Spatial Attention, Precision, and Bayesian Inference: A Study of Saccadic Response Speed

Simone Vossel[1], Christoph Mathys[2,3], Jean Daunizeau[2,3], Markus Bauer[1], Jon Driver[1], Karl J. Friston[1] and Klaas E. Stephan[1,2,3]

[1]Wellcome Trust Centre for Neuroimaging, University College London, WC1N 3BG London, UK [2]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, 8032 Zürich, Switzerland and [3]Laboratory for Social and Neural Systems Research, University of Zurich, 8091 Zürich, Switzerland

Address correspondence to Dr Simone Vossel, Wellcome Trust Centre for Neuroimaging, University College London WC1N 3BG London, UK. Email: s.vossel@ucl.ac.uk

S.V., and C.M. These authors contributed equally.

Inferring the environment's statistical structure and adapting behavior accordingly is a fundamental modus operandi of the brain. A simple form of this faculty based on spatial attentional orienting can be studied with Posner's location-cueing paradigm in which a cue indicates the target location with a known probability. The present study focuses on a more complex version of this task, where probabilistic context (percentage of cue validity) changes unpredictably over time, thereby creating a volatile environment. Saccadic response speed (RS) was recorded in 15 subjects and used to estimate subject-specific parameters of a Bayesian learning scheme modeling the subjects' trial-by-trial updates of beliefs. Different response models—specifying how computational states translate into observable behavior—were compared using Bayesian model selection. Saccadic RS was most plausibly explained as a function of the precision of the belief about the causes of sensory input. This finding is in accordance with current Bayesian theories of brain function, and specifically with the proposal that spatial attention is mediated by a precision-dependent gain modulation of sensory input. Our results provide empirical support for precision-dependent changes in beliefs about saccade target locations and motivate future neuroimaging and neuropharmacological studies of how Bayesian inference may determine spatial attention.

Keywords: cue validity, hierarchical models, variational Bayes, visuospatial processing, volatility

Introduction

Prior beliefs about the location of a behaviorally relevant stimulus facilitate stimulus detection and speed reaction times (RTs). One of the first experimental demonstrations of this effect was provided by Posner's location-cueing paradigm (Posner 1980). In this task, a spatial cue (e.g., an arrow) indicates the most likely position of a behaviorally relevant target stimulus on a trial-by-trial basis. Average RTs are faster on valid trials—where the target appears at the expected or cued location—than on invalid trials, where target location is unexpected. This reflects covert orienting of attention to the cued location in analogy to an attentional spotlight. Attentional orienting enhances information processing at the cued location at the expense of alternative (uncued) locations.

However, there is accumulating evidence that attentional orienting in response to the spatial cue is not an all-or-none phenomenon, but is critically affected by trial history and by the current probabilistic context. For example, RT costs of invalid cueing are larger after a valid than after an invalid trial (Jongen and Smulders 2007)—and RTs to invalid targets increase with the number of preceding valid trials (Vossel et al. 2011). Moreover, the RT difference between invalid and valid trials increases, the higher the proportion of validly cued trials (percentage of cue validity [%CV]; Jonides 1980; Eriksen and Yeh 1985; Giessing et al. 2006; Risko and Stolz 2010). These results imply that subjects infer and predict the current probabilistic context and adjust their behavior accordingly.

The behavioral effects observed in Posner's location-cueing paradigm can be interpreted within recent theoretical frameworks of perception and attention based on Bayesian principles (e.g., Rao 2005; Friston 2009, 2010; Itti and Baldi 2009; Chikkerur et al. 2010; Feldman and Friston 2010). Here, the brain is considered as a Bayesian inference machine (e.g., Dayan et al. 1995; Friston 2009) which maintains and updates a generative model of its sensory inputs. In other words, perception can be framed as an "inverse problem": under a specific generative model, the current state of the world has to be inferred from the noisy signals conveyed by the sensorium. Notably, even when stimuli are presented with a very high signal-to-noise ratio, there are many aspects about the state of the world (i.e., the cause of sensory inputs) that are nontrivial to infer, such as its probabilistic structure (the "laws" that relate causes of stimuli to each other) or nonlinear interactions among causes (e.g., visual occlusion). The overall goal of this architecture is to minimize surprise about sensory inputs and thus underwrite homeostasis—either by updating model-based predictions or by eliciting actions to sample the world according to prior expectations. Notably, because surprise about sensory inputs cannot be evaluated directly, it has been proposed that perception and action optimize a free-energy bound on surprise (Friston et al. 2006; Friston 2009, 2010). Based on this free-energy principle, simulations have demonstrated how spatially selective attention can be understood as a function of precision (confidence or inverse uncertainty) during perceptual inference: attentional selection serves to increase the precision of sensory channels, enabling faster responses to attended stimuli (Feldman and Friston 2010). Physiologically, this attentional effect may be mediated by an increase in the synaptic gain of neuronal populations encoding prediction error. These populations are assumed to project to higher level units in the visual hierarchy where faster changes in neuronal activity are engendered in the context of higher precision (for details, see Feldman and Friston 2010).

An important aspect of Posner's location-cueing task relates to the trial-by-trial uncertainty about the predictive value of the spatial cue (i.e., the probability that the target appears at the cued location in a given trial) (cf. Yu and Dayan 2005).

This becomes particularly important in volatile environments, where the cue predicts the target location with varying probabilities over the course of the experiment—in other words, situations in which probabilistic context changes unpredictably over time. Here, the estimate (representation) of this probability—which we will operationalize in terms of %CV—depends on the integration of information over past events.

A simple description of trial-by-trial learning of cue-target contingencies is provided by reinforcement learning models such as Rescorla–Wagner (Rescorla and Wagner 1972). In these models, the update of the probability estimate (in our case, the probability that the target will appear in the cued hemifield) is the product of a fixed learning rate and a prediction error (i.e., the difference between observed and predicted outcome). The learning rate determines the impact of the prediction error on the belief update and, at the same time, determines to what extent the current belief is affected by past events. In other words, it determines the influence of previous trials (cf. Rushworth and Behrens 2008).

While the Rescorla–Wagner rule describes a variety of human and animal behaviors, it is a heuristic approach that does not follow from principles of probability theory. Moreover, it suffers from some practical limitations that might be overcome by the application of Bayesian principles (Gershman and Niv 2010). For associative learning paradigms, hierarchical Bayesian learning models provide a principled prescription of how beliefs are updated optimally in the presence of new data. These models may provide a more plausible account of behavior than the Rescorla–Wagner rule, particularly in volatile environments where a fixed learning rate is suboptimal (Behrens et al. 2007; den Ouden et al. 2010).

Recently, a generic hierarchical, approximately Bayes-optimal learning scheme was introduced that grandfathers and extends existing normative models (Mathys et al. 2011). This model uses a variational approximation to the optimal Bayesian solution. This approximation results in analytical update equations that 1) minimize free energy, 2) are extremely fast to evaluate, 3) contain parameters allowing for individual differences in learning, and 4) directly express the crucial role of prediction errors (and their weighting by uncertainty) that play such a prominent role in predictive coding schemes based on the free-energy principle described above. Crucially, this Bayesian scheme can be applied to empirical behavioral data, allowing one to compare different models of subject responses and quantify their trial-by-trial estimates of states of the environment that lead to sensory predictions, including the precision of these estimates. This enables formal tests of free-energy-based accounts of attention using empirically observed behavior that complements simulation work (e.g., Feldman and Friston 2010). In particular, one can establish the aspects of a Bayesian learning model that are most influential in determining response speed (RS). While one might hypothesize a relationship between precision and RS in the present attentional cueing task (or even more generally; see, e.g., Whiteley and Sahani 2008), other studies (employing different experimental paradigms) have shown that RTs can be related to the (log) probability estimate per se (Carpenter and Williams 1995; Anderson and Carpenter 2006; Brodersen et al. 2008; den Ouden et al. 2010), or to the amount of surprise that is associated with a particular stimulus (Bestmann et al. 2008). Here, we try to explain observed responses, under these different assumptions. To this end, we formulate competing models that embody different assumptions and formally compare their evidence, using Bayesian model selection (BMS). Practically, in contrast to RTs, RS tend to have a Gaussian distribution (Carpenter and Williams 1995; Brodersen et al. 2008) and provide a better-behaved response measure for modeling.

In particular, we here apply this hierarchal Bayesian learning model to saccadic RS data from a variant of Posner's location-cueing paradigm with changes of probabilistic context (%CV) that are unknown to the subject. Saccadic eye movements and covert spatial attention are closely related and share a common functional neuroanatomy (Corbetta et al. 1998; Nobre et al. 2000; Perry and Zeki 2000; Beauchamp et al. 2001; de Haan et al. 2008). There is strong evidence that eye movements to a given location are inevitably preceded by covert attention shifts to this location, enhancing local perceptual processing (e.g., Deubel and Schneider 1996; Godijn and Theeuwes 2003; Dore-Mazars et al. 2004; Deubel 2008). The "premotor theory of attention" (Rizzolatti et al. 1987) states that attentional orienting may be functionally equivalent to saccade planning and initiation, and that therefore programming a saccade causes a shift of spatial attention. In a related theory, the "Visual Attention Model" (Schneider 1995), a single visual attention mechanism is proposed that controls both the selection for perception and the selection for action. Here, attention shifts are not caused by—but are a precondition for—saccade preparation (Deubel 2008). The obligatory coupling between spatial attention and saccade programming is also evident in a recent computational model of evidence accumulation in the visuomotor cascade: visually responsive neurons that can be found in the frontal eye fields (FEF), the lateral intraparietal area, and superior colliculi (SC) provide the source of drive for motor neurons in FEF and SC to elicit a saccade (Schall et al. 2011).

Saccadic RS have been shown to be affected by the probability of the saccade target location (Carpenter and Williams 1995; Farrell et al. 2010; Chiau et al. 2011), and there is initial evidence that trial-by-trial changes in saccadic RS reflect learning of probabilistic context according to Bayesian principles (Anderson and Carpenter 2006; Brodersen et al. 2008). Anderson and Carpenter (2006) presented 2 subjects with multiple trial blocks, in which targets initially appeared to the left and right side of fixation with equal probability. After 70–120 trials in each block, this probability could change abruptly, so that saccades were more likely to be made to one of the targets. By fitting an exponential function—modeling the trial-by-trial probability of the target location—the authors showed that saccadic RS is related to the learned prior probability of target appearance. Similarly, Brodersen et al. (2008) presented 3 subjects with blocks of left and right targets with different stochastic properties: the targets were either presented with different fixed probabilities, or the probability of the target location was conditional on the target location in the previous trial (first-order Markov sequence). They used 2 different learning models to ask whether the subjects learned and utilized the marginal probabilities of the target locations or their conditional probabilities (and thus a probability transition matrix).

While both studies (Anderson and Carpenter 2006; Brodersen et al. 2008) address the question of intertrial variability in probabilistic beliefs, they do not deal with the effects of the uncertainty (precision) of these beliefs, which have been formally

implicated in spatial attention (Feldman and Friston 2010). Moreover, both studies employed models that are agnostic about environmental volatility, thereby precluding the possibility that the subjects can adapt their learning rates, based on their current belief about the volatility of the environment.

Here, we extend the previous findings in 2 ways. First, we show that trial-by-trial RS in the location-cueing paradigm can be explained as a function of the precision of trialwise beliefs, as inferred using hierarchical Bayesian inference (Mathys et al. 2011). Second, our model accommodates individual learning processes by introducing subject-specific parameters that couple hierarchical levels and thus provides a novel quantification of, and explanation for, individual learning differences. In what follows, we will refer to the hierarchal Bayesian learning model as the "perceptual model," because this model provides a mapping from hidden states (or environmental causes) to sensory inputs (Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. 2010; Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. 2010). Furthermore, we will introduce and compare different "response models" (Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. 2010; Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. 2010) that describe the mapping from the subject's probabilistic representations (beliefs)—as provided by the perceptual model—to the observed responses (i.e., RS).

