better account for sensory samples or by selectively sampling sensations that can be accounted for by the model (cf, perception and action). Attention can be viewed as a selective sampling of sensory data

that have high-precision (signal to noise) in relation to the model's predictions. Crucially, the model is also trying to predict precision. It is this (state-dependent) prediction we associate with attention. In short, perception, attention and action are trying to suppress free-energy, which is an upper bound on (Shannon) surprise (or the negative log-evidence for the brain's model of the world). Under some simplifying assumptions, free-energy is just the amount of prediction error, which means free-energy minimization can be cast as predictive coding. So how does this relate to other formal treatments?

## ATTENTION AND SURPRISE

Rao (2005) has introduced a compelling model of visual attention using Bayesian belief propagation. However, although consistent with Bayesian (free-energy) principles, belief propagation schemes rest on (discrete) representations of hidden causes and states, which are not compatible with the dimensionality of states in the real world (Friston, 2009). Using a more descriptive approach, Itti and Baldi (2006; 2009) proposed that many factors, which influence visual salience, can be integrated with prior expectations by calculating *Bayesian surprise (*Baldi and Itti, 2010*)*. This is (heuristically) related to another measure of saliency, proposed by Bruce and Tsotsos (2009), who suggest that visual searches are attracted to areas of the visual field which maximize the information sampled. Crucially, reducing free-energy or (Shannon) surprise increases Bayesian surprise and increases the changes in the conditional representations afforded by sensory information. This is because Bayesian surprise is the difference (Kullback–Leibler divergence) between the posterior (conditional) and prior densities on hidden causes or states. This difference reports the change in the conditional density after sampling new information. It is also called *complexity* in the Bayesian model comparison literature. Free-energy can be expressed as complexity minus accuracy (Friston, 2009). This means that minimizing (Shannon) surprise by updating conditional representations to increase accuracy (decrease prediction errors), necessarily entails an increase in complexity (Bayesian surprise). In short, increases in Bayesian surprise are necessarily associated with decreases in free-energy (they are the complexity cost of reducing prediction errors) but Bayesian surprise *per se* is not optimized in Bayes-optimal schemes.

## BIASED COMPETITION AND PREDICTIVE CODING

It is becoming increasingly clear that estimates of the precision play an important role in sensory inference. Whiteley and Sahani (2008) demonstrated very neatly that the brain possesses (and uses) a model of sensory uncertainty (i.e., precision) in decision-making, and that this model is available even under intermittent feedback, showing that is estimated internally rather than learnt. Thinking of attention as optimizing representations of uncertainty or precision resolves any potential conflict between biased competition and predictive coding schemes: Spratling (2008) noted the potential difficulty in reconciling these two theories and proposed a variant of predictive coding, in which representations compete via negative feedback. Specifically, he showed that a particular implementation of the biased competition model, in which nodes compete via inhibition that targets the inputs to a cortical region, is mathematically equivalent to linear predictive coding. This scheme relies on a rather complex neural architecture and employs non-linear modifications to prevent cells from having a negative firing rate. These modifications are interesting and relate to important theories based on divisive normalization (Heeger, 1993). This form of (divisive) predictive coding can explain a remarkable range of classical and extraclassical receptive field properties in V1 (see Spratling, 2010).

The formulation in this paper reaffirms that there is no tension between biased competition and predictive coding: it demonstrates that the characteristic behaviors of biased competition emerge naturally under predictive coding. They key thing that reconciles these two theories is to realize that predictive coding can be generalized to cover both states and precisions and that (state-dependent) precision is itself optimized. This leads to non-linear interactions among states implicit in the precision-weighting of prediction errors and provides a simple explanation for attentional gain effects. It will be interesting to relate the ensuing bias or weighting of sensory signals (prediction errors) by precision to the divisive schemes above (e.g., Heeger, 1993; Spratling, 2010).

## BASELINE SHIFTS AND PRECISION

In this paper, we have focussed on reaction time and event-related responses to targets. However, many electrophysiological and neuroimaging studies of attentional paradigms (e.g., Chelazzi et al., 1993; Chawla et al., 1999b; Kastner et al., 1999; Stokes et al., 2009) have demonstrated cue-related increases in the basal firing rate of cells, whose receptive fields correspond to the attended location. A non-invasive electrophysiological correlate of these baseline shifts is called the Contingent Negative Variation component (CNV), which follows a cue that furnishes information about subsequent (imperative) target stimuli (Walter et al., 1964; Rockstroh et al., 1982). Crucially, the cortical sources generating the CNV can include those responsible for processing the stimuli (Gómez et al., 2001). These baseline shifts may be accounted for, in the computational scheme, by the dynamics of expected hidden states, shown in the top left panels of **Figures 2 and 3**. These accumulate evidence from cues and represent changes in context that persist over time. It is possible that the activity of these representational units could contribute to the CNV or baseline shift directly. However, it is also possible that they could modulate baseline activity (caused by ambient sensory signals) in the prediction error-units they modulate. This would be consistent with baseline shifts seen with fMRI in retinotopically mapped areas of directed attention (e.g., Macaluso et al., 2003), and the reduction in non-attended areas (Smith et al., 2000). This suggests that baseline (endogenous) activity may be a quantitative proxy for the expected precision of sensory information in the corresponding sensory area (cf., Hesselmann et al., 2008). This hypothesis was tested recently: using fMRI, Hesselmann et al. (2010) linked perceptual estimates of precision with baseline increases in activity; showing that baseline activity before a (subliminal) stimulus was correlated with the accuracy of deciding if the stimulus was present (and not whether the stimulus was present or absent). This means that baseline activity may reflect the inferred precision of sensory signals. Specifically, they found that neuronal activity in sensory areas (extrastriate visual and early auditory cortex) biases perceptual decisions toward correct inference and not toward a specific percept. They conclude: "In accord with predictive coding models and the free-energy principle, this observation suggests

that cortical activity in sensory brain areas reflects the precision of prediction errors and not just the sensory evidence or prediction errors *per se*."

