PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Modeling the Evolution of Beliefs Using an Attentional Focus Mechanism

Dimitrije Marković[1,2]*, Jan Gläscher[3,4], Peter Bossaerts[4,5,6], John O'Doherty[4,6,7], Stefan J. Kiebel[1,2]

1 Department of Psychology, Technical University Dresden, Dresden, Germany, **2** Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, **3** Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, **4** Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, California, United States of America, **5** Department of Finance, University of Utah, Salt Lake City, United States of America, **6** Computation and Neural Systems, California Institute of Technology, Pasadena, California, United States of America, **7** Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

* dimitrije.markovic@tu-dresden.de

Citation: Marković D, Gläscher J, Bossaerts P, O'Doherty J, Kiebel SJ (2015) Modeling the Evolution of Beliefs Using an Attentional Focus Mechanism. PLoS Comput Biol 11(10): e1004558. doi:10.1371/journal.pcbi.1004558

Editor: Wolfgang Einhäuser, Technische Universitat Chemnitz, GERMANY

Received: December 16, 2014

Accepted: September 1, 2015

Published: October 23, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the US-German Collaboration in Computational Neuroscience of NSF (1207573, to JO) and BMBF (Förderkennzeichen: 01GQ1205, to SJK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

For making decisions in everyday life we often have first to infer the set of environmental features that are relevant for the current task. Here we investigated the computational mechanisms underlying the evolution of beliefs about the relevance of environmental features in a dynamical and noisy environment. For this purpose we designed a probabilistic Wisconsin card sorting task (WCST) with belief solicitation, in which subjects were presented with stimuli composed of multiple visual features. At each moment in time a particular feature was relevant for obtaining reward, and participants had to infer which feature was relevant and report their beliefs accordingly. To test the hypothesis that attentional focus modulates the belief update process, we derived and fitted several probabilistic and non-probabilistic behavioral models, which either incorporate a dynamical model of attentional focus, in the form of a hierarchical winner-take-all neuronal network, or a diffusive model, without attention-like features. We used Bayesian model selection to identify the most likely generative model of subjects' behavior and found that attention-like features in the behavioral model are essential for explaining subjects' responses. Furthermore, we demonstrate a method for integrating both connectionist and Bayesian models of decision making within a single framework that allowed us to infer hidden belief processes of human subjects.

Author Summary

When making decisions in our everyday life (*e.g.* where to eat) we first have to identify a set of environmental features that are relevant for the decision (e.g. the distance to the place, current time or the price). Although we are able to make such inferences almost effortlessly, this type of problems is computationally challenging, as we live in a complex environment that constantly changes and contains an immense number of features. Here we investigated the question of how the human brain solves this computational challenge.

In particular, we designed a new experimental paradigm and derived novel behavioral models to test the hypothesis that attention modulates the formation of beliefs about the relevance of several environmental features. As each behavioral model accounted for a different hypothesis about the underlying computational mechanism we compared them in their ability to explain the measured behavior of human subjects performing the experimental task. The model comparison indicates that an attentional-focus mechanism is a key feature of behavioral models that accurately replicate subjects' behavior. These findings suggest that the evolution of beliefs is modulated by a competitive attractor dynamics that forms prior expectation about future outcomes. Hence, the findings provide interesting and novel insights into the computational mechanisms underlying human behavior when making decisions in complex environments.

Introduction

A typical problem that humans encounter, in our complex environment, is to identify those environmental features that are relevant for achieving a desired outcome in a given task. This is computationally difficult because the real-world environment displays a large number of environmental features. In addition, the relevance of the features can change over time and the observations do not always reflect the relevance of specific features. For example, to increase the chance of catching a fish, a fisherman has to consider various features (*e.g.* time of the day, lightening conditions, water transparency, *etc.*). Depending on the fishing place (*e.g.* pond, lake, or river) only some of these features will be relevant. To perfectly solve such tasks all possible features should be taken into account simultaneously. However, due to an apparent limitation in their cognitive resources, humans dynamically attend only to the most relevant environmental features when deciding what action to pursue [1,2]. Our goal here is to develop a computational model to analyze behavioral data and understand better how attention modulates the update of beliefs about the relevance of features in such complex environments.

An ideal test bed to address these questions is the Wisconsin card sorting task (WCST), as it provides an experimental environment with multiple visual features, in which at any moment of time only a single feature is relevant for correctly solving the task. The WCST was originally designed to test for the damage or dysfunction of the prefrontal cortex, which regulates executive functions [3–6]. More recently it was employed in various behavioral models as a paradigm with which one can investigate computational mechanisms of higher cognitive functions [7].

Here we will focus on the computational mechanisms that underlie update of beliefs about the relevance of various visual features. However, inferring the hidden belief states of subjects performing the standard WCST is difficult, as the only expression of an internal, multidimensional belief space are the behavioral choices [1,8–10]. To address this issue we designed a probabilistic variant of WCST in which we solicited subjects' beliefs [11], that is, we requested from subjects to bet an amount of money proportionally to their beliefs about the relevance of each visual feature. Importantly, various sources of uncertainty made the environment of WCST probabilistic and made the task more difficult, thus allowing us to measure smooth belief trajectories that evolve over single trials. This fine-grained measure provides more direct access to subjects' hidden belief states and thus allowed for improved inference, compared to the standard WCST. Using this novel variant of the WCST, we were able to develop a

probabilistic model for the analysis of behavioral data to provide novel insights into the hidden learning mechanism, which drives human behavior [12–14].

Previous computational models for the WCST can be divided into three groups based on the assumed computational principle that were used to capture human behavior and cognition: (i) functional cognitive models [10], which are motivated by algorithmic properties of the task; (ii) connectionist models [9,15–19], which are motivated by the evidence that the brain is an active and distributed system that constantly generates hypotheses about its environment and tests for their validity [20–25]; and probabilistic Bayesian models [1], which further assume that the brain combines prior knowledge and present sensory information based on their relative precision, that is, in a Bayes-optimal manner [26–32].

The classical connectionist approach provides an elegant framework for defining attention formation in a distributed and dynamical manner. A potential limitation is that one requires additional and rather ad-hoc assumptions to describe the interaction of prediction errors with internal dynamics of beliefs. This issue can be addressed by the Bayesian approach which provides a framework for defining optimal interaction between prediction errors and current belief states. Furthermore, the Bayesian framework provides a computational account of attention [33–36], which the connectionist approach lacks. Here we build upon these past views of attention within the Bayesian framework, with an attentional focus mechanism that relies on competitive and self-organized dynamical principles that guide spontaneous formation of attention. We will fuse the winner-take-all (WTA) dynamics [37–43] with a Bayesian formalism of decision making.

With this combined approach we can investigate, at the same time, the influence of attention and the influence of probabilistic aspects of the environment on the evolution of beliefs during decision making. In addition, this framework allows us to relate our investigation to previous findings of a presumed hierarchical representation in the brain [12,14,44–48]. Importantly, the introduction of such an attentional focus mechanism within a Bayesian framework takes the model away from the rational Bayesian observer that is fully informed about the structure of the probabilistic WCST and which updates beliefs about all features independent of their relevance. However, we expect an attentional focus mechanism to provide a better account for experimentally observed human behavior.