Materials and Methods

Subjects

Sixteen healthy subjects gave written informed consent to participate in the current study. One subject had to be excluded from further analysis due to lack of fixation during the cue-target interval. Therefore, data from 15 subjects were analyzed (9 males, 6 females; age range from 23 to 35 years; mean age 27.4 years). All subjects were right-handed and had normal or corrected to normal vision. The study had been approved by the local ethics committee (University College London).

Stimuli and Experimental Paradigm

We used a location-cueing paradigm with central predictive cueing (Posner 1980). Stimuli were presented on a 19-inch monitor (spatial resolution 1024 × 768 pixels, refresh rate 75 Hz) with a viewing distance of 60 cm. On each trial, 2 peripherally located boxes were

Figure 1. Illustration of the experimental task and the manipulation of %CV over the 612 trials.

shown (1.9° wide and 8° eccentric in each visual field, see Fig. 1) that could contain target stimuli. A central diamond (0.65° eccentric in each visual field) was placed between them, serving as a fixation point. Cues comprised a 200-ms increasing brightness of one side of the diamond—creating an arrowhead pointing to one of the peripheral boxes. After a 1200-ms stimulus onset asynchrony (SOA), a target appeared for 100 ms in one of the boxes. The targets were vertical and horizontal circular sinusoidal gratings (1.3° visual angle). Vertical and horizontal gratings were presented with equal probability.
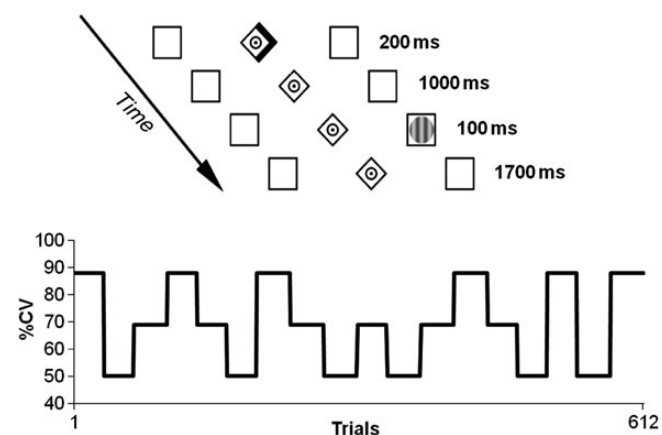
Subjects were instructed to maintain central fixation during the cue period and to make a saccade to the target stimulus as fast as possible. They were encouraged to blink and refixate the central fixation dot after the saccade. After a short practice session of 64 trials—with constant 88%CV—the experiment comprised 612 trials with blockwise changes in %CV that were unknown to the subjects. After half of the trials, the subjects had a short rest of 1 min. Each block with constant %CV contained an equal number of left and right targets, counterbalanced across valid and invalid trials. %CV changed after either 32 or 36 trials switching unpredictably to levels of 88%, 69%, or 50% (see Fig. 1). Subjects were told in advance that there would be changes in %CV over the course of the experiment, but were not informed about the levels of these probabilities or when they would change. Each subject was presented with the same sequence of trials. This is a standard procedure in computational studies of learning processes that require inference on conditional probabilities in time series (cf. Behrens et al. 2007; Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. 2010). In these situations, the parameters of the learning process depend on the exact sequence of trials used. Although this dependency will diminish asymptotically with increasing numbers of trials, for the relatively short sequences (of a few hundred trials at best) that are feasible within a standard experiment, introducing a different sequence for each participant could increase the variability of parameter estimates, over and above the intrinsic interindividual trait-differences per se. We therefore decided to keep trial sequence constant to ensure that differences in model parameters can be attributed to subject-specific rather than task-specific factors.

Eye Movement Data Recording and Analysis

Participants sat in a dimly lit sound-proof cabin with their head stabilized by a chinrest. Eye movements were recorded from the right eye with an EyeLink 1000 desktop mounted eye-tracker (SR Research Ltd) with a sampling rate of 250 Hz. A 9-point eye-tracker calibration and validation was performed at the start of the experiment and after the pause in the middle of the experiment. The validation error was <1° of visual angle.

Eye movement data were analyzed with MATLAB (Mathworks) and ILAB (Gitelman 2002). Blinks were filtered out and pupil coordinates within a time window of 20 ms around the blink were removed. Trials with >20% missing data were discarded from the analyses. To ensure central fixation after presentation of the spatial cue, the period between cue and target was analyzed for gaze deviations from the center. After target appearance, only the first saccade was analyzed. Saccades were identified when the eye velocity exceeded 30°/s (Fischer et al. 1993; Stampe 1993). After this threshold was reached, the beginning of the saccade was defined as the time when the velocity exceeded 15% of the trial-specific maximum velocity (Fischer et al. 1993). Likewise, the end of the saccade was defined as the time when the velocity fell below 15% of the trial-specific maximum velocity. Moreover, the saccade amplitude needed to subtend at least two-thirds of the distance between the fixation point and the actual target location. Saccadic RT was defined as the latency between target and saccade onset. Saccades in which the starting position was not within a region of 1° from the fixation point and saccades with a latency <90 ms were discarded from the analyses. Our analyses focused on inverse RTs (i.e., RS) since, in contrast to RTs, RS are normally distributed (cf. Carpenter and Williams 1995; Brodersen et al. 2008).

To assess the effect of probabilistic context (true %CV), mean RS for each subject and for each %CV condition were entered into a 2 (cue: valid, invalid) × 3 (%CV: 50, 69, 88%) within-subjects analysis of variance (ANOVA). In this analysis, evidence for an impact of probabilistic

context would be reflected in a significant cue × %CV interaction effect —with increasing differences between valid and invalid RS with higher %CV. Results from this analysis are reported in the Results section at a significance level of $P < 0.05$ after Greenhouse–Geisser correction. Condition-specific mean RS was also calculated separately for the 2 halves of the experiment and analyzed with a 2 (cue: valid, invalid) × 3 (%CV: 50, 69, 88%) × 2 (time: first half, second half) within-subjects ANOVA (note that each %CV condition was presented 3 times in each half, cf. Fig. 1).

Having established the significance of the experimental effects, we then sought to model them in terms of hierarchical Bayesian updating:

Perceptual Model

In what follows, we briefly outline the generative perceptual model (for details on the mathematical derivation of the update equations see Appendix section and Mathys et al. 2011). The perceptual model (dark gray panel in Fig. 2) comprises a hierarchy of 3 hidden states (denoted by x), with states 2 and 3 evolving in time as Gaussian random walks.
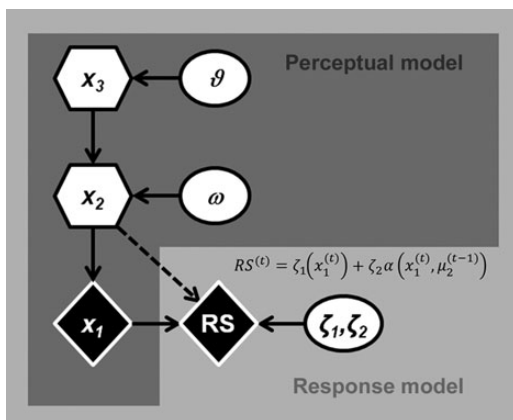
The probability of a target appearing at the cued location in a given trial (t) (represented by the state $x_1^{(t)}$, with $x_1 = 1$ for valid and $x_1 = 0$ for invalid targets) is governed by a state x_2 at the next level of the hierarchy. (Note that in this particular experiment the target stimulus was visible without any ambiguity [very high signal-to-noise ratio]; this means there is a simple deterministic mapping between the [mean of] x_1 and input u of the general model, which allows for situations with perceptual ambiguity [e.g., visual noise].) x_2 is a real number, and the probability distribution of x_1 given x_2 is described by a logistic sigmoid (softmax) function, so that the states $x_1 = 0$ and $x_1 = 1$ are equally probable when $x_2 = 0$.

$$p(x_1^{(t)}|x_2^{(t)}) = s(x_2^{(t)})^{x_1^{(t)}}(1 - s(x_2^{(t)}))^{1-x_1^{(t)}} \quad (1)$$

with $s(x)\underline{def}(1/(1 + e^{-x}))$ and $x_1 \in \{0, 1\}$.

Hence, in the current location-cueing paradigm, x_2 determines the trial-specific estimate for %CV. The probability of x_2 itself changes over time (trials) as a Gaussian random walk, so that the value $x_2^{(t)}$ will be normally distributed around $x_2^{(t-1)}$ from the previous trial, with the variance of the distribution described by the term $e^{x_3^{(t)}+\omega}$. (This is a simplified version of the full model in Mathys et al. 2011, in which a scaling parameter (κ) has been set to 1.)

$$p(x_2^{(t)}) \sim N(x_2^{(t-1)}, e^{x_3^{(t)}+\omega}). \quad (2)$$

Figure 2. Graphical illustration of the perceptual (generative) model with states x_1, x_2, and x_3. The model parameters ω and ϑ impact on the time course of subjects' inferred belief about the states x and are estimated from the individual subject RS data. Circles represent constants, while diamonds represent quantities that change in time (i.e., that carry a time (or trial) index). Hexagons, like diamonds, represent quantities that change in time but that additionally depend on their previous state in time in a Markovian fashion.

Changes in x_2 over time (trials) are thus determined by the quantities x_3 (the 3rd level of the hierarchy) and a subject-specific parameter ω that allows for individual differences in the updating of x_2. Accordingly, x_3 and ω can be regarded as state-dependent and subject-specific (trait-like) measures of log-volatility (trial-by-trial variability in x_2), respectively. The state $x_3^{(t)}$ (Fig. 2) on a given trial is normally distributed around $x_3^{(t-1)}$ with a variance determined by the constant subject-specific parameter ϑ. The parameter ϑ is a measure of meta-volatility (volatility of volatility) that determines the variability of the log-volatility over time.

$$p(x_3^{(t)}) \sim N(x_3^{(t-1)}, \vartheta). \quad (3)$$

To map from the sensory inputs to the probabilistic representations of the subject, the perceptual model needs to be inverted to obtain posterior densities on the hidden states x. In the following, the sufficient statistics of the subject's posterior belief will be denoted by μ (mean) and σ (variance) or $\pi = (1/\sigma)$ (precision). We use the hat symbol (^) to denote predictions before the observation of x_1 on a given trial. Variational inversion under a mean field approximation yields simple analytical update equations—where belief updating rests on precision-weighted prediction errors. The update of the posterior mean at level i in the hierarchy on trial t has the following general form (at the second level of the model in this study, the precision weighting has a slightly different form, i.e., $\hat{\pi}_1^{(t)}/(\hat{\pi}_2^{(t)} \hat{\pi}_1^{(t)} + 1)$, because of the sigmoid transform that relates the second level to the first; see equation A2.):

$$\Delta\mu_i^{(t)} \propto \frac{\hat{\pi}_{i-1}^{(t)}}{\pi_i^{(t)}} \delta_{i-1}^{(t)}. \quad (4)$$

In equation (4), $\hat{\pi}_{i-1}^{(t)}$ is the precision of the prediction about the state at the level below and $\pi_i^{(t)}$ is the precision of the posterior belief about the state at the current level, while $\delta_{i-1}^{(t)}$ is the prediction error about the input from the level below. For the derivation of these updates and their detailed form, see the Appendix section and Mathys et al. (2011). In brief, these equations provide approximately Bayes-optimal rules for the trial-by-trial updating of the representations (beliefs) that determine the subject's estimate of the probability that the target appears at the cued location on a particular trial. Note that this is an individualized Bayes optimality, in reference to the subject-specific values for the parameters ω (determining subject-specific log-volatility) and ϑ (subject-specific meta-volatility).