The neurobiological (resp. computational) mechanisms that might underlie these effects tie several strands of evidence together rather neatly: as noted in the introduction the most plausible candidate for modulating activity-dependent (resp. state-dependent) synaptic gain (resp. precision) are fast synchronous interactions associated with attention (Börgers et al., 2005; Womelsdorf and Fries, 2006; Fries et al., 2008; Zeitler et al., 2008). The associated increase in synchronous gain is necessarily accompanied by increased levels of population activity that are both supported by and support synchrony (Chawla et al., 1999a; Salinas and Sejnowski, 2001). These are manifest as high frequency (gamma) activity and elevated fMRI signals seen in attentional paradigms (Gruber et al., 1999; Sokolov et al., 1999; Steinmetz et al., 2000; Bichot et al., 2005; Pavlova et al., 2006; Vidal et al., 2006; Fries et al., 2008).

### ATTENTION, GAIN AND LEARNING

In closing, we pre-empt a potentially interesting argument about the specificity of gain mechanisms and attention. The idea pursued in this paper is that attention corresponds to inference about uncertainty or precision and that this inference is encoded by dynamic changes in post-synaptic gain. However, non-linear (gain) post-synaptic responses are ubiquitous in the brain; so what is special about the non-linearities associated with attention? We suggest that attention is mediated by gain modulation of prediction error-units (forward or bottom-up information) in contradistinction to gain modulation of prediction units (backward, lateral or top-down information). In other words, non-linearities in the brain's generative model encoding context-sensitive expectations are distinct from non-linearities (gain) entailed by optimal recognition. The distinction may seem subtle but there is a fundamental difference between inferring the context-dependent contingencies and causes of sensations (perception) and their precision (attention). In this sense, there is an implicit distinction between inferring what is relevant for a task (as in classical attention tasks like dichotic listening) and the uncertainty about what is relevant. We have side-stepped this issue with the Posner task, because all cues are task relevant.

There is a final distinction that may be mechanistically important: we have focussed on activity-dependent optimization of gain but have not considered the (slower) learning of how and when this optimization should be deployed. For example, the latency of saccades to a target can be reduced if the target is more likely to appear on one side – and this relationship can be learned in as few as 150 trials (Carpenter and Williams, 1995; Anderson and Carpenter, 2006; Brodersen et al., 2008). This sort of learning corresponds to the optimization of the precision parameters in Eq. 19 and may involve modulatory neurotransmitters. We will pursue this elsewhere and try to relate this learning to the psychopharmacology of attention and related theories about uncertainty (e.g., Yu and Dayan, 2005).

### CONCLUSION

In this paper, we have tried to establish the face validity of optimizing the precision of sensory signals as an explanation for attention in perceptual inference. We started with an established scheme for perception based upon optimizing a free-energy bound on surprise or the log-evidence for a model of the world. Minimizing this bound, using gradient descent, furnishes recognition dynamics that are formally equivalent to evidence accumulation schemes. Under some simplifying assumptions, the free-energy reduces to prediction error and the scheme can be regarded as generalized predictive coding. The key thing that we have tried to demonstrate is that all the quantities required for making an inference have to be optimized. This includes the precisions that encode uncertainty or the amplitude of random fluctuations generating sensory information. By casting attention as inferring precision, we can explain several perspectives on attentional processing that fit comfortably with their putative neurobiological mechanisms. Furthermore, by considering how states of the world influence uncertainty, one arrives at a plausible architecture, in which conditional expectations about states modulate their own precision. This leads naturally to competition and other non-linear phenomena during perception. We have tried to illustrate these ideas in the context of a classical paradigm (the Posner paradigm) and relate the ensuing behavior to biased competition evident in electrophysiological responses recorded from awake, behaving monkeys. In future work, we will use the theoretical framework in this paper to model empirical psychophysical and electrophysiological data and pursue this hypothesis using formal model comparison.

## ACKNOWLEDGMENTS

The Wellcome Trust funded this work. We would like to thank Marcia Bennett for helping prepare this manuscript. We are very grateful to Jon Driver and Kia Nobre for invaluable help in formulating these ideas.

## GLOSSARY OF TERMS

**Bayesian surprise**: A measure of salience based on the (Kullback–Leibler) divergence between the recognition and prior densities. It measures the information in the data that can be recognized.

**Conditional density**: Conditional density or posterior density is the probability distribution of causes or model parameters, given some data; i.e., a probabilistic mapping from observed data (consequences) to causes.

**(Kullback–Leibler) Divergence**: Information divergence, information gain or relative entropy is a non-commutative measure of the difference between two probability distributions.

**Empirical prior**: Priors that are induced by hierarchical models; they provide constraints on the recognition density is the usual way but depend on the data.

**Entropy**: The average surprise of outcomes sampled from a probability distribution or density. A density with low entropy means, on average, the outcome is relatively predictable (certain).

**Free-energy**: An information theory measure that bounds (is greater than) the surprise on sampling some data, given a generative model.

**Generalized coordinates**: Generalized coordinates of motion cover the value of a variable, its motion, acceleration, jerk and higher orders of motion. A point in generalized coordinates corresponds to a path or trajectory over time.