To test whether subjects' behavior reflects the assumption that the update of beliefs is modulated by attentional focus we compared multiple variants of the behavioral models, both with and without an attentional focus mechanism, in their ability to generate behavioral data. In particular, we used a recently described meta-Bayesian approach, the so-called 'Observing the observer' (OTO) framework to infer the hidden belief states and their influence on behavioral responses of human subjects [49,50]. Importantly, using the OTO framework enabled us to put perception and action (i.e., subjects' responses) into a single behavioral model and to compare various variants of both perceptual and response models. Each variant of the perceptual model tested for different assumptions about the mechanisms that underlie the update of beliefs. Similarly different variants of the response model tested for evidence regarding sub-optimality in human decision making, caused by a potentially stochastic representation of posterior beliefs in the brain [51–53].

In what follows, we will first describe the experimental paradigm, briefly introduce the OTO framework, and derive the update equations of several variants of the behavioral models. Then we will describe the data analysis technique that relies on Bayesian model selection using a random effects metric [54,55], and present the results of the analysis that we performed on a behavioral, multi-subject, data set obtained from a probabilistic WCST paradigm. In the last section of the article we discuss the relevance of the proposed attentional-focus mechanism and its relation to past works.

Methods

In this section we will first describe the experimental task, a probabilistic Wisconsin card sorting task with belief solicitation. Afterwards, we will give a brief description of the OTO framework 'Observing the observer' [49] and we will introduce the variants of perceptual and response models that we used to model the update of the hidden belief states and the corresponding solicited responses. Finally, we will outline the methods that we applied to estimate the posterior distribution of model parameters and the corresponding model evidence, which we used to perform Bayesian model comparison.

Ethics statement

The experiment was approved by the Caltech Institutional Review Board and all subjects gave informed consent before participating in the study.

Probabilistic Wisconsin card sorting task

We designed the experimental task with the aim to access the hidden belief states of the subjects. For this purpose we instructed the subjects to infer, by observing a series of an experimenter's choices, which one of the three different visual features is relevant for the current choice, and to report their beliefs about the relevance of each of the features. Participants in the experiment were all healthy volunteers recruited from the Caltech student population.

The visual stimuli that we presented to subjects consisted of a pair of cards (top and bottom), where each card contained three visual features (color, motion, shape). In turn, each visual feature was represented by one of the two possible exemplars (red-green, left-right, circle-square). As each card had to contain a distinct exemplar, there were eight distinct configurations of card pairs. Thus at each experimental trial the visual stimulus was randomly selected from one of the eight configurations (e.g., a red right-moving circle and green left-moving square; see Fig 1A).

Each out of $n = 22$ pre-trained subjects (14 male and 8 female) was exposed to an experimental session divided into six blocks consisting of $T = 40$ trials each. In three randomly selected blocks the relevant feature remained fixed (no-switch condition), whereas in the other

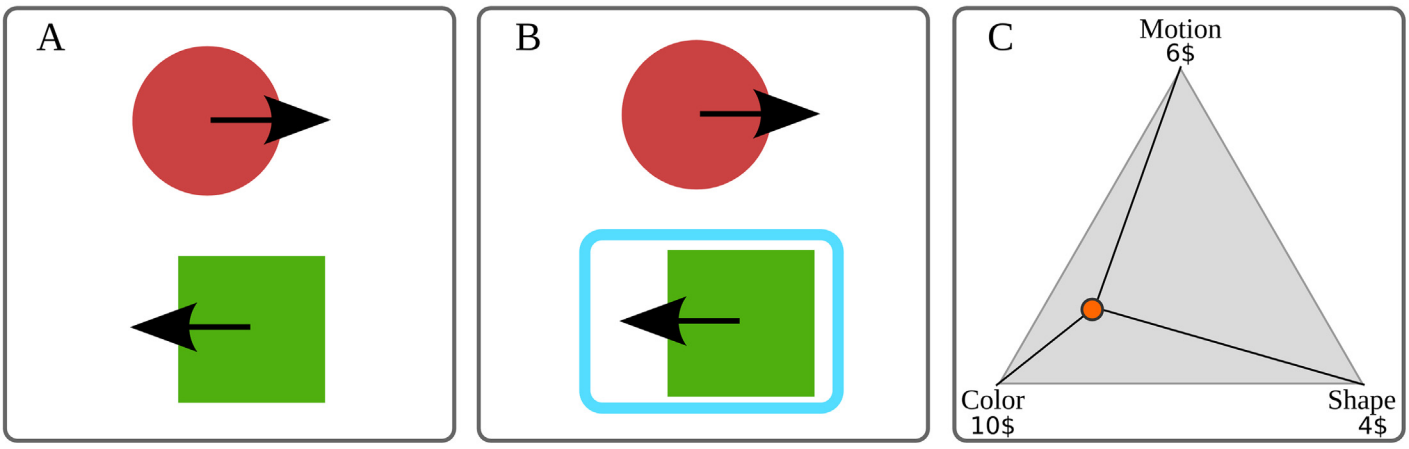

Fig 1. Experimental design. A trial consists of three subsequent steps: (A) The visual stimuli shown in a single trial as two cards. Note that each of the three visual features (color, shape, and motion) has two exemplars (e.g red and green for color) which are assigned either to the top or to the bottom card. (B) The experimenter selects one of the cards, here shown as a blue rectangle. (C) The subject distributes 20$ over three visual features by moving a cursor (red circle) within a triangle. The closer the cursor was to one of the corners of the triangle, the more money was assigned to the corresponding visual feature.

doi:10.1371/journal.pcbi.1004558.g001

three blocks the relevant feature would change with a probability $p = 0.35$ (switch condition). After each switch the relevant feature would remain constant for 8 trials before another switch could occur. Importantly, to make the otherwise quite simple task more difficult for healthy subjects we introduced observation uncertainty: the experimenter would select a wrong card (a card not containing the relevant exemplar) with probability $\varepsilon = 0.2$ in the no-switch condition, and with probability $\varepsilon = 0.3$ in the switch condition. The error rate ε was set to values that induced the most distinct behavioral responses between two experimental conditions, while rendering the switch condition informative enough to induce betting responses in subjects.

At the beginning of each experimental block we informed the subjects about the block type, but we did not inform them about the exact values of the error rates ε or switch probabilities; they had to infer these probabilities during the training phase. Each subject went through three training sessions, where each subsequent session slightly increased the difficulty of the task in the following manner: In the first session subjects were exposed to a no switch environment with error rate of experimenters choices set to zero. In the second session the switches in the selection rule where announced with error rate still being set at zero. The third session consisted of the no-switch environment with $\varepsilon = 0.2$. Afterwards, we explained to subjects the condition in the final switch environment with non-zero error rate.

During a single trial subjects were first exposed to one of the eight possible visual stimuli (see Fig 1A). After one second the presentation program would select a card containing the relevant exemplar with probability $1 - \varepsilon$ (see Fig 1B). After observing the selected choice for 5 seconds subjects had a 4 second period to respond by distributing 20$ on the three visual features depending on their belief about the relevance of each feature for the selection process. The response was generated by moving a cursor within a triangle presented on the screen (see Fig 1C). The closer the cursor was to one of the corners of the triangle the more money was assigned to the corresponding visual feature. Importantly, subjects were told that at the end of the experiment a single trial will be randomly selected and that subjects will gain the amount of money that they assigned to the relevant feature in that trial. This ensured that participants were motivated to provide an accurate rendering of their beliefs over the features.

For clarification of the task we present at this point some of the key behavioral results (see Fig 2). We quantified the performance of subjects as the median amount of their money bets on a truly relevant visual feature over an experimental block. The maximal performance would correspond to betting the full amount of 20$ to the truly relevant feature at each trial. As expected, the median of subjects' performance was higher during the no-switch condition (Kruskal-Wallis test, $p < 10^{-14}$), whereas the median reaction times were lower (Kruskal-Wallis test, $p < 10^{-12}$) during the same experimental condition which reflects the increased difficulty of the switch condition.