It is interesting to note that the general update equations (4) arising from the variational hierarchical Bayesian scheme are formally similar to reinforcement learning models such as the Rescorla–Wagner rule (Rescorla and Wagner 1972). As described in detail in Mathys et al. (2011), the precision weighting of the updates at the second level can be understood as a time-varying learning rate, which varies with the state-dependent component μ_3 of the log-volatility (see Appendix section for details). An alternative—but equally useful—perspective on the generic update scheme in equation (4) is in terms of Bayesian filtering, for example, Kalman filtering. The Kalman filter can be regarded as an extension of the Rescorla–Wagner rule. It formalizes the predictive relationship between events, but also comprises expectations about how this relationship is expected to change over time and takes into account the uncertainty about this prediction (Dayan 2000). In this context, the precision–dependent weighting of prediction errors in our scheme corresponds to something called the Kalman gain that is applied to prediction errors to provide optimal predictions about the future. These perspectives on precision (reinforcement learning rates and Kalman gain) illustrate the formal equivalence between reinforcement learning, predictive coding, and Bayesian filtering, disclosed by the general scheme used here.

In addition to the full hierarchical Bayesian model, we employed 2 reduced versions of the perceptual model. This allowed us to evaluate whether the relatively complex hierarchical model was needed to explain our subjects' behavior. Specifically, the full hierarchical model assumes that 1) subjects are capable of learning the hierarchical structure of the probabilities in this experiment and 2) exploit this
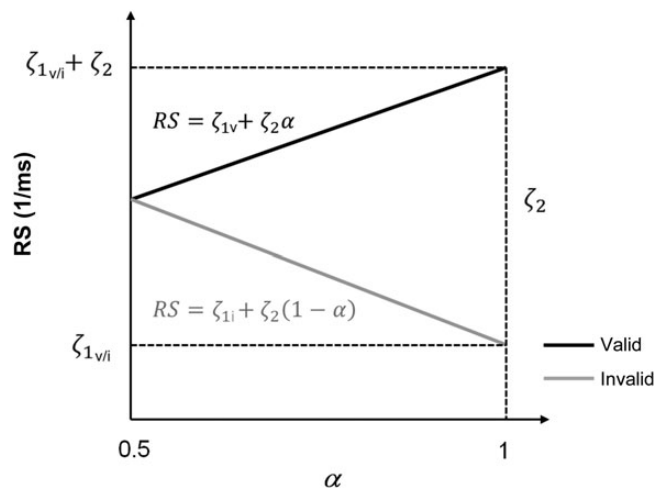
knowledge to dynamically adapt the speed at which they update beliefs (i.e., learning rate) by using precision-weighted prediction errors. Although these assumptions are theoretically well founded (cf. Mathys et al. 2011), it needs to be shown that equivalent explanations of the data could not be afforded by simpler, nonhierarchical learning models. Therefore, we specified 2 alternative perceptual Bayesian models that eschewed assumptions about hierarchically structured learning, but in different ways. The first alternative model assumed that subjects ignored the instructions that the environment was volatile, expecting negligible changes in log-volatility (third level): ϑ was thus fixed to zero, and only ω was estimated. The second perceptual model did not use estimates of environmental volatility to adapt learning. In this model, the influence of x_3 on the variance of x_2 was therefore fixed to zero (cf. eq. 2), so that levels 2 and 3 of the model became decoupled and rendered the values at the third level of the model irrelevant (an equivalent effect is obtained by fixing $x_3^{(t)}$ to zero).

Response Models

To map from the subject's posterior beliefs to observed responses, 3 different response models were compared. A detailed analysis and motivation of their functional forms can be found in the Appendix section. All response models predict inverse RT (RS), since the distribution of RS is typically normal, in contrast to RTs themselves (Carpenter and Williams 1995). Furthermore, all response models describe trialwise RS as a linear function of an attentional factor α, based on the posterior beliefs of the perceptual model. This factor can be regarded as the proportion of attentional resources allocated to the cued location (i.e., α is normalized to the unit interval):

$$RS = \begin{cases} \zeta_{1_{valid}} + \zeta_2 & \text{for } x_1 = 1 \text{ (i.e., valid trial)}, \\ \zeta_{1_{invalid}} + \zeta_2(1 - \alpha) & \text{for } x_1 = 0 \text{ (i.e., invalid trial)}. \end{cases} \quad (5)$$

Note that in all cases, RS is the same function of attentional resources allocated to the outcome location: on valid trials, this is the amount of attentional resources α allocated to the cued location, while—on invalid trials—it is the amount of attentional resources $1 - \alpha$ allocated to the uncued location (cf. Fig. 3). Here, $\zeta_{1_{valid}}$, $\zeta_{1_{invalid}}$, and ζ_2 are subject-specific parameters that are estimated from the data. Minimal and maximal RS for valid and invalid trials are then defined by $\zeta_{1_{valid/invalid}}$ and $\zeta_{1_{valid/invalid}} + \zeta_2$, respectively.

Figure 3. Illustration of the relationship between RS (inverse RT) and the quantity α representing the amount of attentional resources allocated to the cued location. For each response model, RS were assumed to be linearly related to α (which differs between the 3 models, see Appendix). Note the opposite behaviour of RS for increasing α on valid (black line and equation) and invalid (gray line and equation) trials (cf. eq. 5).

Crucially, the 3 competing response models differ in how they specify the dependence of α on computational quantities from the perceptual model: these are precision, belief, and surprise about the sensory signal, respectively. All 3 models respected the same boundary conditions, i.e., α remained confined to the unit interval with $\alpha = 0.5$ when $\hat{\mu}_1 = 0.5$ (cf. Appendix section and Fig. 4).

The first response model focused on the precision estimate at the first level of the perceptual model—following the recent proposal by Feldman and Friston (2010) concerning the role of precision for spatially selective attention in the location-cueing paradigm. Here, we assumed that on a given trial t, the attentional factor $\alpha^{(t)}$ was determined by a sigmoid transformation (s) of $\hat{\pi}_1^{(t)}$, the precision of the prediction at the first level, relative to its minimal value (i.e., 4 when $\hat{\mu}_1 = 0.5$):

$$\alpha^{(t)} = s(\hat{\pi}_1^{(t)} - 4). \quad (6)$$

In the second response model, the "belief" model, the attentional factor α depended on the strength of the prediction about CV:

$$\alpha^{(t)} = \hat{\mu}_1^{(t)} = s(\mu_2^{(t-1)}). \quad (7)$$

The third response model (surprise) was based upon the (Shannon) surprise associated with the target stimulus. The Shannon surprise (Shannon 1948) is the negative logarithm of a probability (here $\hat{\mu}_1^{(t)}$). This response model was inspired by a previous study on cueing of motor responses in which RTs were examined in relation to trialwise surprise (Bestmann et al. 2008). Here, we defined α as a nonlinear function of Shannon surprise:

$$\alpha^{(t)} = \frac{1}{(1 + surprise(\hat{\mu}_1^{(t)}))} \quad (8)$$

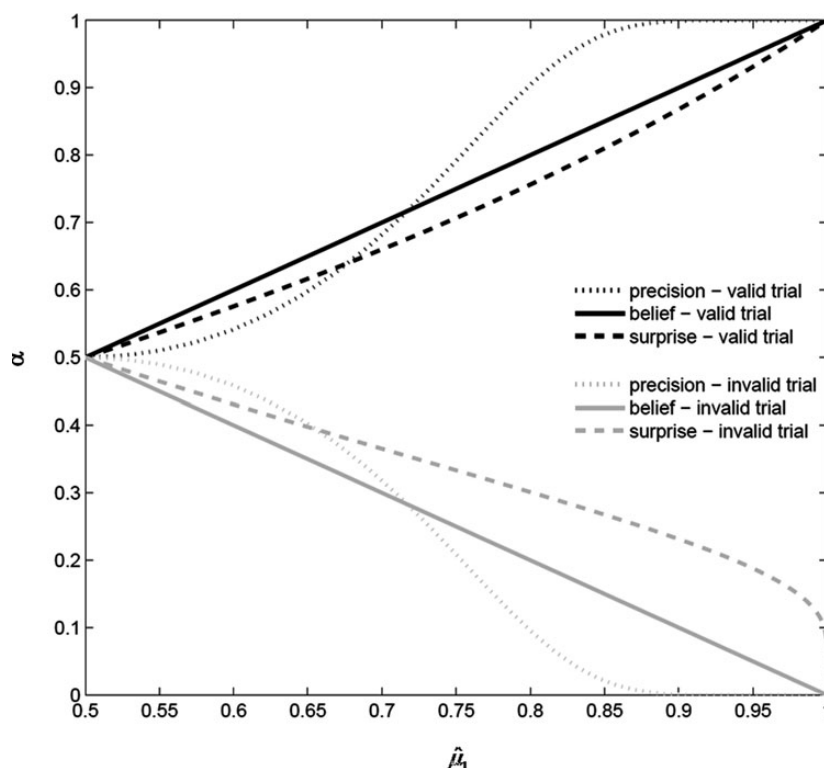with surprise$(\hat{\mu}_1^{(t)}) = -\log_2 p(x_1^{(t)} = 1 | \hat{\mu}_1^{(t)}) = -\log_2(\hat{\mu}_1^{(t)})$.

In summary, we specified 3 alternative perceptual models and 3 alternative response models. This resulted in a 3×3 factorial model space. We compared the relative plausibility of these models using a random effects BMS procedure at the group level, both for individual models and model families (Stephan et al. 2009; Penny et al. 2010). In addition, we compared these models to a standard Rescorla–Wagner learning model as well as to a model assuming that the true underlying (categorical) probabilities were known to subjects—in other words, they did not have to be inferred on the basis of experience. In the latter 2 models, trialwise RS was supposed to be linearly related to the estimated or true %CV, respectively.

Estimation of the Model Parameters

The perceptual model parameters ω and ϑ, as well as the response model parameters $\zeta_{1_{valid}}$, $\zeta_{1_{invalid}}$, and ζ_2 were estimated from the trialwise RS measures using variational Bayes. This enabled us to obtain an estimate of the log model evidence for model comparison and to evaluate the posterior densities of the model parameters. In short, variational Bayes optimizes the (negative) free-energy F as a lower bound on the log-evidence, such that maximizing F minimizes the Kullback–Leibler divergence between exact and approximate posterior distributions (for details, see Friston et al. 2007; Penny et al. 2007). MATLAB functions for the variational Bayes scheme were derived from the DAVB toolbox (Daunizeau et al. 2009; dl.dropbox. com/u/18527014/CODE/DAVB.zip). This approach is analogous to the Bayesian inversion of Dynamic Causal Models for functional imaging or electrophysiological data (dynamic causal modeling [DCM], Friston et al. 2003; Daunizeau et al. 2011).

As any Bayesian approach, variational Bayesian inversion requires the definition of priors on the parameters. Importantly, the prior (co)variance influences the estimability of parameters, e.g., their degree of independence; also by choosing a very small prior variance (very high prior precision) one can effectively fix the value of a parameter. Table 1 provides the priors used for inverting the full hierarchical model. In the perceptual model, starting values for μ and σ of states 2

Figure 4. Illustration of the amount of attentional resources α for the 3 different theoretical response models as a function of $\hat{\mu}_1$.