**Generative model**: Generative model or forward model is a probabilistic mapping from causes to observed consequences (data). It is usually specified in terms of the likelihood of getting some data given their causes (parameters of a model) and priors on the parameters.

**Gradient descent**: An optimization scheme that finds a minimum of a function by changing its arguments in proportion to the negative of the gradient of the function at the current value.

**Helmholtz (inference) machine**: Device or scheme that uses a generative model to furnish a recognition density. They learn hidden structure in data by optimizing the parameters of generative models.

**Precision**: (In general statistical usage) means the inverse variance or dispersion of a random variable. The precision matrix of several variables is also called a concentration matrix. It quantifies the degree of certainty about the variables.

**Prior**: The probability distribution or density on the causes of data that encode beliefs about those causes prior to observing the data.

**Recognition density**: Recognition density or approximating conditional density is a probability distribution over the causes of data. It is the product of (approximate) inference or inverting a generative model. It is sometimes referred to as a proposal or ensemble density in machine learning.

**Surprise**: Surprisal or self-information is the negative log-probability of an outcome. An improbable outcome is therefore surprising.

**Stochastic**: The successive states of stochastic processes are governed by random effects.

**Uncertainty**: A measure of unpredictability or expected surprise (cf, entropy). The uncertainly about a variable is often quantified with its variance (inverse precision).

## REFERENCES


Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nat. Neurosci.* 7, 1057–1058.

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262.

Anderson, A. J., and Carpenter, R. H. S. (2006). Changes in expectation consequent on experience, modeled by a simple, forgetful neural circuit. *J. Vis.* 6, 822–835.

Baldi, P., and Itti, L. (2010). Of bits and wows: a Bayesian theory of surprise with applications to attention. *Neural Netw.* 23, 649–666.

Ballard, D. H., Hinton, G. E., and Sejnowski, T. J. (1983). Parallel visual computation. *Nature* 306, 21–26.

Bartolomeo, P., Caroline Decaix, C., and Siéroff, E. (2007). The phenomenology of endogenous orienting. *Conscious. Cogn.* 16, 144–161.

Bauer, F., Cheadle, S. W., Parton, A., Müller, H. J., and Usher, M. (2009). Gamma flicker triggers attentional selection without awareness. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1666–1671.

Beck, D. M., and Kastner, S. (2005). Stimulus context modulates competition in human extrastriate cortex. *Nat. Neurosci.* 8, 1110–1116.

Bichot, N. P., Rossi, A. F., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308, 529–534.

Börgers, C., Epstein, S., and Kopell, N. J. (2005). Background gamma rhythmicity and attention in cortical local circuits: a computational study. *Proc. Natl. Acad. Sci. U.S.A* 102, 7002–7007.

Broadbent, D. E. (1952a). Listening to one of two synchronous messages. *J. Exp. Psychol.* 44, 51–55.

Broadbent, D. E. (1952b). Failures of attention in selective listening. *J. Exp. Psychol.* 44, 428–433.

Broadbent, D. E. (1958). *Perception and Communication*. New York: Pergamon Press.

Brodersen, K. H., Penny, W. D., Harrison, L. M., Daunizeau, J., Ruff, C. C., Duzel, E., Friston, K. J., and Stephan, K. E. (2008). Integrated Bayesian models of learning and decision making for saccadic eye movements. *Neural Netw.* 21, 1247–1260.

Bruce, N. D., and Tsotsos, J. K. (2009). Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* 9, 1–24.

Buhl, E. H., Tamás, G., and Fisahn, A. (1998). Cholinergic activation and tonic excitation induce persistent gamma oscillations in mouse somatosensory cortex in vitro. *J. Physiol.* 513, 117–126.

Buia, C., and Tiesinga, P. (2006). Attentional modulation of firing rate and synchrony in a model cortical network. *J. Comput. Neurosci.* 20, 247–264.

Carpenter, R. H. S., and Williams, M. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377, 59–62.

Cave, C. R., and Bichot, N. P. (1999). Visuospatial attention: beyond a spotlight model. *Psychon. Bull. Rev.* 6, 204–223.

Chawla, D., Lumer, E. D., and Friston, K. J. (1999a). The relationship between synchronization among neuronal populations and their mean activity levels. *Neural Comput.* 11, 1389–1411.

Chawla, D., Rees, G., and Friston, K. J. (1999b). The physiological basis of attentional modulation in extrastriate visual areas. *Nat. Neurosci.* 2, 671–676.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature* 363, 345–347.

Chikkerur, S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a Bayesian inference theory of attention. *Vision Res.* 50, 2223–2247.

Clark, C. R., Geffen, G. M., and Geffen, L. B. (1989). Catecholamines and the covert orientation of attention in humans. *Neuropsychologia* 27, 131–139.

Coull, J. T. (1998). Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology. *Prog. Neurobiol.* 55, 343–361.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 81, 4586–4590.

Dalley, J. W., McGaughy, J., O'Connell, M. T., Cardinal, R. N., Levita, L., and Robbins, T. W. (2001). Distinct changes in cortical acetylcholine and noradrenaline efflux during contingent and noncontingent performance of a visual attentional task. *J. Neurosci.* 21, 4908–4914.

Daunizeau, J., David, O., and Stephan, K. E. (2009). Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage* [Epub ahead of print].

Davidson, M. C., and Marrocco, R. T. (2000). Local infusion of scopolamine into intraparietal cortex slows covert orienting in rhesus monkeys. *J. Neurophysiol.* 83, 1536–1549.

Dayan, P., Hinton, G. E., and Neal, R. M. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904.