'Observing the Observer' framework

Our goal is to infer, from the behavioral data, the hidden belief states of each subject that are conditioned on the past sequence of visual stimuli and experimenter choices. By deriving an adequate mapping of observations onto internal belief states (the perceptual model) and the mapping of the internal belief states onto desired responses (response model), we can define a generative model of the whole observation-response process [49,50] as (see Fig 3 for a graphical representation):

$$p(\vec{r}_t, \gamma, \theta | \vec{e}_t, m^{(p)}, m^{(r)}) = p(\vec{r}_t | b_t(b_{t-1}, \vec{e}_t, \gamma),\ \theta, m^{(p)},\ m^{(r)}) p(\gamma, \theta | m^{(r)}, m^{(p)}), \quad (1)$$

where $p(\vec{r}_t | b_t(b_{t-1}, \vec{e}_t, \gamma),\ \theta, m^{(p)}, m^{(r)})$ denotes the probability of observing a response $\vec{r}_t$ given the hidden belief states b_t (that depend on past beliefs, current sensory observations $\vec{e}_t$, and a

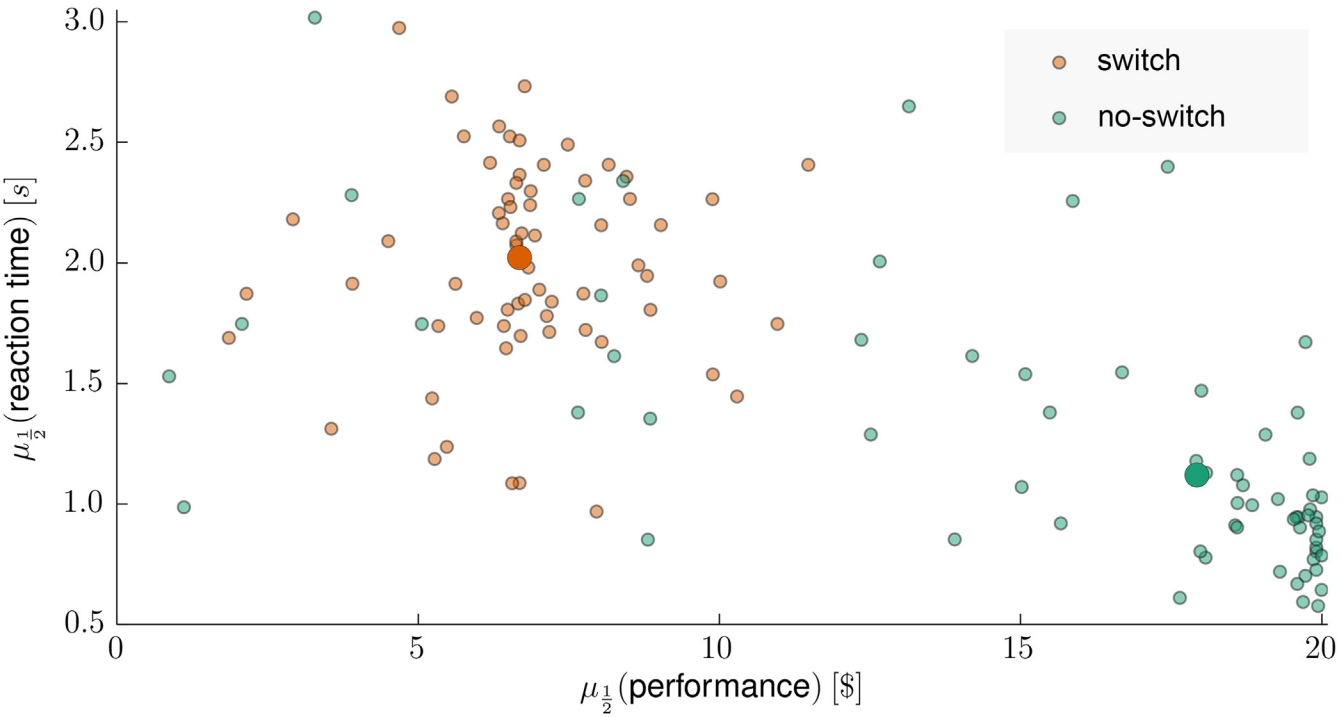

Fig 2. Reaction times and task performance. Median reaction time plotted against median performance of 22 subjects for each of three experimental blocks of the switch (orange circles) and no-switch condition (green circles). The two large circles denote the median values across all experimental blocks within the two experimental conditions. We defined the median performance as the median money gain within an experimental block, that is, the median amount of money assigned to the truly relevant visual feature within an experimental block.

doi:10.1371/journal.pcbi.1004558.g002

set γ of free parameters of the perceptual model $m^{(p)}$) and a set θ of free parameters of the response model $m^{(r)}$. The last term $p(\gamma, \theta|m^{(r)}, m^{(p)})$ in Eq (1) denotes a prior distribution over the space of free parameters.

Thus, to infer the hidden belief states of a subject we have to invert the generative model (Eq (1)) for the given set of behavioral responses $r_{1\ldots t}$ and sensory stimuli $e_{1\ldots t}$, and compute the posterior distribution over the model parameters

$$p(\gamma, \theta|e_{1\ldots t}, r_{1\ldots t}) = \frac{p(\gamma, \theta)\prod_{k=1}^{t} p(\vec{r}_k|\gamma, \theta, e_{1\ldots k})}{p(r_{1\ldots t}|e_{1\ldots t})}, \quad (2)$$

where we omitted $m^{(r)}$, $m^{(p)}$ for better readability. Knowing the posterior distribution one can either compute the most likely belief state at trial t as $b_t(\hat{\gamma})$—where $\hat{\gamma}$ denotes the mode of the posterior—or an expected belief state at trial t, as $\bar{b}_t = E_{p(\gamma|e_{1\ldots t}, r_{1\ldots t})}[b_t(\gamma)]$.

To test the hypothesis that subjects focus their attention on a subset of environmental features when updating their beliefs about the features' relevance, it is essential to compare multiple models in their ability to replicate the behavioral data and select the most appropriate model. Bayesian model comparison uses model evidence, that is, marginal likelihood $p(r_{1\ldots t}|e_{1\ldots t})$, to estimate the probability that a specific model has generated the data. The advantage of such a procedure, compared to standard goodness of fit approaches, is that more complex models are penalized automatically. The model evidence, for any pair of perceptual and