Table 1

Prior mean and variance for the parameters of the perceptual and response models, and the noise parameter

Parameter	Prior mean	Prior variance
Perceptual model		
ω	−6	100
ϑ	0.1	100
Response model		
$\zeta_{1_{valid}}$	0.0052	0.1
$\zeta_{1_{invalid}}$	0.0052	0.1
ζ_2	0.0006	0.001
Noise parameters		
ζ_3	0.001	1000

Note: ϑ is estimated in logit-space, while ζ_{1_valid}, $\zeta_{1\_invalid}$, and ζ_2 are estimated in log-space.

and 3 were fixed and an upper bound of 1 was defined for the parameter ϑ. In the response model, the prior variance for ζ_2, which parameterizes the relationship between the attentional factor α and RS (Fig. 3), was set to a fairly small value (10^{-3}). In other words, we assumed that the relation between RS and α (see eq. 5) did not differ greatly across subjects. In contrast, to account for individual baseline differences in RS (i.e., the intercept of the linear slope); the response model parameters $\zeta_{1_{valid}}$ and $\zeta_{1_{invalid}}$ were given a larger prior variance, allowing for substantial individual differences between subjects.

While trials with missing responses did not contribute to parameter estimation, they did contribute to estimating the evolution of the states x, since they still provided the subject with an observation about the cue-target contingency. In other words, we used what the subject saw to estimate the Bayes-optimal estimate of hidden states over the experiment—under a particular set of parameters and used subject responses to optimize the parameters of the perceptual and response models.

Bayesian Model Selection

BMS evaluates the relative log-evidence (or log-marginal likelihood) of alternative models. The log-evidence of a model is the negative surprise about the data, given a model, and represents a generic trade-off between the accuracy and complexity of a model that can be derived from first principles of probability theory. Over the past decade, BMS has become a standard approach to assess the relative plausibility of competing models that describe how neurophysiological or behavioral responses are generated (cf. Stephan et al. 2009; Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. 2010, Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. 2010). Here, we use it to disambiguate different hypotheses about how learning (as described by the perceptual models) and decision making (as described by the response models) evolve across and within trials.

Above, we introduced 3 perceptual models and 3 response models ("precision", "belief", and "surprise"). Combining these alternatives provides 9 models in a 3×3 factorial model space, plus the additional 2 control models (standard Rescorla–Wagner model and a model assuming that the true probabilities were known to the subjects). To assess the relative plausibility of our models at the group level, we used random effects BMS (Stephan et al. 2009) and report both posterior probabilities and the exceedance probabilities of the competing models. Importantly, random effects BMS treats the model itself as being selected probabilistically by each subject in the population; i.e., as a random effect following a Dirichlet distribution. In brief, this enables group-level inference while accounting for interindividual differences (e.g., the optimal model can vary across subjects). Critically, random effects BMS not only assesses the relative goodness of competing models but also quantifies (via the Dirichlet parameter estimates) the degree of heterogeneity in the sample studied (Stephan et al. 2009).

The exceedance probability of a model is the probability that it is more likely than any other model considered, given the data. For example, an exceedance probability of 95% for a particular model means that one has 95% confidence that this model has a greater

posterior probability than any other model tested (Stephan et al. 2009). Both posterior probabilities and exceedance probabilities sum to unity over all models tested.

Reproducibility of Results

To examine the reproducibility and hence generalizability of our findings, we performed an additional analysis, using an independent set of subjects ($n = 16$, 8 males, 8 females: age range from 19 to 30 years; mean age 23.4 years). Again, all subjects were right-handed and had normal or corrected to normal vision. The subjects were tested as part of a separate psychopharmacological study employing a within-subject cross-over design. The data presented here were taken from the placebo session only, during which the subjects received a multivitamin tablet. This study was approved by the NHS Research Ethics Committee.

The subjects were presented with exactly the same trial sequence as in the original study. The within-trial structure was also almost identical, with slight modifications to the timing of the task: the cue-target SOA was reduced to 800 ms and the target was presented for 200 ms. Moreover, the trials were interspersed with 108 "null-trials" where only the baseline display (the fixation point and peripheral boxes) was shown. The task lasted 35 min and comprised 4 short rest periods. Finally, the subjects received a slightly longer training than the original group (one session with 100 trials with constant 80%CV and one session with 121 trials with changes in %CV). The same procedures and analyses as outlined above were applied to the eye movement data, except that the data here were recorded with a sampling rate of 1000 Hz. Using trialwise RS, we again fitted the parameters of the perceptual and response models outlined above.
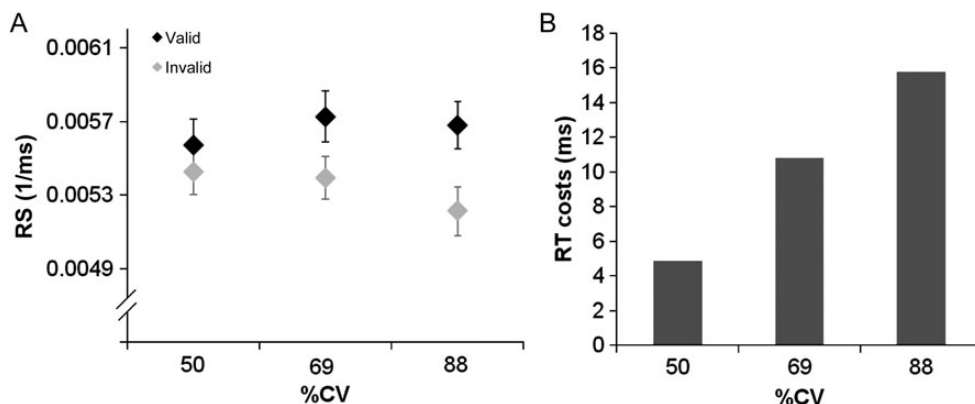
Results

Fixation During the Cue-Target Interval and Missing Trial Data

Between the appearance of the cue and the target, the subjects fixated the center of the display in $87.7 \pm 2.3\%$ (mean ± SEM) of the trials—within a region of interest of 1°—and in $95.4 \pm 1.2\%$ of the trials, within in a region of 2° from the fixation point. The proportion of trials with missing eye data or missing or incorrect saccades amounted to $20.0 \pm 3\%$, so that on average 80% of the trials (487 of 612 trials) were analyzed. Trials excluded from analysis were due to anticipated responses ($3 \pm 1\%$), incorrect or absent saccades ($5 \pm 1\%$), saccades not starting from the fixation zone ($8 \pm 1\%$), or missing data points, e.g., due to blinks ($4 \pm 1\%$). There was no significant difference in the percentage of correct trials between the first and second half of the experiment (paired t-test, $P = 0.895$).

Classical Inference About the Effects of Probability on RS

The 2 (cue: valid, invalid) × 3 (%CV: 50, 69, 88%) ANOVA on RS data revealed a significant main effect of cue ($F_{1,14} = 8.8$, $P = 0.01$) reflecting faster responses (higher RS) on valid than on invalid trials. The main effect of %CV was not significant—in other words, averaging over valid and invalid trials removed any effect of probability. Crucially, we observed a significant cue × %CV interaction effect ($F_{1.9,26.6} = 9.5$, $P = 0.001$) reflecting a differential impact of %CV on valid and invalid trials (Fig. 5). A separate analysis also considered general trends in the data over time, e.g., due to fatigue, by including time (first vs. second half of the experiment) as additional factor. This resulted in a 3-factorial cue (valid, invalid) × %CV (50, 69, 88%) × time (first, second half) ANOVA. Again, this analysis revealed a main effect of cue ($F_{1,14} = 8.2$, $P = 0.013$) and a significant cue × %CV interaction ($F_{1.6,22.5} = 10.5$, $P = 0.001$). The main effect of %CV was not significant. Importantly, there was neither a significant main effect of time nor interaction effects of the factor time with any of the other factors (all P's > 0.4).

The cue × %CV interaction effect indicates a significant influence of probabilistic context on the subjects' responses, with stronger attentional orienting to the cue (and higher RT costs after invalid cueing) with higher %CV. However, Figure 5 does not show a strictly monotonic relationship between RS and true %CV for valid cues. This probably results from the fact that the underlying probabilistic structure (i.e., %CV) was unknown to the subjects and was changing in time fairly rapidly. It therefore had to be inferred by the subjects online, and these subject-specific and dynamic estimates should be the relevant predictors of observed RS, not %CV. In other words, the ANOVAs above (and the results in Fig. 5) average across trials that are heterogeneous in terms of subjective probability estimates, and a model predicting the subjective estimates should be superior in explaining behavior (cf. Fig. 9). In what follows, we test this hypothesis, asking whether the empirically observed RS might reflect trial-by-trial updating of the subjects' beliefs according to our Bayesian perceptual model. Additionally, we compare a systematic set of models that combine different putative learning processes (perceptual models) with different ways in which the learned quantities drive behavior (response models).

Figure 5. (*A*) Average RS in valid and invalid trials for the 3 (true) %CV levels. Error bars depict standard errors of the mean (SEM) (*B*) Illustration of how the observed RS costs after invalid cueing translate into RT differences (in ms).

Bayesian Model Selection

Random effects BMS among the 3 perceptual model families (i.e., the full models and the 2 reduced model versions for each of the 3 response models) revealed that the full hierarchical Bayesian model had substantially higher model evidence than the 2 reduced (null) versions (Table 2).

Comparing the 3 response model families (i.e., the precision, belief and surprise models for each of the 3 versions of the perceptual model) showed that the response model based upon precision was clearly superior to the belief and the surprise model (Table 2). Finally, comparison of all 11 individual models revealed that the full hierarchical Bayesian model combined with the precision response model was clearly superior to all other models we considered (Table 2, Supplementary Fig. 1).

Parameters of the Winning Model

The subject-specific values for log-volatility ω and meta-volatility ϑ derived from the full hierarchical perceptual model—based upon precision—are depicted in Figure 6A. Figure 6B shows the minimal and maximal RS for each subject as derived from the response model parameters $\zeta_{1_{\text{valid}}}$, $\zeta_{1_{\text{invalid}}}$, and ζ_2 in relation to the subject's overall (mean) RS. The graph shows that there were considerable differences in the absolute speed of responding across subjects, as parameterized by averaged values for $\zeta_{1_{\text{valid}}}$ and $\zeta_{1_{\text{invalid}}}$, which were estimated from the individual datasets.

In our hierarchical Bayesian scheme, the precision-weighting $\hat{\pi}_1^{(t)}/(\hat{\pi}_2^{(t)}\hat{\pi}_1^{(t)}+1)$ at the second level plays the role of a (time-varying) learning rate that depends on the log-volatility, determined by ω and μ_3. As shown previously (Mathys et al. 2011), this dependence—on higher order knowledge about change points in the environment—enables more adaptive learning in volatile environments, such as our paradigm. This is also reflected by the BMS results described above, where the hierarchical Bayesian model clearly outperformed a standard Rescorla–Wagner model with a fixed
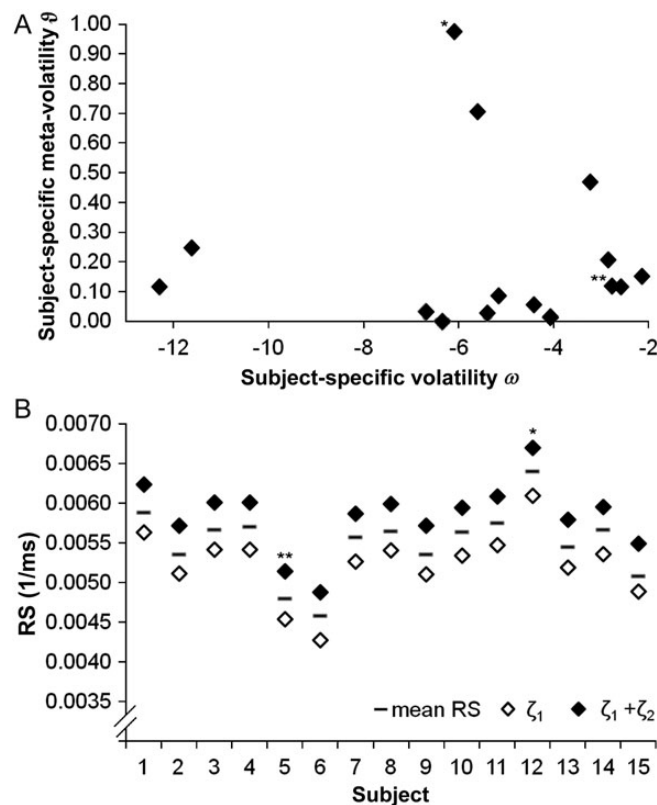
learning rate. However, given the formal similarity of the 2 models, one may expect to find a correlation between the fixed learning rate of the Rescorla–Wagner model and the parameters determining the learning rate of our hierarchical Bayesian model. Figure 7 depicts this relationship between the perceptual parameters ω and ϑ, and the learning rate ε derived from the Rescorla–Wagner model. While there was a significantly positive correlation between the subject-specific volatility estimate ω and learning rate ε ($r = 0.69$; $P = 0.004$), no relationship was observed between ε and the meta-volatility ϑ ($P > 0.25$) (Fig. 7).