Decaix, C., Siéroff, E., and Bartolomeo, P. (2002). How voluntary is 'voluntary' orienting of attention? *Cortex* 38, 841–845.

Deco, G., and Rolls, E. T. (2005). Neurodynamics of biased competition and cooperation for attention: a model with spiking neurons. *J. Neurophysiol.* 94, 295–313.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13494–13499.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. B* 353, 1245–1255.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Desimone, R., and Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Res.* 178, 363–380.

Deutsch, J. A., and Deutsch, D. (1963). Attention: some theoretical considerations. *Psychol. Rev.* 70, 80–90.

Donchin, E., and Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behav Brain Sci.* 11, 355–372.

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96, 433–458.

Eckstein, M. P., Shimozaki, S. S., and Abbey, C. K. (2002). The footprints of visual attention in the Posner cueing paradigm revealed by classification images. *J. Vis.* 2, 25–45.

Eimer, M. (1993). Spatial cueing, sensory gating and selective response preparation: an ERP study on visuo-spatial orienting. *Electroencephalogr. Clin. Neurophysiol.* 88, 408–420.

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

Fernandez-Duque, D., and Posner, M. I. (1997). Relating the mechanisms of orienting and alerting. *Neuropsychologia* 35, 477–486.

Formankiewicz, M. A., and Mollon, J. D. (2009). The psychophysics of detecting binocular discrepancies of luminance. *Vision Res.* 49, 1929–1938.

Friedman, D., Cycowicz, Y. M., and Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neurosci. Biobehav. Rev.* 25, 355–373.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.* 9, 474–480.

Fries, P., Womelsdorf, T., Oostenveld, R., and Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. *J. Neurosci.* 28, 4823–4835.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301.

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1211–1221.

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010a). Action and behavior: a free-energy formulation. *Biol. Cybern.* 102, 227–260.

Friston, K., Stephan, K. E., Li, B., and Daunizeau, J. (2010b). Generalised filtering. *Math. Probl. Eng.* 2010, Article ID 621670.

Fründ, I., Busch, N. A., Schadow, J., Körner, U., and Herrmann, C. S. (2007). From perception to action: phase-locked gamma oscillations correlate with reaction times in a speeded response task. *BMC Neurosci.* 8, 27. doi: 10.1186/1471-2202-8-27.

Gómez, C. M., Delinte, A., Vaquero, E., Cardoso, M. J., Vazquez, M., Crommelynck, M., and Roucoux, A. (2001). Current source density analysis of CNV during temporal gap paradigm. *Brain Topogr.* 13, 149–159.

Gómez, C. M., Flores, A., Digiacomo, M. R., Ledesma, A., and González-Rosa, J. (2008). P3a and P3b components associated to the neurocognitive evaluation of invalidly cued targets. *Neurosci. Lett.* 430, 181–185.

Gordon, I. E. (1967). Stimulus probability and simple reaction time. *Nature* 215, 895–896.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. R. Soc. Lond. B* 171, 179–196.

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

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35, 96–111.

Gruber, T., Müller, M. M., Keil, A., and Elbert, T. (1999). Selective visual-spatial attention alters induced gamma band responses in the human EEG. *Clin. Neurophysiol.* 110, 2074–2085.

Hasselmo, M. E., and Giocomo, L. M. (2006). Cholinergic modulation of cortical function. *J. Mol. Neurosci.* 30, 133–135.

Heeger, D. J. (1993). Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *J. Neurophysiol.* 70, 1885–1898.

Herrero, J. L., Roberts, M. J., Delicato, L. S., Gieselmann, M. A., Dayan, P., and Thiele, A. (2008). Acetylcholine contributes through muscarinic receptors to attentional modulation in V1. *Nature* 454, 1110–1114.

Hesselmann, G., Kell, C. A., and Kleinschmidt, A. (2008). Ongoing activity fluctuations in hMT+ bias the perception of coherent visual motion. *J. Neurosci.* 28, 14481–14485.

Hesselmann, G., Sadaghiani, S., Friston, K. J., and Kleinschmidt, A. (2010). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE* 5, e9926. doi: 10.1371/journal.pone.0009926.

Hirayama, J., Yoshimoto, J., and Ishii, S. (2004). Bayesian representation learning in the cortex regulated by acetylcholine. *Neural Netw.* 17, 1391–1400.

Hommel, B., Pratt, J., Colzato, L., and Godijn, R. (2001). Symbolic control of visual attention. *Psychol. Sci.* 12, 360–365.

Itti, L., and Baldi, P. (2006). "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems,* Vol. 18, eds Y. Weiss, B. Schölkopf, and J. Platt (Cambridge, MA: MIT Press), 1–8.

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306.

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

James, W. (1890). *The Principles of Psychology*, Vol. 1. New York: Henry Holt, 403–404.

Jaramillo, S., and Pearlmutter, B. A. (2007). Optimal coding predicts attentional modulation of activity in neural systems. *Neural Comput.* 19, 1295–1312.

Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science* 282, 108–111.

Kastner, S., De Weerd, P., Pinsk, M. A., Elizondo, M. I., Desimone, R., and Ungerleider, L. G. (2001). Modulation of sensory suppression: implications for receptive fields sizes in the human visual cortex. *J. Neurophysiol.* 86, 1398–1411.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* 22, 751–761.

Kiebel, S. J., von Kriegstein, K., Daunizeau, J., and Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Comput. Biol.* 5, e1000464. doi: 10.1371/journal.pcbi.1000464.

Knill, D. C., and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.* 43, 2539–2558.

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

Langer, M. S., and Bulthoff, H. H. (2001). A prior for global convexity in local shape-from-shading. *Perception* 30, 403–410.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 451–468.