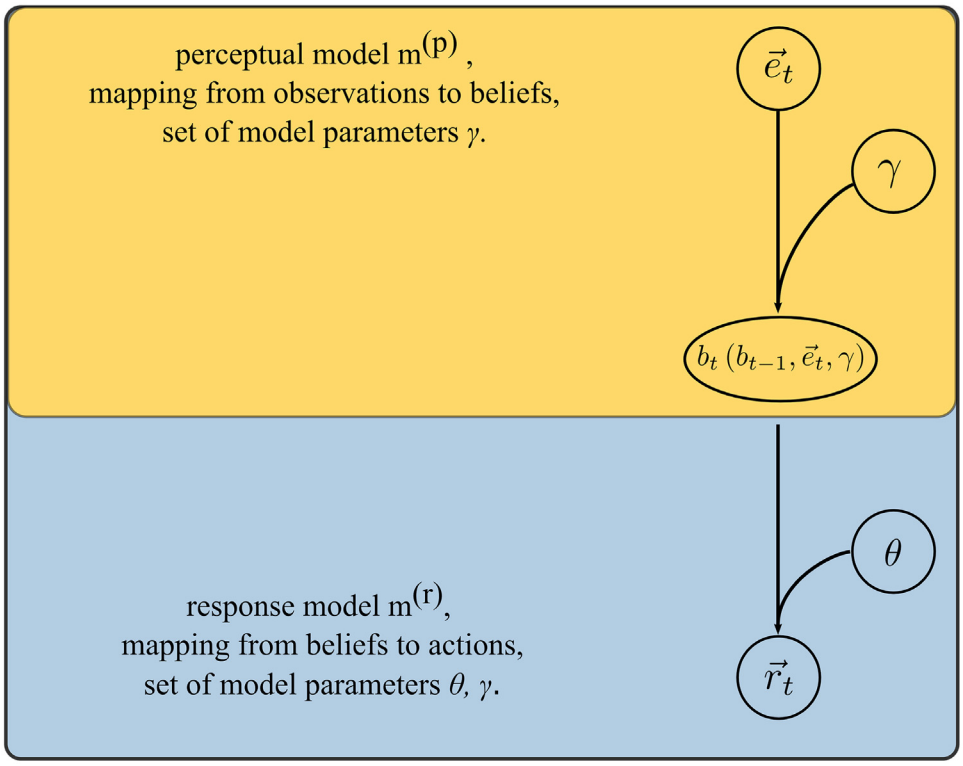

Fig 3. Schematic of implicit generative model as formulated under the Observing-the-observer framework. The full generative model consists of a combined perceptual (orange box) and response model (blue box). The perceptual part of the generative model defines the mapping from current observations $\vec{e}_t$, past beliefs b_{t-1}, and a set of model parameters γ, onto current beliefs b_t. The response part of the generative model defines the mapping from current beliefs b_t and a set of model parameters θ, onto responses $\vec{r}_t$. Figure adapted from [49].

doi:10.1371/journal.pcbi.1004558.g003

response models, is given as

$$p\left(r_{1\ldots t}|e_{1\ldots t},\ m^{(p)}, m^{(r)}\right) = \int d\gamma d\theta p(\gamma,\theta) \prod_{k=1}^{t} p\left(r_k,\ |\gamma, \theta, e_{1\ldots k}, m^{(p)}, m^{(r)}\right). \quad (3)$$

To estimate the model evidence and obtain the posterior distribution over model parameters $p(\gamma,\theta|e_{1\ldots t}, r_{1\ldots t})$ any approximate inference scheme can be applied. In particular, Daunizeau *et. al.* [49,50] proposed the use of a variational scheme where the model log-evidence is approximated with the variational free-energy and the posterior distribution over the model parameters is selected as the maximizer of the free-energy obtained through variational calculus. However, this method requires the computation of the gradients of the log-joint probability distributions (natural logarithm of the joint probability distribution given in Eq (1)), which in our case are not obtainable analytically as the derivatives affect the parameters of the non-linear equations of the belief process. Furthermore, a small change in the parameters of the update equations of beliefs (Eq (11), see below) can have a large influence on the shape of the trajectory, thus the log-joint probability distribution can be ill-conditioned with respect to model parameters. Therefore, even if the gradient, with respect to model parameters, would be computable at every point of the trajectory, a gradient ascent method would have difficulties to converge to a global mode of the joint probability distribution, as the underlying landscape might have a multimodal, non-linear, and non-convex structure.

Thus, we use a numerical gradient-free scheme to find the mode of the log-joint probability distribution and apply a numerical method to compute the Hessian matrix at that mode [56–58]. With the obtained values of the mode and the Hessian we compute the Laplace approximation to the model evidence [59]. We will discuss the specifics of the numerical estimates in the final subsection of the methods. In what follows we will first introduce the behavioral models.

Perceptual model. To derive the perceptual model, which maps sensory cues onto beliefs we followed previous accounts in making three important assumptions [60–62]. First, we will assume that subjects combine prior beliefs and sensory information in a Bayes optimal fashion (Bayesian observer assumption). Note that this assumption will later be relaxed to obtain a non-Bayesian approximation to the update equations. Second, we assume that the update of beliefs can be represented as a Markov process, that is, future belief states depend only on the present beliefs. Third, we will assume that subjects perform counterfactual inference [35], that is, they try to infer which of the several hypothesis (explanations of experimenter's choices) is currently correct. A single hypothesis would correspond to saying that the experimenter selects cards containing a specific exemplar (e.g. color red). As each visual feature has two exemplars (red-green color, leftward-rightward motion, and round-square shape), there are in total six hypotheses.

Starting with these three assumptions we will define a generative model of the sensory observations in the form of a hierarchical state space model [63], that captures the dynamics of the transient probability that one of the six possible selection rules is currently active. Inversion of the generative model will provide us with the required mapping from sensory cues onto posterior probability about the correctness of each hypothesis, that is, the posterior beliefs about the relevance of different visual features and exemplars.

However, to specify the structure of the hierarchical generative model, a few additional assumptions are required. First, we can assume that the probability $p(H_t)$ of hypothesis H_t being correct is represented in a factorized from, that is, $p(H_t)$ equals to the product of the probability $p(F_t)$ that one of the visual features F_t is currently relevant and of the conditional probability $P(E_t|F_t)$ that one of the two exemplars E_t is currently relevant (given the fact that the corresponding visual feature F_t is relevant for the selection process). Alternatively, we can assume that only the probability $p(H_t)$ of hypothesis H_t being correct is explicitly represented and that the marginal probability $p(F_t)$ is computed only implicitly via the integration of corresponding beliefs.

Depending on the starting assumption one will end up with slightly different structure of the corresponding hierarchical generative model. Here we will describe in detail only the generative model based on the assumption that only the joint hypothesis probability $p(H_t)$ is explicitly represented and actively updated within the belief space. The reason for this is that model comparisons (see below) suggest that such representation better captures subject behavior. Nevertheless, the detailed derivation and the analysis of the behavioral data based on the alternative assumption, mentioned above, are provided in the supplementary material (S1 Text).

Here we will define the generative model as a three-level hierarchy (see Fig 4 for graphical representation): (i) the 1st level of the hierarchy encodes the hidden selection rule, that is, the currently correct hypothesis H_t (see Eq (5)); (ii) the 2nd level of the hierarchy encodes the probability, in the form of a state space vector $\vec{h}_t^{(e)}$, that each of the possible exemplar-feature pairs is currently relevant for the experimenter's choices (see Eq (6)), and (iii) the 3rd level of the hierarchy encodes the probability, in the form of the state space vector $\vec{h}_t^{(f)}$, that each visual feature is currently relevant for the experimenter's choices (see Eq (7)).