To illustrate different individual learning styles, Figure 8 shows the exemplary time courses of the third and first levels of the Bayesian model for 2 subjects with distinct updating behavior. The 2 subjects show differences in the volatility estimate ω as well as the meta-volatility estimate ϑ (cf. Fig. 6 where these subjects are indicated by stars). Although the meta-volatility estimate ϑ is higher in subject A than in subject B, subject B shows faster updating due to a higher volatility estimate ω. In other words, our model shows that the first subject perceives the environment as substantially less volatile than the second subject. As the updates of $\mu_2^{(t)}$ (the estimated CV) are coupled to the estimated log-volatility $\mu_3^{(t-1)}$, this translates into a higher learning rate and quicker updating behavior in the second subject, when the true underlying %CV changes.

Table 2
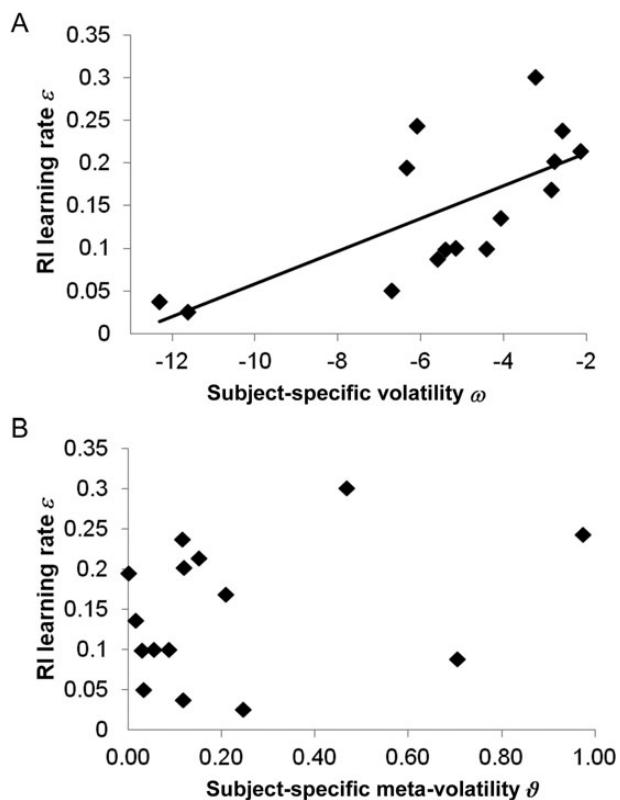Results of the Bayesian model selection (BMS)

Model	Main dataset ($n = 15$)		Replication dataset ($n = 16$)	
	PP	XP	PP	XP
Model family comparison—perceptual models				
Full hierarchical Bayesian family	0.873	0.999	0.777	0.997
Reduced model family ($\vartheta = 0$)	0.064	<0.001	0.105	0.001
Reduced model family ($x_3^{(t)} = 0$)	0.063	<0.001	0.118	0.002
Model family comparison—response models				
"Precision" family	0.756	0.991	0.642	0.930
"Belief" family	0.076	0.001	0.251	0.066
"Surprise" family	0.168	0.008	0.107	0.004
Model comparison of all 11 models				
Full hierarchical Bayesian model "Precision"	0.499	0.995	0.381	0.914
Reduced model ($\vartheta = 0$) "Precision"	0.006	<0.001	0.182	0.074
Reduced model ($x_3^{(t)} = 0$) "Precision"	0.119	0.004	0.047	<0.001
Full hierarchical Bayesian model "Belief"	0.040	<0.001	0.041	<0.001
Reduced model ($\vartheta = 0$) "Belief"	0.040	<0.001	0.042	<0.001
Reduced model ($x_3^{(t)} = 0$) "Belief"	0.040	<0.001	0.074	0.004
Full hierarchical Bayesian model "Surprise"	0.040	<0.001	0.079	0.004
Reduced model ($\vartheta = 0$) "Surprise"	0.040	<0.001	0.039	<0.001
Reduced model ($x_3^{(t)} = 0$) "Surprise"	0.040	<0.001	0.039	<0.001
Rescorla–Wagner model	0.040	<0.001	0.038	<0.001
True categorical probability model	0.040	<0.001	0.038	<0.001

Note: *PP*, posterior probability; *XP*, exceedance probability.

Figure 6. (*A*) Illustration of the subject-specific patterns for the values of the volatility estimate ω and the meta-volatility estimate ϑ. (*B*) Illustration of minimal and maximal RS (as derived from the response model parameters ζ_1 (averaged for valid and invalid trials) and ζ_2) in relation to overall (mean) RS. The symbols single and double asterisks denote the data from subjects A and B depicted in Figure 8, respectively.

Figure 7. Relationship between the perceptual parameters ω and ϑ, and the Rescorla–Wagner learning rate ε.

To illustrate how RTs are related to the precision-based attentional factor α, we pooled RS over different bins of the attentional factor (using bins of 0.1, separately for valid and invalid trials) using estimates of trial-specific α based on the group average values for ω and ϑ. Figure 9 depicts the binned RS over subjects as a function of α. A 2 (cue: valid, invalid) × 6 (precision-based quantity α: 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0) ANOVA revealed a significant main effect of cue ($F_{1,14} = 11.8$, $P = .004$) and a significant cue × α interaction effect ($F_{3.44,48.09} = 10.5$, $P < .001$). We compared these empirical RS values to the RS predicted by the model. For this, we computed the expected RS as a function of α on the basis of the group average values for $\zeta_{1_{valid}}$ and $\zeta_{1_{invalid}}$ and ζ_2 (see Fig. 9). It can be seen that the observed RS shows a similar pattern as the predicted RS. As expected, as precision (confidence in the validity of the cue) increases, there is a RT benefit for valid trials and an equivalent cost for invalid trials. This illustrates that one can explain attention formally in terms of optimizing or learning the relative precision of competing sensory channels.

Reproducibility of Results

In the independent replication, the proportion of trials with missing eye data or missing or incorrect saccades amounted to $7.6 \pm 2\%$, so that on average 92.4% of the trials (566 of 612 trials) were analyzed. Excluded trials were due to anticipated responses ($0.6 \pm 0.2\%$), incorrect or absent saccades ($0.6 \pm 0.2\%$), saccades not starting from the fixation zone ($3.3 \pm 1\%$), or missing data points (e.g., due to blinks) ($3.1 \pm 1\%$). Note that due to the extended training and the

increased number of resting periods, the amount of usable trials was higher than in the original study.

The 2 (cue: valid, invalid) × 3 (%CV: 50, 69, 88%) ANOVA on RS data gave the same results as for the original dataset. Specifically, it revealed a significant main effect of cue ($F_{1,15} = 17.6$, $P = 0.001$) reflecting faster responses (higher RS) on valid than on invalid trials. As before, the main effect of % CV was not significant but we observed a significant cue × % CV interaction effect ($F_{1.99,29.88} = 4.7$, $P = 0.017$). As the data were derived from a within-subject cross-over design (where half of the subjects received the placebo tablet in the first session, while the placebo session for the other half of subjects was the second experimental session), we additionally tested for an effect of session order by adding this variable as a between-subject factor to the ANOVA. No main effect of session order ($P = 0.15$) or interaction of session order with any of the other factors (all $P > 0.28$) was observed.

The results of the Bayesian model comparison are shown in Table 2. Again, the full Bayesian model based upon precision showed the highest exceedance probability (0.914) when compared with alternative models. For the winning model, we again observed a significant positive correlation between the Rescorla–Wagner learning rate ε and ω ($r = 0.59$, $P = 0.017$), while no such relationship was observed between ε and ϑ ($P = 0.97$). In summary, this second dataset provided a full replication of our original results.

Discussion

The present study analyzed saccadic RTs in a location-cueing paradigm with a volatile probabilistic context, probing Bayesian theories of perceptual inference. Extending previous theoretical work (Feldman and Friston 2010), we were able to provide empirical evidence for the free-energy formulation of attention in the context of a Posner paradigm—where CV changed unpredictably in time, thus requiring the subject to learn about environmental volatility. Specifically, using a generic hierarchical Bayesian scheme (Mathys et al. 2011), we compared 3 alternative models of how subjects might update estimates of CV across trials (perceptual models) and crossed these with 3 alternative hypotheses about how posterior beliefs (precision, belief, and surprise) might inform decision making within trials (response models). The resulting 9 models—and 2 control models—were optimized using empirical measures of saccadic RS and their relative plausibility was evaluated using BMS. The results of this model comparison provided strong evidence in favor of the hierarchical Bayesian model combined with the precision response model (Table 2) and this finding was replicated in an independent dataset. This supports the notion that attention can be formulated as optimizing the confidence in (or precision of) the inference on sensory input (Friston 2009). In the following, we examine our results in more detail, discuss them in the context of previous work, and outline future extensions.
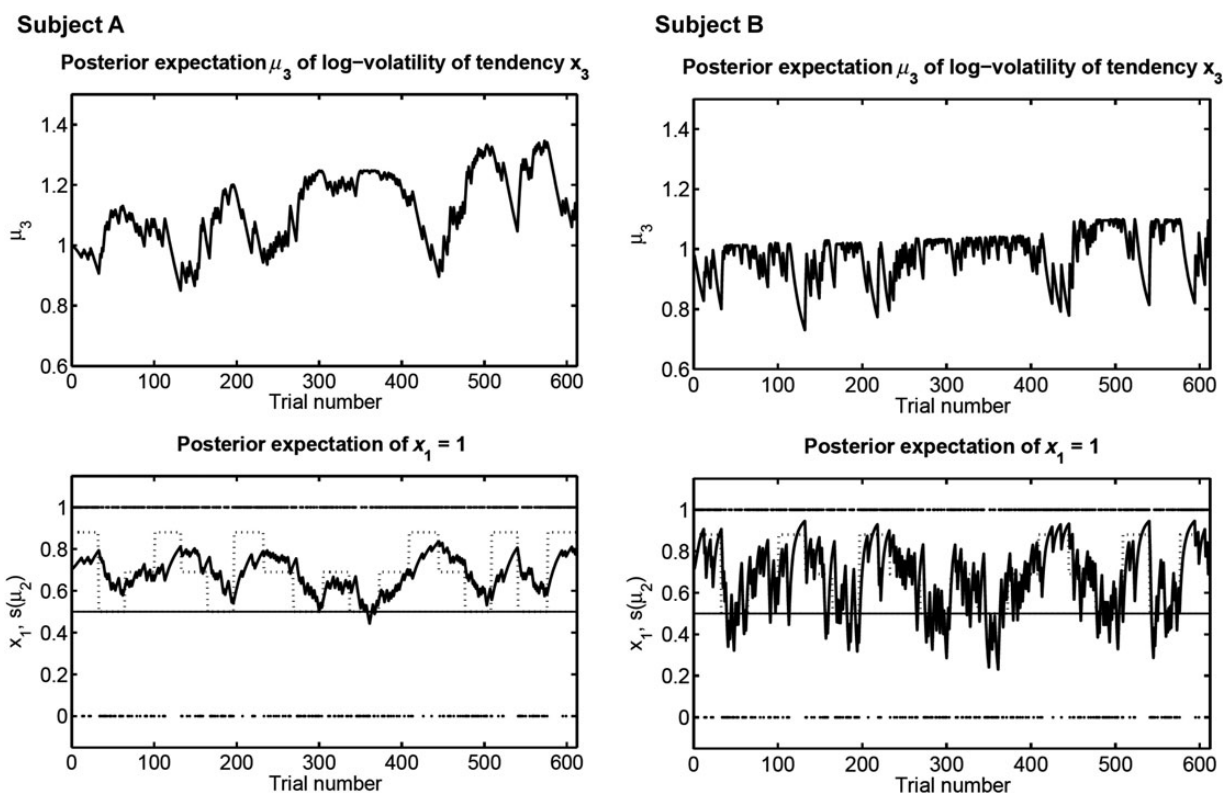
Our experimental paradigm differed from a conventional Posner task, in that the spatial cues predicted the target location with different probabilities at different times during the experiment, thus requiring the subject to infer CV while accounting for environmental volatility. Indeed, a conventional ANOVA showed that the subjects' RS varied as a function of the (unknown) true probabilities, reflecting adaptation to the changing environmental statistics. In other words,

probabilistic context significantly influenced saccadic latencies, although the probabilistic structure of the task was changing in a way that was unknown to subjects.