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Luck, S. J., Heinze, H. J., Mangun, G. R., and Hillyard, S. A. (1990). Visual event-related potentials index focused attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. *Electroencephalogr. Clin. Neurophysiol.* 75, 528–542.

Macaluso, E., Eimer, M., Frith, C. D., and Driver, J. (2003). Preparatory states in crossmodal spatial attention: spatial specificity and possible control mechanisms. *Exp. Brain Res.* 149, 62–74.

Mangun, G. R., and Hillyard, S. A. (1991). Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 1057–1074.

Marrocco, R. T., Witte, E. A., and Davidson, M. C. (1994). Arousal systems. *Curr. Opin. Neurobiol.* 4, 166–170.

Maunsell, J. H., and Treue, S. (2006). Feature-based attention in visual cortex. *Trends Neurosci.* 29, 317–322.

Mazurek, M. E., Roitman, J. D., Ditterich, J., and Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cereb. Cortex.* 13, 1257–1269.

McCormick, D. A., and Prince, D. A. (1985). Two types of muscarinic response to acetylcholine in mammalian cortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6344–6348.

McCormick, D. A., and Prince, D. A. (1986). Mechanisms of action of acetylcholine in the guinea-pig cerebral cortex in vitro. *J. Physiol.* 375, 169–194.

Moran, J., and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784.

Moray, N. (1959). Attention in dichotic listening: affective cues and the influence of instructions. *Q. J. Exp. Psychol.* 11, 56–60.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* 66, 241–251.

Nobre, A., Correa, A., and Coull, J. (2007). The hazards of time. *Curr. Opin. Neurobiol.* 17, 465–470.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., and Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nat. Neurosci.* 5, 1203–1209.

Ozaki, T. (1992). A bridge between nonlinear time-series models and nonlinear stochastic dynamical systems: a local linearization approach. *Stat. Sin.* 2, 113–135.

Parikh, V., Kozak, R., Martinez, V., and Sarter, M. (2007). Prefrontal acetylcholine release controls cue detection on multiple timescales. *Neuron* 56, 141–154.

Pavlova, M., Birbaumer, N., and Sokolov, A. (2006). Attentional modulation of cortical neuromagnetic gamma response to biological movement. *Cereb. Cortex* 16, 321–327.

Perchet, C., Revol, O., Fourneret, P., Mauguière, F., and Garcia-Larrea, L. (2001). Attention shifts and anticipatory mechanisms in hyperactive children: an ERP study using the Posner paradigm. *Biol. Psychiatry* 50, 44–57.

Pessoa, L., Kastner, S., and Ungerleider, L. G. (2002). Attentional control of the processing of neutral and emotional stimuli. *Cogn. Brain Res.* 15, 31–45.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148.

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25.

Posner, M. I. (2008). Measuring alertness. *Ann. N. Y. Acad. Sci.* 1129, 193–199.

Posner, M. I., and Cohen, Y. (1984). "Components of visual orienting," in *Attention and Performance, Vol. 10*, eds H. Bouma and D. G. Bouwhuis (Hillsdale, NJ: Erlbaum), 531–556.

Posner, M. I., Nissen, M. J., and Ogden, W. C. (1978). "Attended and unattended processing modes: the role of set for spatial location," in *Modes of Perceiving and Processing Information,* eds H. L.