Assuming that the k th hypothesis is the correct one ($k \in \{1,\ldots,6\}$), the corresponding exemplar will be selected with probability $1-\varepsilon$, where ε denotes the error rate of experimenter's choices. We will encode the experimenter's choice with a binary vector $\vec{e}_t \in \{0,1\}^6$ whose

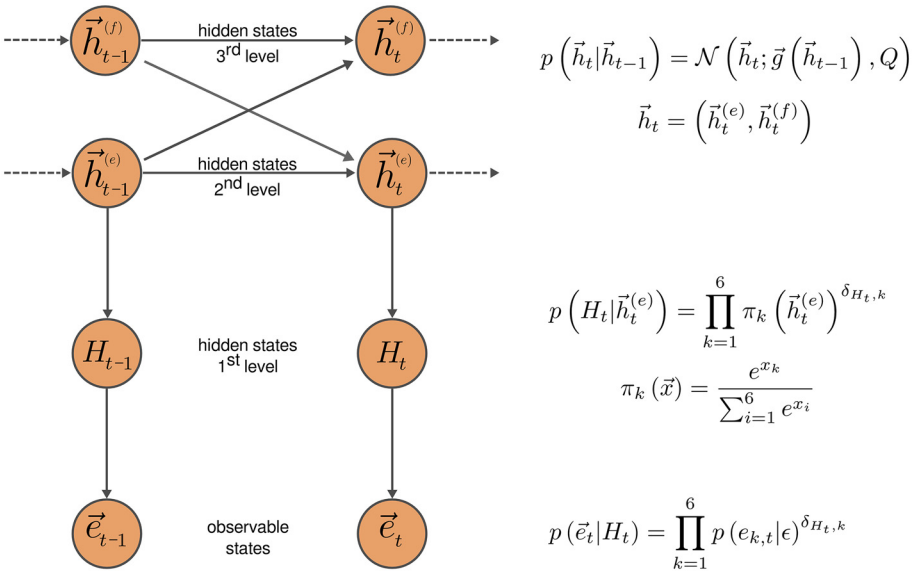

Fig 4. A graphical representation of the hierarchical generative model of percepts. The highest 3rd level hierarchy describes the dynamics of the three dimensional state space vector $\vec{h}_t^{(f)}$ that encodes the relevance of the three visual features. Similarly, the 2nd level of the hierarchy describes the dynamics of the six dimensional state space vector $\vec{h}_t^{(e)}$ that encodes the relevance for the selection process of the six exemplar-feature pairs. The functional form of the state transition probability $p(\vec{h}_t|\vec{h}_{t-1})$ is shown on the right hand side of the plot. The 1st level of the hierarchy encodes the currently active selection rule, that is, a currently correct hypothesis H_t, where the currently correct hypothesis is drawn from a conditional probability $p(H_t|\vec{h}_t^{(e)})$ shown on the right hand side of the plot. Finally, the observable states are denoted with the six dimensional vector $\vec{e}_t$, that encodes currently selected exemplars. On the right hand side of the plot we show the conditional probability $p(\vec{e}_t|H_t)$ of selecting the k th exemplar given the active selection rule H_t. For details, see the Eqs (4) to (7) and the accompanying text.

doi:10.1371/journal.pcbi.1004558.g004

elements are set to 1 or 0 depending on the presence or absence of the corresponding exemplar on the selected card. Thus, we can write the observation likelihood as

$$p(\vec{e}_t|H_t)=\prod_{k=1}^{6}p(e_{k,t}|\boldsymbol{\varepsilon})^{\delta_{H_t,k}};\quad p(e_{k,t}|\boldsymbol{\varepsilon})=\ (1-\boldsymbol{\varepsilon})^{e_{k,t}}\boldsymbol{\varepsilon}^{1-e_{k,t}},\tag{4}$$

where $\delta_{H_t,k}$ denotes Kronecker's delta and $e_{k,t}$ denotes the kth component of $\vec{e}_t$.

At the 1st (lowest) level of the hierarchy, we defined the probability that a hypothesis $H_t \in \{1,\ldots,6\}$ is the correct one as a categorical probability distribution

$$p(H_t|\vec{h}_t)=\prod_{k=1}^{6}\pi_k(\vec{h}_t^{(e)})^{\delta_{H_t,k}},\tag{5}$$

where the $\pi_k(\vec{h}_t^{(e)})$ denotes the probability of the k th hypothesis. These probabilities are encoded at the 2nd level of the hierarchy (see Fig 4) with the real valued vector $\vec{h}_t^{(e)} \in \mathbb{R}^6$, where

we defined the mapping to the space of categorical probabilities as the softmax transform

$$\pi_k(\vec{h}_t^{(e)}) = \frac{e^{h_{k,t}^{(e)}}}{\sum_{j=1}^{6} e^{h_{j,t}^{(e)}}}.$$

To incorporate an attention-like mechanism within the perceptual model, we make the state transition of $\vec{h}_t^{(e)}$ to follow a winner-take all (WTA) dynamics. We used this type of dynamics for three reasons:

1. The WTA dynamics is characterized by a set of stable fixed points that can be arranged in such a way that at each fixed point only one component of $\vec{h}_t^{(e)}$ is set to a high value (which encodes a high relevance of the corresponding exemplar), while all other components have low values. Such attractor state captures the structure of the WCST environment, in which at any moment in time only one exemplar-feature pair can be relevant.
2. Adding uncorrelated noise to the WTA dynamics mediates the switching between stable attractors. The larger the noise term the more probable is the transition between attractors. Thus, we can use a single parameter that defines the level of noise in the WTA dynamics to capture different experimental conditions.
3. WTA networks were successfully used before as a hierarchical neural model of higher cognitive functions [8,15,16,18,21,25,64,65], and as a model of attention spontaneously emerging from competitive neural dynamics [66].

Thus, assuming the WTA dynamics, the time evolution of $\vec{h}_t^{(e)}$ becomes

$$\vec{h}_{t+1}^{(e)} = \tau_e \vec{h}_t^{(e)} + \kappa_e + W_{lat}^{(e)} \varphi(\vec{h}_t^{(e)} - \kappa_e) + W_{dist}^{(f)} \varphi(\vec{h}_t^{(f)} - \kappa_f) + \vec{\omega}_t^{(e)}. \quad (6)$$

Here $\varphi(x) = \frac{1}{1+e^{-x}}$ and $\varphi(\vec{y}) = (\varphi(y_1), \cdots, \varphi(y_n))$; $\vec{\omega}_t^{(e)}$ denotes a vector of i.i.d. random variables drawn from normal distribution $\mathcal{N}(\vec{\omega}_t^{(e)}; 0,\ q_e I_6)$ with zero mean and variance q_e; τ_e denotes the time scale of the update equations, and κ_e an additive constant. Importantly, the dynamics of the $\vec{h}_t^{(e)}$ is influenced by the state vector $\vec{h}_t^{(f)} \in \mathbb{R}^3$ at the 3rd level of hierarchy, which encodes the relevance of the three visual features (see Fig 4). The time evolution of $\vec{h}_t^{(f)}$ is defined by an analogous set of equations

$$\vec{h}_{t+1}^{(f)} = \tau_f \vec{h}_t^{(f)} + \kappa_f + W_{lat}^{(f)} \varphi(\vec{h}_t^{(f)} - \kappa_f) + W_{dist}^{(e)} \varphi(\vec{h}_t^{(e)} - \kappa_e) + \vec{\omega}_t^{(f)}. \quad (7)$$

Importantly, the connectivity matrices $W_{lat}^{(e)}$, $W_{lat}^{(f)}$ denote the inhibitory interactions within levels, which are essential for the realization of attractor dynamics and $W_{dist}^{(e)}$, $W_{dist}^{(f)}$ denote the excitatory interactions between the levels of the hierarchy. This allows for integrating the beliefs about hypothesis relevance into beliefs about feature relevance.