This relates to previous work in so far as it has been shown that (inverse) saccadic RTs are sensitive to the probability of the saccade target location when abrupt changes in location probability occur within an experimental block (Anderson and Carpenter 2006), or when different blocks employ saccade targets with different probabilities and/or stochastic properties (Brodersen et al. 2008). In contrast to our task, both these studies presented targets without preceding cues, and the latter study also examined learning of sequential (conditional) dependencies between successive stimuli according to a first-order Markov sequence. The present task used explicit cues to elicit spatial attention shifts and investigated how the impact of these cues depended on the subject's current belief (and its precision) about the cue-target contingency. Moreover, instead of presenting different experimental blocks with different probabilistic contexts, here we introduced a volatile environment with frequent but hidden changes of probabilistic context within one continuous trial sequence. A natural modeling framework for explaining the ensuing saccadic reactions is a hierarchical Bayesian learning model— where the subject's belief about the environment's volatility affects the updating of beliefs about the most likely saccade target location. Indeed, comparison of competing perceptual models showed that a full hierarchical perceptual model had higher evidence than reduced models; assuming either that subjects ignored prior knowledge about the volatile nature of

the environment or that they did not use them for updating beliefs about current CV. Moreover, the optimal full hierarchical Bayesian learning model showed higher model evidence than a Rescorla–Wagner learning model or a model which assumes that the subjects knew the true underlying probabilities. Interestingly, however, the subject-specific volatility parameter ω significantly correlated with the learning rate ε of the Rescorla–Wagner model, while no such relationship was observed for the meta-volatility parameter ϑ. The effects of the BMS as well as the relationship to the learning parameter of a Rescorla–Wagner model could be replicated in an independent dataset.

Hierarchical Bayesian models have been used previously to successfully explain various aspects of human behavior under uncertainty, such as binary choices (Behrens et al. 2007) or RTs (den Ouden et al. 2010). These studies, however, assumed an ideal Bayesian observer with no interindividual variation in the learning process per se. In contrast, we followed the meta-Bayesian approach of Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. (2010), Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. (2010) and inferred subject-specific parameters of a Bayes-optimal learning scheme (Mathys et al. 2011) from empirical responses. Our results showed that there is considerable interindividual variability, even within our group of young healthy subjects (cf. Figs. 6 and 8). An obvious and important extension of the present work is to relate this variability to demographic or neurobiological factors. In fact, the work reported here is a prelude to future psychopharmacological and patient studies,
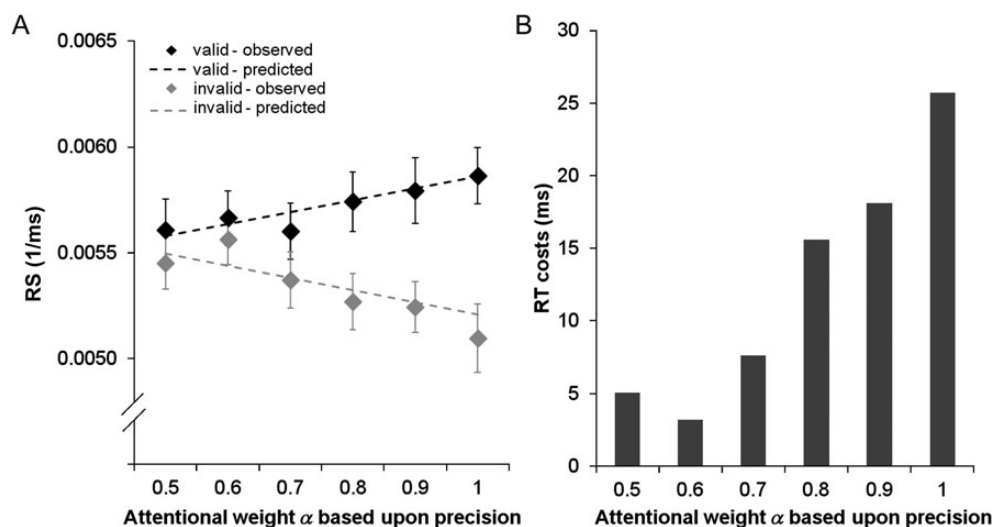
Figure 8. Illustration of the time course of μ_3 (upper panels) and $s(\mu_2)$ (lower panels) during observation of x_1 (black diamonds) for 2 exemplary subjects with different parameters for ω and ϑ. The true %CV is depicted as a dotted line. It can be seen that subject A ($\omega = -6.09$; $\vartheta = 0.97$) shows slower updating of the probability estimate that the target will appear at the cued location than subject B ($\omega = -2.78$; $\vartheta = 0.12$). This can be attributed to subject A's lower value of ω (reflecting the subject's belief in a less volatile environment).

in which we will examine the putative relationship between individual differences in learning and attention (as encoded by our model parameters) and individual differences in neuromodulatory processes (as induced by medication, aging, or disease). In this context, the current results can be seen as trying to establish the construct validity or our paradigm and its modeling.

Moreover, we introduced and tested different response models, i.e., mappings from posterior beliefs provided by the perceptual model to observable behavior. These response models account for the individual variability in the overall speed of responding (Fig. 6B), but differ in assuming whether precision of predictions, strength of the prediction about CV or surprise, respectively, determine saccadic RS. Our results showed that model evidence was highest for the response model in which RS was determined by the precision of the prediction.

In one sense, our findings from the Bayesian model comparison—that precision was the most plausible account for RT benefits—should not be surprising. This is because precision plays the role of a rate constant in evidence accumulation schemes based upon predictive coding (Feldman and Friston 2010). In other words, precision modulates the gain of prediction error in driving changes in conditional representations or expectations. This means that sensory channels that enjoy greater precision will engender faster changes in high-level representations and lead to more rapid perceptual convergence. Behaviorally, this should be manifest in speeded up RTs. Exactly the same theme is seen at higher levels of the hierarchy—that concern slower timescales—such as inference about the probabilistic (trial-to-trial) contingencies, we manipulated in our volatility paradigm. Here, the rate constant corresponds to a learning rate in conventional (reinforcement learning) formulations. In short, sensory evidence and empirical priors that are afforded greater precision have preferential access to higher levels in hierarchical inference. This is expressed as more efficient and faster convergence in those processing streams—and provides a nice metaphor for attention.

In other words, attention corresponds to optimizing estimates of precision in sensory hierarchies and is implemented by changing the postsynaptic gain of neuronal prediction error units. Hence, attention determines which part of the sensorium is treated as furnishing precise information. In this respect, this approach is perfectly congruent with spotlight or zoom lens theories of attention (Posner 1980; Eriksen and St James 1986) as well as with the biased competition model (Desimone and Duncan 1995): the limitation of processing capacities demands a selection of stimulus locations or features so that only the most relevant receive full attention. Neurobiologically, this is likely reflected in increased synaptic gain and neuronal synchronization, manifesting as enhanced firing rates (e.g., Luck et al. 1997) or blood-oxygen-level–dependent responses (e.g., Brefczynski and DeYoe 1999; Kastner et al. 1999) in visual cortex, when attention is directed to a particular spatial location. It may also be noteworthy that, at the synaptic level, precision-dependent synaptic gain (e.g., at superficial pyramidal cells) may be controlled by classical neuromodulators such as dopamine or acetylcholine (Friston 2009). In predictive coding schemes, increased gain boosts the sensitivity of principal cells sending forward afferents to higher levels (such as the intraparietal sulcus [IPS] or the FEF), so that evidence accumulates more rapidly and saccades are elicited more quickly. This notion resonates with findings from several recent studies. For example, Saproo and Serences (2010) showed that spatial attention increases the mutual information of population responses in early visual cortex and suggested that this should enable higher visual areas to read out this information more quickly and efficiently. This is similar to the proposals by Feldman and Friston (2010) and in this article, where higher precision at lower levels induces more rapid changes in the activity of higher level areas. Others have suggested that attention produces behavioral improvements by efficiently selecting the "relevant" sensory signals (Pestilli et al. 2011); the suggested mechanism (focusing on the magnitudes of signals and employing pooling operations) differs in detail

Figure 9. (A) Observed and predicted average RS in valid and invalid trials as a function of the precision-dependent attentional weight parameter α (attention to cued location; calculated for the group average values). Error bars depict standard errors of the mean (SEM). The lines correspond to the predictions using the average response model parameters, over subjects. (B) Illustration of how the observed RS costs after invalid cueing translate into RT differences (in ms).

from mechanisms assumed in Feldman and Friston (a simple modulation of postsynaptic gain) but both call upon nonlinear (pooling and selection) mechanisms. It would be interesting to see whether the results obtained by Pestilli et al. on behavioral contrast-discrimination performance could be replicated when trials are grouped according to precision estimates. Finally, it has been shown that electrical stimulation of direction-selective neurons in MT elicits faster perceptual decisions due to faster evidence accumulation (Ditterich et al. 2003).

According to predictive coding implementations of hierarchical Bayesian inference, the gain of prediction error associated with bottom-up signals corresponds to the precision of those prediction errors. Physiologically, this means that precision may be encoded by the gain of superficial pyramidal cells (Brown and Friston 2012). Accordingly, our computational model would predict that during spatial attention, activity in hierarchically related visual areas should exhibit precision-dependent modulatory effects that result from the enhanced gain of superficial pyramidal cells. This hypothesis —as well as questions about where in the spatial attention/ saccade network precision exerts this effect—could be tested with DCM of electroencephalographic or magnetoencephalographic data (Bastos et al. 2011; Brown and Friston 2012). Interestingly, a recent fMRI study, using a simpler DCM for fMRI, has highlighted the importance of the modulation of inhibitory self-connections in visual areas by attention and prediction (Kok et al. 2012). This type of modulation corresponds (phenomenologically) to a simple gain control mechanism that may reflect the precision-dependent modulation of pyramidal cells described above.

Given the involvement of common areas (FEF and IPS) in both covert attentional orienting of attention and overt eye movements (Corbetta et al. 1998; Nobre et al. 2000; Perry and Zeki 2000; Beauchamp et al. 2001; de Haan et al. 2008), the psychophysical evidence for an inherent link between attention shifts and saccade programming (Deubel and Schneider 1996; Godijn and Theeuwes 2003; Dore-Mazars et al. 2004; Deubel 2008), and the existence of both visual and motor neurons in key structures such as the FEF (e.g., Bruce and Goldberg 1985; Schall and Hanes 1993), it seems plausible that precision should affect both sensory-perceptual as well as motor preparatory processes (cf. the model proposed by Schall et al. 2011). Hence, one could also frame the processes studied here in the broader context of visual-saccadic decision making (see Glimcher 2001, 2003 for comprehensive reviews).