Pick and N. J. Saltzman (Hillsdale, NJ: Lawrence Erlbaum Associates).
Rafal, R. D., Calabresi, P. A., Brennan, C. W., and Sciolto, T. K. (1989). Saccade preparation inhibits reorienting to recently attended locations. *J. Exp. Psychol Hum. Percept. Perform.* 15, 673–685.
Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16, 1843–1848.
Rao, R. P., and Ballard, D. H. (1998). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nat. Neurosci.* 2, 79–87.
Recanzone, G. H., Wurtz, R. H., and Schwarz, U. (1997). Responses of MT and MST neurons to one and two moving objects in the receptive field. *J. Neurophysiol.* 78, 2904–2915.
Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* 19, 1736–1753.
Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185.
Risko, E. F., and Stolz, J. A. (2010). The proportion valid effect in covert orienting: strategic control or implicit learning? *Conscious. Cogn.* 19, 432–442.
Rockstroh, B., Elbert, T., Birbaumer, N., and Lutzenberger, W. (1982). *Slow Brain Potentials and Behavior*. Baltimore-Munich: Urban and Schwarzenberg.
Rolls, E. T., and Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Exp. Brain Res.* 103, 409–420.
Salinas, E., and Sejnowski, T. J. (2001). Gain modulation in the central nervous system: where behavior, neurophysiology, and computation meet. *Neuroscientist* 7, 430–440.
Sara, S. J. (1998). Learning by neurones: role of attention, reinforcement and behaviour. *C. R. Acad. Sci. III, Sci. Vie* 321, 193–198.
Schroeder, C. E., Mehta, A. D., and Foxe, J. J. (2001). Determinants and mechanisms of attentional modulation of neural processing. *Front. Biosci.* 6, D672–D684.
Seydell, A., McCann, B. C., Trommershauser, J., and Knill, D. C. (2008). Learning stochastic reward distributions in a speeded pointing task. *J. Neurosci.* 28, 4356–4367.
Shulman, G. L., Remington, R. W., and McLean, J. P. (1979). Moving attention through visual space. *J. Exp. Psychol. Hum. Percept. Perform.* 3, 522–526.
Smith, A. T., Singh, K. D., and Greenlee, M. W. (2000). Attentional suppression of activity in the human visual cortex. *Neuroreport* 11, 271–277.
Sokolov, A., Lutzenberger, W., Pavlova, M., Preissl, H., Braun, C., and Birbaumer, N. (1999). Gamma-band MEG activity to coherent motion depends on task-driven attention. *Neuroreport* 10, 1997–2000.
Spratling, M. W. (2008). Predictive-coding as a model of biased competition in visual attention. *Vision Res.* 48, 1391–1408.
Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* 30, 3531–3543.
Steinmetz, P. N., Roy, A., Fitzgerald, P. J., Hsiao, S. S., Johnson, K. O., and Niebur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* 404, 187–190.
Stokes, M., Thompson, R., Nobre, A. C., and Duncan, J. (2009). Shape-specific preparatory activity mediates attention to targets in human visual cortex. *Proc. Natl. Acad. Sci. U S A.* 106, 19569–19574.
Theeuwes, J. (1991). Exogenous and endogenous control of attention: the effect of visual onsets and offsets. *Percept. Psychophys.* 49, 83–90.
Treisman, A. (1964). Verbal cues, language and meaning in selective attention. *Am. J. Psychol.* 77, 206–209.
Treisman, A. (1998). Feature binding, attention and object perception. *Philos. Trans. R. Soc.Lond. B* 353, 1295–1306.
Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136.
Treisman, A., and Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cogn. Psychol.* 14, 107–141.
Treue, S., and Maunsell, J. H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* 382, 539–541.
Trommershauser, J., Maloney, L. T., and Landy, M. S. (2003b). Statistical decision theory and the selection of rapid, goal-directed movements. *J. Opt. Soc. Am. A* 20, 1419–1433.
Vibell, J., Klinge, C., Zampini, M., Spence, C., and Nobre, A. C. (2007). Temporal order is coded temporally in the brain: early event-related potential latency shifts underlying prior entry in a cross-modal temporal order judgment task. *J. Cogn. Neurosci.* 19, 109–120.
Vidal, J. R., Chaumon, M., O'Regan, J. K., and Tallon-Baudry, C. (2006). Visual grouping and the focusing of attention induce gamma-band oscillations at different frequencies in human magnetoencephalogram signals. *J. Cogn. Neurosci.* 18, 1850–1862.
Vossel, S., Thiel, C. M., and Fink, G. R. (2006). Cue validity modulates the neural correlates of covert endogenous orienting of attention in parietal and frontal cortex. *Neuroimage* 32, 1257–1264.
Vossel, S., Thiel, C. M., and Fink, G. R. (2008). Behavioral and neural effects of nicotine on visuospatial attentional reorienting in non-smoking subjects. *Neuropsychopharmacology* 33, 731–738.
Voytko, M. L., Olton, D. S., Richardson, R. T., Gorman, L. K., Tobin, J. R., and Price, D. L. (1994). Basal forebrain lesions in monkeys disrupt attention but not learning and memory. *J. Neurosci.* 14, 167–186.
Walter, W. G., Cooper, R., Aldridge, W. J., and McCallum, W. C. (1964). Contingent negative variation: an electrophysiological sign of sensoriomotor association and expectancy in the human brain. *Nature* 203, 380–384.
Whiteley, L., and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *J. Vis.* 8, 2.1–15.
Witte, E. A., Davidson, M. C., and Marrocco, R. T. (1997). Effects of altering brain cholinergic activity on covert orienting of attention: comparison of monkey and human performance. *Psychopharmacology (Berl.)* 132, 324–334.
Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882.
Womelsdorf, T., and Fries, P. (2006). Neuronal coherence during selective attentional processing and sensory-motor integration. *J. Physiol. (Paris)* 100, 182–193.
Womelsdorf, T., Fries, P., Mitra, P. P., and Desimone, R. (2006). Gamma-band synchronization in visual cortex predicts speed of change detection. *Nature* 439, 733–736.
Wonnacott, S. (1997). Presynaptic nicotinic ACh receptors. *Trends Neurosci.* 20, 92–98.
Xiang, Z., Huguenard, J. R., and Prince, D. A. (1998). Cholinergic switching within neocortical inhibitory networks. *Science* 28, 985–988.
Yantis, S., and Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *J. Exp. Psychol Hum. Percept. Perform.* 10, 601–621.
Yu, A. J., and Dayan, P. (2005). Uncertainty, neuromodulation and attention. *Neuron* 46, 681–692.
Zeitler, M., Fries, P., and Gielen, S. (2008). Biased competition through variations in amplitude of gamma-oscillations. *J. Comput. Neurosci.* 25, 89–107.
Zilles, K., Palomero-Gallagher, N., and Schleicher, A. (2004). Transmitter receptors and functional anatomy of the cerebral cortex. *J. Anat.* 205, 417–432.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 04 August 2010; paper pending published: 24 September 2010; accepted: 18 October 2010; published online: 02 December 2010.*
*Citation: Feldman H and Friston KJ (2010) Attention, uncertainty, and free-energy. Front. Hum. Neurosci. **4**:215. doi: 10.3389/fnhum.2010.00215*