In what follows, to simplify the notation, we will merge the state vectors $\vec{h}_t^{(e)}$ and $\vec{h}_t^{(f)}$ into a single state vector $\vec{h}_t = (\vec{h}_t^{(e)},\ \vec{h}_t^{(f)})$ whose update equation is denoted by $\vec{g}(\vec{x})$, that is,

$$\vec{h}_{t+1} = \vec{g}(\vec{h}_t).$$

Bayesian inference. Given the observation likelihood Eq (4), hypothesis probability Eq (5) and the transition probabilities Eqs (6) and (7) we write the full generative model as

$$p(\vec{e}_t, H_t, \vec{h}_t, \vec{h}_{t-1}|e_{1\ldots t-1}) = p(\vec{e}_t|H_t)p(H_t|\vec{h}_t)p(\vec{h}_t|\vec{h}_{t-1})p(\vec{h}_{t-1}|e_{1\ldots t-1}), \quad (8)$$

where $e_{1\ldots t-1} = (\vec{e}_1, \ldots, \vec{e}_{t-1})$ denotes all past observations. As we are interested in obtaining the posterior probability of the hidden states $p(H_t, \vec{h}_t|e_{1\ldots t})$, we require a compact form of the generative model

$$p(\vec{e}_t, H_t, \vec{h}_t|e_{1\ldots t-1}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\vec{e}_t, H_t, \vec{h}_t, \vec{h}_{t-1}|e_{1\ldots t-1}) d\vec{h}_{t-1}.$$

To obtain this compact form it is necessary to calculate the following integral

$$p(\vec{h}_t|e_{1\ldots t-1}), = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\vec{h}_t|\vec{h}_{t-1})p(\vec{h}_{t-1}|e_{1\ldots t-1}) d\vec{h}_t.$$

Assuming that $p(\vec{h}_{t-1}|e_{1\ldots t-1})$ is a normal distribution with mean $\vec{\mu}_{t-1}$ and covariance matrix Σ_{t-1}, we can approximate the integral on the right hand side as

$$p(\vec{h}_t|\vec{e}_{1\ldots t-1}) = \mathcal{N}(\vec{h}_t; \vec{g}(\vec{\mu}_{t-1}), \partial_{\vec{h}}\vec{g}\, \Sigma_{t-1} \partial_{\vec{h}}\vec{g}^T + Q), \quad Q = q_e I_6 \oplus q_f I_3$$

$$p(\vec{h}_0) = \mathcal{N}(\vec{h}_0; \vec{\mu}_0, \ \Sigma_0), \ \ \vec{\mu}_0 = (\vec{\mu}_e^0, \vec{\mu}_f^0) \text{ and } \Sigma_0 = \sigma_e^0 I_6 \oplus \sigma_f^0 I_3,$$

where the approximate predictive distribution $p(\vec{h}_{t-1}|e_{1\ldots t-1})$ is obtained by linearizing $\vec{g}(\vec{x})$ around the currently known mean $\vec{\mu}_{t-1}$, and where $\oplus$ denotes direct sum of matrices which constructs a block diagonal matrix from the elements of the sum.

To invert the generative model we apply the variational Bayesian method and the mean-field approximation in which the posterior distribution is approximated by a variational distribution. Thus, we write the posterior probability over the hidden states as a product of approximate posterior distributions, that is

$$p(H_t, \vec{h}_t|\vec{e}_{1\ldots t}) = q(H_t)q(\vec{h}_t),$$

where we chose the functional forms of the approximate posteriors as the distribution with maximum entropy given the specified mean and variance. This procedure allows for minimal assumptions about the form of the approximate posterior [46]. Hence, for the posterior probability over the discrete space of hypotheses we selected again a categorical probability

$$q(H_t) = \prod_{k=1}^{6} \rho_{t,k}{}^{\delta_{H_t,k}},$$

whereas for the posterior beliefs about the relevance of exemplars and visual features we selected a multivariate normal distribution

$$q(\vec{h}_t) = N(\vec{h}_t; \vec{\mu}_t, \Sigma_t).$$

Note that in this formulations the posterior belief is fully defined by the tuple of the posterior expectations and the posterior covariance, that is, posterior uncertainty; hence we will denote beliefs as a set $b_t = \{\vec{\mu}_t, \Sigma_t\}$.

Following variational calculus, the approximate posterior, given the mean-field approximation, is proportional to the exponential of the variational energy [67]. The variational energies for the given generative model and the above mentioned factorization of approximate posterior are defined as

$$I(H_t) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} q(\vec{h}_t) \ln p(\vec{e}_t,\ H_t, \vec{h}_t | \vec{e}_{1\dots t-1}) d\vec{h}_t,$$

$$I(\vec{h}_t) = \sum\nolimits_{H_t \in \{H_1, \dots,\ H_6\}} q(H_t) \ln p(\vec{e}_t, H_t, \vec{h}_t | \vec{e}_{1\dots t-1}).$$

To find the dependency of current beliefs b_t on prior beliefs b_{t-1} and current observation $\vec{e}_t$ we used a series of approximations previously described in [46], which we extended to the multidimensional case.

First, to compute $I(H_t)$, we need to know the beliefs b_t, whose computations require knowing $I(\vec{h}_t)$, which is a functional of $q(H_t)$, thus leading to a circular problem. We break the circularity by computing $I(H_t)$ with the expected beliefs $\hat{b}_t \in \{\vec{g}(\vec{\mu}_{t-1}),\ \partial_{\vec{h}} \vec{g}\, \Sigma_{t-1} \partial_{\vec{h}} \vec{g}^T + Q\}$; hence, we assume that the information about the observation $\vec{e}_t$ first changes the 1st level of the model's hierarchy and then propagates to the 2nd and 3rd level. As the exponential of the $I(H_t)$ has the form of a categorical distribution, one can show with simple algebraic manipulations that

$$\rho_{t,k} = \frac{p(e_{t,k}|\boldsymbol{\varepsilon}) e^{g_k(\vec{\mu}_{t-1})}}{\sum_{j=1}^{6} p(e_{t,j}|\boldsymbol{\varepsilon}) e^{g_j(\vec{\mu}_{t-1})}}. \tag{9}$$

With the known $\vec{\rho}_t$ one can compute the $I(\vec{h}_t)$, where the difficulty is that the variational energy does not have a quadratic form, that is, the exponential of $I(\vec{h}_t)$ is not a Gaussian distribution. Thus, to obtain a Gaussian form of the approximate posterior we need an additional quadratic approximation to the variational energy

$$\hat{I}(\vec{h}_t) = I(\vec{g}(\vec{\mu}_{t-1})) + \partial_{\vec{h}_t} I(\vec{g}(\vec{\mu}_{t-1}))(\vec{h}_t - \vec{g}(\vec{\mu}_{t-1}))$$
$$+ \frac{1}{2}(\vec{h}_t - \vec{g}(\vec{\mu}_{t-1}))^T \left[\partial^2_{\vec{h}_t} I(\vec{g}(\vec{\mu}_{t-1}))\right](\vec{h}_t - \vec{g}(\vec{\mu}_{t-1})),$$

where we made a second order Taylor expansion of $I(\vec{h}_t)$ around the predictive mean $\vec{g}(\vec{\mu}_{t-1})$, that is, the anticipated position of the posterior expectation. Finally, having the quadratic form we get the posterior mean $\vec{\mu}_t$ as the argument of the maximum of $\hat{I}(\vec{h}_t)$. The maximum is obtained with the Newton's method

$$\vec{\mu}_t = \operatorname{argmax} \hat{I}(\vec{h}_t) = \vec{h}_t - \left[\partial^2_{\vec{h}_t} \hat{I}(\vec{h}_t)\right]^{-1} \partial_{\vec{h}_t} \hat{I}(\vec{h}_t). \tag{10}$$