The focus of the present study was on explaining observed trialwise saccadic RS using a generative (hierarchical Bayesian) model and on using model selection to disambiguate among different ways of updating beliefs about upcoming target locations in a volatile environment. While our analyses suggest a precision-based mechanism for spatial attention, it remains to be investigated where these precision estimates are computed within the hierarchical visual attention/saccade network. The present behavioral-modeling results are a foundation for future imaging studies that will exploit the across-trial and between-subject variation in model states and parameters to identify the network of regions in which precision plays a role for belief updating in spatial attention. We

imagine that neuroimaging studies could use the time series of the states of our perceptual model as predictor variables to identify their neuronal correlates (cf. Behrens et al. 2007; den Ouden et al. 2010). Furthermore, as mentioned above, subject-specific estimates of the parameters encoding individual learning style can be used at the between-subject level to reveal the neuronal substrates of interindividual differences.

Conclusion

We have used a new formal framework for characterizing Bayes-optimal trial-by-trial updating of probabilistic beliefs under uncertainty for explaining attentional mechanisms. Specifically, we characterized saccadic RS during an extended Posner paradigm with variable CV. Comparing 11 alternative models, we found that empirical responses are most plausibly explained as a function of precision (of the beliefs about the causes of sensory input). This finding is consistent with attention theories derived from Bayesian theories of brain function (the free-energy principle) that equate spatial attention to a precision-dependent gain modulation of sensory input. Future neuroimaging work could use the modeling approach introduced in this article to identify the neural and neurochemical basis of attentional selection and saccadic eye movements, in relation to probabilistic expectancies.

Supplementary Material

Supplementary material can be found at: http://www.cercor. oxfordjournals.org/.

Funding

Notes

Appendix

Update Equations of the Perceptual Model

The variational inversion method introduced in Mathys et al. (2011) yields closed-form one-step update equations for the sufficient statistics of the posterior distributions representing beliefs about the hidden states x of the agent's environment. In the specific perceptual model depicted in Figure 2, state x_1 is observed, whereas x_2 and x_3 remain hidden. As posteriors are assumed to be Gaussian, the relevant sufficient statistics are the means μ_2, μ_3 and precisions (inverse variances) π_2, π_3 of the distributions for x_2 and x_3. It turns out that the

updates of the means take the form of precision-weighted prediction errors:

$$\Delta\mu_3 = \mu_3^{(t)} - \mu_3^{(t-1)} = \frac{1}{2} v_2^{(t)} \frac{\hat{\pi}_2^{(t)}}{\pi_3^{(t)}} \delta_2^{(t)}, \tag{A1}$$

$$\Delta\mu_2 = \mu_2^{(t)} - \mu_2^{(t-1)} = \sigma_2^{(t)} \delta_1^{(t)} = \frac{\hat{\pi}_1^{(t)}}{\hat{\pi}_2^{(t)} \hat{\pi}_1^{(t)} + 1} \delta_1^{(t)}, \tag{A2}$$

where for clarity and to reveal the conceptual meaning of these equations, we make use of the following definitions:

$$v_2^{(t)} \underset{\text{def}}{=} \exp(\mu_3^{(t-1)} + \omega), \tag{A3}$$

$$\hat{\pi}_2^{(t)} \underset{\text{def}}{=} \frac{1}{\sigma_2^{(t-1)} + v_2^{(t)}}, \tag{A4}$$

$$\delta_2^{(t)} \underset{\text{def}}{=} \frac{\sigma_2^{(t)} + (\mu_2^{(t)} - \mu_2^{(t-1)})^2}{\sigma_2^{(t-1)} + v_2^{(t)}} - 1, \tag{A5}$$

$$\hat{\pi}_1^{(t)} \underset{\text{def}}{=} \frac{1}{s\underbrace{(\mu_2^{(t-1)})}_{\underset{\text{def}}{=}\hat{\mu}_1^{(t)}}(1 - s(\mu_2^{(t-1)}))} = \frac{1}{\hat{\mu}_1^{(t)}(1 - \hat{\mu}_1^{(t)})}, \tag{A6}$$

$$\delta_1^{(t)} \underset{\text{def}}{=} x_1^{(t)} - \hat{\mu}_1^{(t)}. \tag{A7}$$

$v_2^{(t)}$ is the agent's estimate of the variance of the random walk in x_2 before receiving input $x_1^{(t)}$; $\hat{\pi}_i^{(t)}$ are the precisions of the predictions about states $x_i^{(t)}$ and $\delta_i^{(t)}$ are the prediction errors. Note that $\delta_2^{(t)}$ is a prediction error referring not to the value of x_2 but to its log volatility; therefore, it is determined by the ratio of observed to predicted total variance in x_2.

It is obvious from equations (A1) and (A2) that updates are always proportional to the prediction error about the input from the level below δ_{i-1} and to the precision $\hat{\pi}_{i-1}$ of the prediction about the state at the level below.

The update equations for the precisions are

$$\pi_3^{(t)} = \hat{\pi}_3^{(t)} + \frac{1}{2}(v_2^{(t)}\hat{\pi}_2^{(t)})^2 \left(1 + \left(1 - \frac{\sigma_2^{(t-1)}}{v_2^{(t)}}\right)\delta_2^{(t)}\right), \tag{A8}$$

$$\pi_2^{(t)} = \hat{\pi}_2^{(t)} + \frac{1}{\hat{\pi}_1^{(t)}} \tag{A9}$$

with

$$\hat{\pi}_3^{(t)} \underset{\text{def}}{=} \frac{1}{\sigma_3^{(t-1)} + \vartheta}. \tag{A10}$$

For the derivation of these equations, see Mathys et al. (2011).

Response Models

In the following, we explain the functional form of our 3 response models in more detail. All models assume a linear

relationship between α and RS, parameterized by the 2 parameters ζ_1 and ζ_2 (cf. eq. 5 and Fig. 3). α represents the proportion of total attentional capacity that is allocated to the cued location (and therefore lies in the unit interval) and should amount to 0.5 if both target locations are equally likely. These constraints, which all response models conform to, can be summarized as:

C1: $\quad 0 \le \alpha \le 1,$

C2: $\quad \alpha = 0.5$ for $\hat{\mu}_1 = 0.5.$

Given these constraints, our response models differ in which attribute of the predicted validity of the cue maps to the attentional factor α (and thus determines RS in eq. 5). The functional forms of these models are motivated in the following and are depicted graphically in Figure 4. (Note that the vertical axis in Fig. 4 is attention to outcome location. For valid trials, this is equal to attention to cued location α, while for invalid trials it is $1 - \alpha$.)

The "precision" model (eq. 6) links attention to the precision of predictions as suggested by Feldman and Friston (2010). In our specific case, the precision of the prediction at the first level ($\hat{\pi}_1$) has a minimal value of 4 when $\hat{\mu}_1 = 0.5$ and approaches infinity as $\hat{\mu}_1$ approaches 1 (cf. eq. A6). The most parsimonious way to meet the above constraints C1 and C2 is to define α as the logistic sigmoid of $\hat{\pi}_1$, minus its minimum (cf. eq. 6):

$$\alpha = s(\hat{\pi}_1 - 4). \tag{A11}$$

Note that as the cue becomes a counter indication of outcome location when μ_2 falls below 0 (or equivalently, when $\hat{\mu}_1$ drops below 0.5), a suitable definition of α for the whole range of $\hat{\mu}_1$ is

$$\alpha^{(t)} = s(\text{sign}(\mu_2^{(t-1)})(\hat{\pi}_1^{(t)} - 4)). \tag{A12}$$

This ensures that attention to the cued location falls to 0 as $\hat{\mu}_1$ approaches 0.

A simpler model of attention allocation given a cue-induced belief about outcome is that attention is proportional to predicted probability of outcome: if the agent believes that the probability of seeing outcome "left" is P (e.g., 80%), then it will allocate proportion P (i.e., 80%) of its attentional resources to location left. We call this the belief model (cf. eq. 7). In terms of our perceptual model, the predicted probability of a valid trial is simply $\hat{\mu}_1$:

$$\alpha = \hat{\mu}_1. \tag{A13}$$

Finally, the surprise model describes the attentional factor α as a function of the Shannon surprise (the negative logarithm of the probability of the outcome being the cued location given the prediction $\hat{\mu}_1$) (cf. Bestmann et al. 2008). For a predicted probability $\hat{\mu}_1 = 1$ of a valid trial, surprise is zero, whereas for $\hat{\mu}_1 = 0$, it is infinite. In the first case, attention is therefore allocated exclusively to the cued location (i.e., $\alpha = 1$), whereas in the second case, attention is allocated exclusively to the noncued location (i.e., $\alpha = 0$). The simplest way this can be achieved under consideration of constraints

C1 and C2 is (cf. eq. 8):

$$\alpha^{(t)} = \frac{1}{1 + \text{surprise}(\hat{\mu}_1^{(t)})}, \qquad (A14)$$

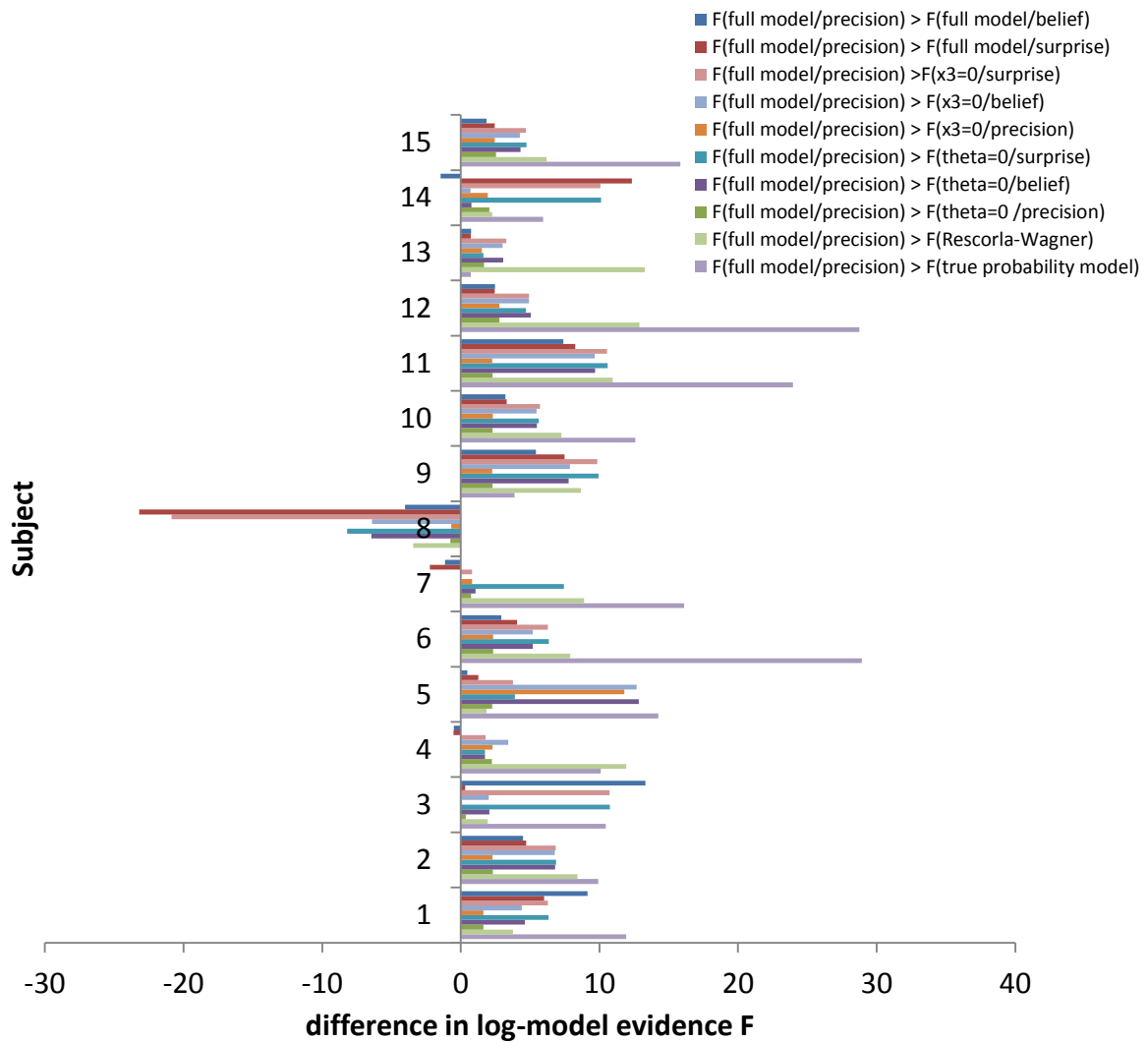where $\text{surprise}(\hat{\mu}_1^{(t)}) = -\log_2 p(x_1^{(t)} = 1 | \hat{\mu}_1^{(t)}) = -\log_2(\hat{\mu}_1^{(t)})$. Note that we make use of $\log_2 x = \ln x / \ln 2$ to ensure we meet constraint C2.