## APPENDIX

### INTEGRATING THE RECOGNITION DYNAMICS (GENERALIZED FILTERING)

Generalized filtering (Friston et al., 2010b) involves integrating the ordinary differential Eqs 7 and 8 to optimize the conditional means. We can simplify the numerics for hierarchical dynamic models by first collapsing over the hierarchy, then over generalized motion and finally over hidden causes and states:

$$\mu^{(v)} = \begin{bmatrix} \mu^{(1,v)} \\ \mu^{(2,v)} \\ \vdots \end{bmatrix} \quad f^{(v)} = \begin{bmatrix} f^{(1,v)} \\ f^{(2,v)} \\ \vdots \end{bmatrix} \quad \tilde{\varepsilon}^{(v)} = \begin{bmatrix} \tilde{s} \\ \tilde{\mu}^{(v)} \end{bmatrix} - \begin{bmatrix} \tilde{f}^{(v)} \\ \tilde{\eta}^{(v)} \end{bmatrix}$$
$$\mu^{(x)} = \begin{bmatrix} \mu^{(1,x)} \\ \mu^{(2,v)} \\ \vdots \end{bmatrix} \quad f^{(x)} = \begin{bmatrix} f^{(1,x)} \\ f^{(2,x)} \\ \vdots \end{bmatrix} \quad \tilde{\varepsilon}^{(x)} = \mathcal{D}\tilde{\mu}^{(x)} - \tilde{f}^{(u)} \tag{A.1}$$
$$\tilde{\varepsilon}^{(u)} = \begin{bmatrix} \tilde{\varepsilon}^{(v)} \\ \tilde{\varepsilon}^{(x)} \end{bmatrix} \quad \tilde{\Pi}^{(u)} = \begin{bmatrix} R^{(x)} \otimes diag\left(\exp\left(\pi^{(x)}\right)\right) & \\ & R^{(x)} \otimes diag\left(\exp\left(\pi^{(x)}\right)\right) \end{bmatrix} \tilde{u} = \begin{bmatrix} \tilde{v} \\ \tilde{x} \end{bmatrix}$$

This gives a simple form for the (Gibbs) energy that comprises a log-likelihood and prior

$$\begin{aligned} \mathcal{L} &= \mathcal{L}^{(u)} + \mathcal{L}^{(\varphi)} \\ \mathcal{L}^{(u)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} - \tfrac{1}{2}\ln\left|\tilde{\Pi}^{(u)}\right| \\ \mathcal{L}^{(\varphi)} &= \tfrac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \tfrac{1}{2}\ln\left|\tilde{\Pi}^{(\varphi)}\right| \end{aligned} \tag{A.2}$$

with the following integration scheme

$$\dot{y} = \begin{bmatrix} \dot{\tilde{s}} \\ \dot{\tilde{\mu}} \end{bmatrix} = \begin{bmatrix} \dot{\tilde{s}} \\ \dot{\tilde{\mu}}^{(u)} \\ \dot{\mu}^{(\theta)} \\ \dot{\mu}^{(\gamma)} \\ \dot{\mu}'^{(\theta)} \\ \dot{\mu}'^{(\gamma)} \end{bmatrix} = \begin{bmatrix} \mathcal{D}\tilde{s} \\ \mathcal{D}\tilde{\mu}^{(u)} - \mathcal{F}_{\tilde{u}} \\ \mu'^{(\theta)} \\ \mu'^{(\gamma)} \\ -\mathcal{F}_{\theta} - \kappa\mu'^{(\theta)} \\ -\mathcal{F}_{\gamma} - \kappa\mu'^{(\gamma)} \end{bmatrix} \quad \Im = \frac{\partial \dot{y}}{\partial y}$$
$$= \begin{bmatrix} \mathcal{D} & & & & & \\ -\mathcal{F}_{\tilde{u}\tilde{s}} & \mathcal{D} - \mathcal{F}_{\tilde{u}\tilde{u}} & & & & \\ & & & & I & \\ & & & & & I \\ -\mathcal{F}_{\theta\tilde{s}} & -\mathcal{F}_{\theta\tilde{u}} & -\mathcal{F}_{\theta\theta} & -\mathcal{F}_{\theta\gamma} & -\kappa & \\ -\mathcal{F}_{\gamma\tilde{s}} & -\mathcal{F}_{\gamma\tilde{u}} & -\mathcal{F}_{\gamma\theta} & -\mathcal{F}_{\gamma\gamma} & & -\kappa \end{bmatrix} \tag{A.3}$$

This system can be solved (integrated) using a local linearization (Ozaki, 1992) with updates $\Delta y = (\exp(\Delta t\,\Im) - I)\Im(t)^{-1}\dot{y}$ over time steps $\Delta t$, where $\Im(t)$ the filter's Jacobian. Note that we have omitted terms that mediate changes in the motion of state estimates due to changes in parameter estimates. This is because changes in parameter estimates are negligible at the time scale of changes in states. The requisite gradients (evaluated at the conditional expectation) are, with a slight abuse of notion when dealing with derivatives with respect to vectors

$$\begin{aligned} \mathcal{F}_{\tilde{u}} &= \tfrac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)} + \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} - \tfrac{1}{2}tr\left(\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\Sigma}^{(u)}\right) + \tfrac{1}{2}tr\left(\mathcal{P}_{\tilde{u}}\mathcal{C}\right) \\ \mathcal{F}_{\gamma} &= \tfrac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)} + \Pi^{(\gamma)}\mu^{(\gamma)} - \tfrac{1}{2}tr\left(\tilde{\Pi}^{(u)}_{\gamma}\tilde{\Sigma}^{(u)}\right) + \tfrac{1}{2}tr\left(\mathcal{P}_{\gamma}\mathcal{C}\right) \\ \mathcal{F}_{\theta} &= \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} + \Pi^{(\theta)}\mu^{(\theta)} + \tfrac{1}{2}tr\left(\mathcal{P}_{\theta}\mathcal{C}\right) \end{aligned} \tag{A.4}$$