As Eq (10) is valid for any point $\vec{h}_t$ of the quadratic function $\hat{I}(\vec{h}_t)$, we can select again the expansion point $\vec{g}(\vec{\mu}_{t-1})$ as the starting value. In this way we obtain the following update

equations for the expected relevance of the hidden states

$$\begin{aligned}\vec{\mu}_t &= \vec{g}(\vec{\mu}_{t-1}) + \Sigma_t \vec{\delta}_t,\\ \vec{\delta}_t &= (\vec{\rho}_t - \vec{\pi}(\vec{g}^{(e)}(\vec{\mu}_{t-1})), \vec{0}_3),\end{aligned} \tag{11}$$

where $\vec{0}_3$ denotes the three-dimensional zero vector and where the posterior covariance Σ_t is given as the inverse of the negative Hessian at the expansion point $\vec{g}(\vec{\mu}_{t-1})$, that is,

$$\Sigma_t = -[\partial^2_{\vec{h}_t} I(\vec{g}(\vec{\mu}_{t-1}))]^{-1}.$$

The posterior covariance is updated as

$$\Sigma_t = \frac{\hat{\Sigma}_{t-1}}{I + \hat{\Sigma}_{t-1} Y_t};\ \hat{\Sigma}_{t-1} = \partial_{\vec{h}} \vec{g}(\vec{\mu}_{t-1})\, \Sigma_{t-1} \partial_{\vec{h}} \vec{g}^T(\vec{\mu}_{t-1}) + Q,$$

$$Y_t = \left[\oplus_{i=1}^{6} \pi_i(\vec{g}^{(e)}(\vec{\mu}_{t-1})) - \vec{\pi}(\vec{g}^{(e)}(\vec{\mu}_{t-1}))\vec{\pi}(\vec{g}^{(e)}(\vec{\mu}_{t-1}))^T\right] \oplus 0_{3,3}, \tag{12}$$

where $0_{3,3}$ denotes squared null matrix and $\partial_{\vec{h}} \vec{g}(\vec{\mu}_{t-1})$ denotes the Jacobian matrix of $\vec{g}(\vec{h})$ computed at prior expectations $\vec{\mu}_{t-1}$.

There are two interesting features of these update equations:

- The update equation for the posterior expectation Eq (11) have the form of a WTA neural network, with the key feature that the external input is proportional to the prediction error. This is similar to the hierarchical neuronal network models used in [15,25] to model behavioral planning in prefrontal cortex. The important difference is that in our model the update equations are derived from a probabilistic generative model (see Eq (8)), and therefore there is an adaptive influence of prediction errors on the internal dynamics of the WTA network; as expected from the Bayesian observer assumption.
- The hypothesis evidence $p(\vec{e}_t|H_t)$ is modulated by the predicted relevance of that hypothesis $\vec{g}^{(e)}(\vec{\mu}_{t-1})$ when the posterior hypothesis probability $\vec{\rho}_t$ is computed (see Eq (9)). Effectively, the evidence in favor of a hypothesis is neglected if the expectation about its relevance is low. This is similar to the effect that attention has on the processing of sensory information, as only the currently relevant features of the stimuli are being processed at any moment of time. Importantly, in the presence of competitive inhibitory dynamics the expectations of all but the most likely hypothesis will be suppressed. In other words, internal dynamics of beliefs leads to selection of prior expectation [34].

As the derivation of the perceptual model required multiple assumptions, which are not directly motivated by the behavioral data, it is important to test which of the assumption is actually essential for describing and predicting behavioral responses. Thus, in what follows we will describe several variants of the perceptual model that are obtained by relaxing some of the assumption made in the derivations presented above.

Structured models. To reduce the number of free parameters in the perceptual model described above we will assume that between the 2nd and the 3rd level there are only symmetric excitatory connections with equal values and that these connections exist only between components encoding the relevance of exemplars and corresponding visual features, thus

$$\left[W^{(f)}_{dist}\right]_{i,j} = \left[W^{(e)}_{dist}\right]_{j,i} = \begin{cases} w_{dist}, & \text{for } \nu(i) = j \\ 0, & \text{for } \nu(i) \neq j \end{cases},$$

where $\nu(i)$ maps the i th exemplar to the corresponding visual feature. Furthermore, we will assume that within the 2nd and the 3rd level there are only symmetric inhibitory connections with equal values, thus

$$\left[W_{lat}^{(e,f)}\right]_{i,j} = \begin{cases} -w_{lat}^{(e,f)}, & \text{for } i \neq j \\ 0, & \text{for } i = j \end{cases}.$$

Importantly, we will constrain the WTA dynamics to attractor states in which only single component of $\vec{h}_t^{(e)}$ and $\vec{h}_t^{(f)}$ have high values while all other components are set to zero or lower values. This is achieved by setting $w_{lat}^{(e,f)} = 2\kappa_{e,f}$, as suggested in [41].

However, removing the lateral inhibition form either the 2nd or the 3rd level would not disrupt completely the attractor dynamics as long as there are excitatory connections between levels. Thus, we will also consider two additional variants of the structured model in which we set either κ_e or κ_f to zero.

Therefore, the full set of parameters of the structured perceptual models is given by $\gamma = \{\varepsilon, \tau_{e,f}, \kappa_{e,f}, q_{e,f}, w^{dist}, \vec{\mu}_{e,f}^0, \sigma_{e,f}^0\}$, where in the first variant, denoted by w_1, we have that $\kappa_{e,f} \neq 0$, in the second variant, w_2, we set $\kappa_e = 0$, and in the third variant, w_3, we set $\kappa_f = 0$. The graphical representation of all structured model variants is shown in Fig 5A–5C.

Structure-free model. To explicitly test whether a complex attractor dynamics is necessary to describe subjects' behavior, that is, to test whether an attention-like mechanism modulates the update of beliefs, we require an alternative model without such an attentional focus mechanism. Hence, by setting both κ_e and κ_f to zero we obtain a structure-free model, denoted by d, in which the state transition of $\vec{h}_t$ is described with a diffusive dynamics (Fig 5D). The effect of removing the lateral inhibition is that a feature considered relevant will not inhibit other features, that is, there is no attentional focus effect. Note that setting $\kappa_{e,f} = 0$ also reduces the number of free parameters, thus the model complexity. Critically, by employing a model with lower complexity enables us to test whether the attentional focus model may be too complex for the behavioral data.

Note that both the structured and the structure-free models are able to capture the transient relevance of visual features. However, one expected difference is that the structured model, as it encodes a key constraint of the task environment, requires less evidence to form strong beliefs about relevance of visual features.

Reduced structured and structure-free models. To further simplify both structured and structure-free models note that the 3rd level of the hierarchy encodes the beliefs about the relevance of a visual feature. The importance of the 3rd level is to provide, as a dynamical implementation, the integration of the beliefs from the 2nd level of the hierarchy. The expectations at the 3rd level of the hierarchy are then used to generate responses, as described in the text below. In addition, one can also generate responses by using directly the expectations provided at the 2nd level of the hierarchy. In such a case the 3rd level of hierarchy is obsolete and can be removed.

In this way we obtain two reduced variants of the perceptual model defined by the following set of the free parameters $\{\varepsilon, \tau_e, \kappa_e, q_e, \vec{\mu}_e^0, \sigma_e^0\}$. For the reduced structured model, denoted by *rw*, κ_e is a free parameter (Fig 5E), while for the reduced structure-free model, denoted by *rd*, κ_e is fixed to zero (Fig 5F).