References

Anderson AJ, Carpenter RHS. 2006. Changes in expectation consequent on experience, modeled by a simple, forgetful neural circuit. J Vis. 6:822–835.

Bastos A, Moran R, Litvak V, Fries P, Friston K, 2011. A Dynamic Causal Model of how inter-areal synchronization is achieved in canonical microcircuits. Program No. 622.16. NeuroscienceMeeting Planner. Society for Neuroscience, Washington, DC. (Online).

Beauchamp MS, Petit L, Ellmore TM, Ingeholm J, Haxby JV. 2001. A parametric fMRI study of overt and covert shifts of visuospatial attention. Neuroimage. 14:310–321.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. 2007. Learning the value of information in an uncertain world. Nat Neurosci. 10:1214–1221.

Bestmann S, Harrison LM, Blankenburg F, Mars RB, Haggard P, Friston KJ, Rothwell JC. 2008. Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. Curr Biol. 18:775–780.

Brefczynski JA, DeYoe EA. 1999. A physiological correlate of the 'spotlight' of visual attention. Nat Neurosci. 2:370–374.

Brodersen KH, Penny WD, Harrison LM, Daunizeau J, Ruff CC, Duzel E, Friston KJ, Stephan KE. 2008. Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw. 21:1247–1260.

Brown H, Friston KJ. 2012. Dynamic causal modelling of precision and synaptic gain in visual perception – an EEG study. Neuroimage. 63:233–231.

Bruce CJ, Goldberg ME. 1985. Primate frontal eye fields. I. Single neurons discharging before saccades. J Neurophysiol. 53:603–635.

Carpenter RHS, Williams MLL. 1995. Neural computation of log likelihood in control of saccadic eye movements. Nature. 377:59–62.

Chiau H-Y, Tseng P, Su J-H, Tzeng OJL, Hung DL, Muggleton NG, Juan C-H. 2011. Trial type probability modulates the cost of antisaccades. J Neurophysiol. 106:515–526.

Chikkerur S, Serre T, Tan C, Poggio T. 2010. What and where: a Bayesian inference theory of attention. Vision Res. 50:2233–2247.

Corbetta M, Akbudak E, Conturo TE, Snyder AZ, Ollinger JM, Drury HA, Linenweber MR, Petersen SE, Raichle ME, Van Essen DC et al. 1998. A common network of functional areas for attention and eye movements. Neuron. 21:761–773.

Daunizeau J, David O, Stephan KE. 2011. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. Neuroimage. 58:312–322.

Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Friston KJ, Stephan KE. 2010. Observing the observer (II): deciding when to decide. PLoS ONE. 5:e15555.

Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. 2010. Observing the observer (I): meta-Bayesian models of learning and decision-making. PLoS ONE. 5:e15554.

Daunizeau J, Friston KJ, Kiebel SJ. 2009. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Physica D. 238:2089–2118.

Dayan P. 2000. Learning and selective attention. Nat Rev Neurosci. 3:1218–1223.

Dayan P, Hinton GE, Neal RM, Zemel RS. 1995. The Helmholtz machine. Neural Comput. 7:889–904.

De Haan B, Morgan BS, Rorden C. 2008. Covert orienting of attention and overt eye movements activate identical brain regions. Brain Res. 1204:102–111.

den Ouden HEM, Daunizeau J, Roiser J, Friston KJ, Stephan KE. 2010. Striatal prediction error modulates cortical coupling. J Neurosci. 30:3210–3219.

Desimone R, Duncan J. 1995. Neural mechanisms of selective visual attention. Ann Rev Neurosci. 18:193–222.

Deubel H. 2008. The time course of presaccadic attention shifts. Psychol Res. 72:630–640.

Deubel H, Schneider WX. 1996. Saccade target selection and object recognition: evidence for a common attentional mechanism. Vis Res. 36:1827–1837.

Ditterich J, Mazurek ME, Shadlen MN. 2003. Microstimulation of visual cortex affects the speed of perceptual decisions. Nat Neurosci. 6:891–898.

Dore-Mazars K, Pought P, Beauvillain C. 2004. Attentional selection during preparation of eye movements. Psychol Res. 69:67–76.

Eriksen CW, St James JD. 1986. Visual attention within and around the field of focal attention: a zoomlens model. Percept Psychophys. 40:225–240.

Eriksen CW, Yeh YY. 1985. Allocation of attention in the visual field. J Exp Psychol Human Percep Perform. 11:583–587.

Farrell S, Ludwig CJH, Ellis LA, Gilchrist ID. 2010. Influence of environmental statistics on inhibition of saccadic return. PNAS. 107:929–934.

Feldman H, Friston KJ. 2010. Attention, uncertainty, and free-energy. Front Hum Neurosci. 4:215.

Fischer B, Biscaldi M, Otto P. 1993. Saccadic eye movements of dyslexic adult subjects. Neuropsychologia. 31:887–906.

Friston KJ. 2009. The free energy principle: a rough guide to the brain? Trends Cogn Sci. 13:293–301.

Friston KJ. 2010. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 11:127–138.

Friston KJ, Harrison L, Penny W. 2003. Dynamic causal modelling. Neuroimage. 19:1273–1302.

Friston KJ, Kilner J, Harrison L. 2006. A free energy principle for the brain. J Physiol Paris. 100:70–87.

Friston KJ, Mattout J, Trujillo-Barreto N, Ashburner J, Penny WD. 2007. Variational free energy and the Laplace approximation. Neuroimage. 34:220–234.

Gershman SJ, Niv Y. 2010. Learning latent structure: carving nature at its joints. Curr Opin Neurobiol. 20:251–256.

Giessing C, Thiel CM, Fink GR. 2006. The modulatory effects of nicotine on parietal cortex activity in a cued target detection task depend upon cue reliability. Neuroscience. 137:853–864.

Gitelman DR. 2002. ILAB: a program for postexperimental eye movement analysis. Behav Res Methods Instrum Comput. 34:605–612.

Glimcher PW. 2001. Making choices: the neurophysiology of visual-saccadic decision making. TiNS. 24:654–659.

Glimcher PW. 2003. The neurobiology of visual-saccadic decision making. Ann Rev Neurosci. 26:133–179.

Godijn R, Theeuwes J. 2003. Parallel allocation of attention prior to the execution of saccade sequences. J Exp Psychol Human Percep Perform. 29:882–896.

Itti L, Baldi P. 2009. Bayesian surprise attracts human attention. Vision Res. 49:1295–1306.

Jongen EMM, Smulders FTY. 2007. Sequence effects in a spatial cueing task: endogenous orienting is sensitive to orienting in the preceding trial. Psychol Res. 71:516–523.

Jonides J. 1980. Towards a model of the mind's eye's movement. Can J Psychol. 34:103–112.

Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG. 1999. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron. 22:751–761.

Kok P, Rahnev D, Jehee JFM, Lau HC, de Lange FP. 2012. Attention reverses the effect of prediction in silencing sensory signals. Cereb Cortex. 22:2197–2206.

Luck SJ, Chelazzi L, Hillyard SA, Desimone R. 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J Neurophysiol. 77:24–42.

Mathys C, Daunizeau J, Friston KJ, Stephan KE. 2011. A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci. 5:39.

Nobre AC, Gitelman DR, Dias EC, Mesulam MM. 2000. Covert visual spatial orienting and saccades: overlapping neural systems. Neuroimage. 11:210–216.

Penny WD, Kiebel S, Friston KJ. 2007. Variational Bayes. In: Friston KJ, Ashburner J, Kiebel S, Nichols T, Penny WD, editors. Statistical parametric mapping: the analysis of functional brain images. London: Academic Press. p. 303–312.

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP. 2010. Comparing families of dynamic causal models. PLoS Comput Biol. 6:e1000709.

Perry RJ, Zeki S. 2000. The neurology of saccades and covert shifts in spatial attention: an event-related fMRI study. Brain. 123: 2273–2288.

Pestilli F, Carrasco M, Heeger DJ, Gardner JL. 2011. Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. Neuron. 72:832–846.

Posner MI. 1980. Orienting of attention. Q J Exp Psychol. 32:3–25.

Rao RP. 2005. Bayesian inference and attentional modulation in the visual cortex. Neuroreport. 16:1843–1848.

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement. In: Black AH, Prokasy WF. editors. Classical conditioning II: current research and theory. New York: Appleton-Century-Crofts. p. 64–99.

Risko EF, Stolz JA. 2010. The proportion valid effect in covert orienting: strategic control or implicit learning? Conscious Cogn. 19:432–442.

Rizzolatti G, Riggio L, Dascola I, Umilta C. 1987. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. Neuropsychologia. 25:31–40.

Rushworth MF, Behrens TEJ. 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Rev Neurosc. 11:389–397.

Saproo S, Serences JT. 2010. Spatial attention improves the quality of population codes in human visual cortex. J Neurophysiol. 104:885–895.

Schall JD, Hanes DP. 1993. Neural basis of saccade target selection in frontal eye field during visual search. Nature. 366:467–469.

Schall JD, Purcell BA, Heitz RP, Logan GD, Palmeri TJ. 2011. Neural mechanisms of saccade target selection: gated accumulator model of the visual–motor cascade. Eur J Neurosci. 33:101–2002.

Schneider WX. 1995. VAM: a neuro-cognitive model for attention control of segmentation, object recognition and space-based motor action. Vis Cogn. 2:331–374.

Shannon CE. 1948. A mathematical theory of communication. Bell Syst Techn J. 27:379. –423, 623–656.

Stampe DM. 1993. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behav Res Methods Instrum Comput. 25:137–142.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. 2009. Bayesian model selection for group studies. Neuroimage. 46:1004–1017.

Vossel S, Weidner R, Fink GR. 2011. Dynamic coding of events within the inferior frontal gyrus in a probabilistic selective attention task. J Cogn Neurosci. 23:414–424.

Whiteley L, Sahani M. 2008. Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. J Vis. 8:1–15.

Yu A, Dayan P. 2005. Uncertainty, neuromodulation, and attention. Neuron. 46:681–692.

Supplementary Figure S1. Differences in log-model evidence for the winning model as compared to alternative models. Positive values reflect higher model evidence for the full Bayesian model based upon precision.