The corresponding curvatures are (neglecting second-order terms involving states and parameters and second-order derivatives of the conditional entropy)

$$\begin{aligned} \mathcal{F}_{\tilde{u}\tilde{s}} &\approx \mathcal{L}_{\tilde{u}\tilde{s}} = \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{s}} \\ \mathcal{F}_{\gamma\tilde{s}} &\approx \mathcal{L}_{\gamma\tilde{s}} = \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{s}} \\ \mathcal{F}_{\theta\tilde{s}} &\approx \mathcal{L}_{\theta\tilde{s}} = \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{s}} \\ \mathcal{F}_{\tilde{u}\tilde{u}} &\approx \mathcal{L}_{\tilde{u}\tilde{u}} = \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{u}} + \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)} + \tilde{\varepsilon}^{(u)}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\tilde{u}} \\ \mathcal{F}_{\tilde{u}\theta} &\approx \mathcal{L}_{\tilde{u}\theta} = \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\theta} + \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} \\ \mathcal{F}_{\theta\theta} &\approx \mathcal{L}_{\theta\theta} = \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\theta} + \Pi^{(\theta)} \\ \mathcal{F}_{\gamma\theta} &\approx \mathcal{L}_{\gamma\theta} = \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} \approx \mathcal{F}^{T}_{\theta\gamma} \\ \mathcal{F}_{\tilde{u}\gamma} &\approx \mathcal{L}_{\tilde{u}\gamma} = \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)} \\ \mathcal{F}_{\gamma\gamma} &\approx \mathcal{L}_{\gamma\gamma} = \tfrac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)} + \Pi^{(\gamma)} \\ \tilde{\Pi}^{(v)}_{w} &= R^{(v)} \otimes diag\left(\pi^{(v)}_{w} \times \exp\left(\pi^{(v)}\right)\right) \\ \tilde{\Pi}^{(x)}_{w} &= R^{(x)} \otimes diag\left(\pi^{(x)}_{w} \times \exp\left(\pi^{(x)}\right)\right) \end{aligned} \tag{A.5}$$

Finally, the conditional precision and its derivatives are given by the curvature of the (Gibbs) energy

$$\begin{aligned} \mathcal{C}^{-1} = \mathcal{P} = \mathcal{L}_{\mu\mu} &\approx \begin{bmatrix} \mathcal{L}_{\tilde{u}\tilde{u}} & \mathcal{L}_{\tilde{u}\theta} & \mathcal{L}_{\tilde{u}\gamma} \\ \mathcal{L}_{\theta\tilde{u}} & \mathcal{L}_{\theta\theta} & \mathcal{L}_{\theta\gamma} \\ \mathcal{L}_{\gamma\tilde{u}} & \mathcal{L}_{\gamma\theta} & \mathcal{L}_{\gamma\gamma} \end{bmatrix} \\ \mathcal{P}_{\tilde{u}} &\approx \begin{bmatrix} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\theta\tilde{u}} & \\ \tilde{\varepsilon}^{(u)T}_{\theta\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & 2\tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\theta\tilde{u}} & \\ & & 0 \end{bmatrix} \\ &\quad + \begin{bmatrix} 3\tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & 2\tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} \\ 2\tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} \\ \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} \end{bmatrix} \\ \mathcal{P}_{\theta} &\approx \begin{bmatrix} 2\tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{u}\theta} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\tilde{u}\theta} & \\ \tilde{\varepsilon}^{(u)T}_{\tilde{u}\theta}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)}_{\theta} & & \\ & & 0 \end{bmatrix} \\ &\quad + \begin{bmatrix} 2\tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} \\ \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\tilde{u}}\tilde{\varepsilon}^{(u)}_{\theta} & & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} \\ \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)}_{\theta} \end{bmatrix} \\ \mathcal{P}_{\gamma} &\approx \begin{bmatrix} \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\tilde{u}}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)} \\ \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}_{\theta}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)} \\ \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)}_{\tilde{u}} & \tilde{\varepsilon}^{(u)}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)}_{\theta} & \tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}_{\gamma\gamma}\tilde{\varepsilon}^{(u)} \end{bmatrix} \end{aligned} \tag{A.6}$$

Note that we have simplified the numerics here by neglecting conditional dependencies between the precisions and the states or parameters. These equations may look complicated but can be evaluated automatically using numerical derivatives. All the simulations in this paper used just one routine – *spm_LAP.m*. Demonstrations of this scheme are available as part of the SPM software (http://www.fil.ion.ion.ucl.ac.uk/spm; *DEM_demo.m*) and reproduce the examples in the figures.

### STATE-DEPENDENT NOISE AND WEBER'S LAW

Sensory signals are invariably registered as non-negative quantities (e.g., firing rates of photoreceptors). If we assume the sensory signals $s \approx \ln\varsigma$ are an approximate log-transform of some non-negative variables $\varsigma \in \Re^+$ sampled from a Poisson distribution with rate $\lambda$, we have from Eq. 9 (and using a first-order Taylor expansion):

$$
\begin{aligned}
& s = f^{(v)} + z^{(v)} : s \sim \mathcal{N}\left(f^{(v)}, \Sigma^{(v)}\right) \\
& s = \ln\lambda + \frac{(\varsigma - \lambda)}{\lambda} \approx \ln\varsigma : \varsigma \sim Pois(\lambda) \Rightarrow s \sim \mathcal{N}\left(\ln\lambda, \lambda^{-1}\right) \Rightarrow \\
& f^{(v)} = \ln\lambda \\
& \Sigma^{(v)} = \lambda^{-1}
\end{aligned}
\tag{A.7}
$$

This means that as the expected amplitude of the sensory input increases, $f^{(v)} = \ln\lambda$, so does its precision $\Pi^{(v)} = \lambda = \exp(f^{(v)})$.