Non-Bayesian perceptual models. All the previous variants of the perceptual model were based on the same form of the update equations as provided in Eqs (11) and (12). The only difference so far between them is that certain parameters were removed, that is, fixed to zero. Importantly, these update equations are based on the assumption that subjects combine prior beliefs and sensory information in a Bayes optimal fashion. This requires the representation of

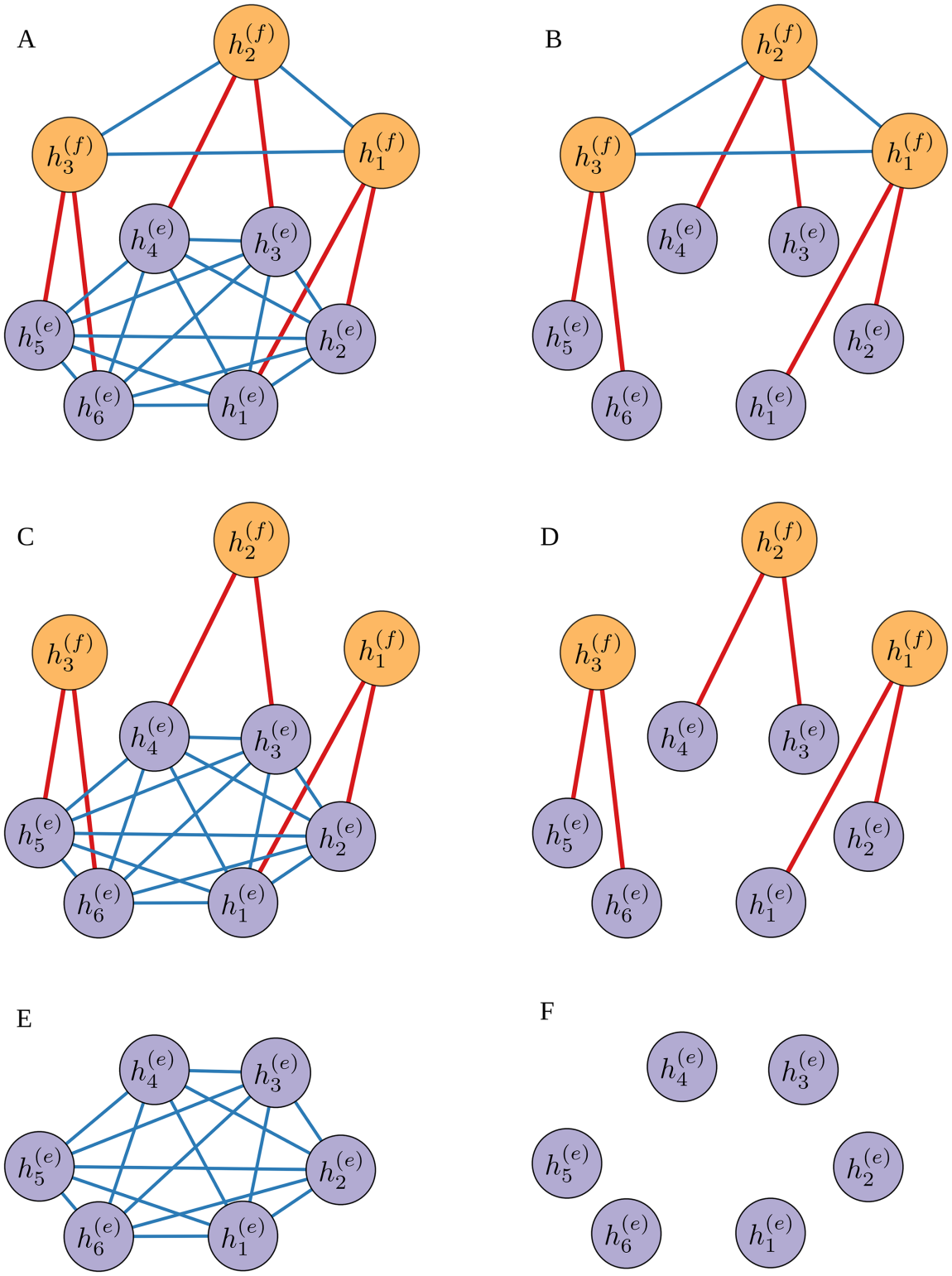

Fig 5. Visualization of six different model structures. Graphical representations of the connectivity matrix *W* of all variants of the perceptual model: the three variants of the structured model denoted with (A) w_1, (B) w_2, and (C) w_3; (D) the structure-free model variant denoted with *d*; and the two reduced variants of the perceptual model denoted with (E) *rw*, and (F) *rd* (for formal definition please see the accompanying text). The relevance of visual features and exemplars encoded by the vector $\vec{h}_t = (\vec{h}_t^{(e)}, \vec{h}_t^{(f)})$ (see Fig 4) corresponds to the activity levels at the nine nodes of the neural network. The orange nodes encode the relevance of three visual features $\vec{h}_t^{(f)}$ (color, motion, and shape). The purple nodes encode the relevance of six exemplar-feature pairs $\vec{h}_t^{(e)}$ (red-

color, green-color, leftward-motion, rightward-motion, circle-shape and square-shape). The structured models incorporate symmetrical lateral inhibition $w_{lat}^{(e,f)}$ (depicted with blue lines) that implements a winner-take-all dynamics (see Eq (6) and Eq (7)) and symmetrical excitation between levels w_{dist} (depicted with red lines), that implement integration of relevance between levels of hierarchy. Note that the structure-free model has only symmetrical excitation w_{dist} (red lines) from the level of exemplar-feature pairs to the level of visual features. In the case of the reduced perceptual models, the level of visual features is removed.

doi:10.1371/journal.pcbi.1004558.g005

both the expectations about the true state of the world and the uncertainties about these expectations. This assumption might not be correct in our case, and potentially the only relevant quantity, both for update of beliefs and for generating responses, might be the expectations about the relevance of exemplars and visual features. Thus, to test for this possibility we considered a non-Bayesian variant of the perceptual model described above, in which we fix the values of prior and posterior uncertainty on all levels of the hierarchy. This effectively makes the perceptual model non-Bayesian, as the sensory observations are not combined with the prior knowledge in a Bayes-optimal fashion. Thus, in the non-Bayesian variant of the perceptual model, we will set the posterior covariance matrix to a fixed value, $\Sigma_t = \alpha I_9$, which leads to the following update equations of expectations

$$\begin{aligned} \vec{\mu}_t^{(e)} &= \vec{g}^{(e)}(\vec{\mu}_{t-1}) + \alpha(\vec{\rho}_t - \vec{\pi}(\vec{g}^{(e)}(\vec{\mu}_{t-1}))), \\ \vec{\mu}_t^{(f)} &= \vec{g}^{(f)}(\vec{\mu}_{t-1}). \end{aligned} \tag{13}$$

Furthermore, in this formulation the evidence $\rho_{t,k} = 1 - \epsilon$ if the exemplar supporting kth hypothesis was selected and $\rho_{t,k} = \epsilon$ otherwise, where $\epsilon \in \left[0, \frac{1}{2}\right]$ denotes a free parameter which is not equivalent to the experimenters error rate ε, but only related to it. Note that the update equations shown in Eq (13) have a functional form similar to the Rescorla-Wagner model which is often used in reinforcement learning models [68,69].

Response model. Having obtained the update equation for the hidden belief states, the next step is to define an appropriate response model (see Fig 3). Thus, the question we will answer here is what would be an optimal response in an experimental trial t given the hidden beliefs b_t? Note first that the posterior probability that the ith visual feature is currently relevant is defined as

$$p_{t,i} = \frac{e^{\mu_{t,i}^{(f)}}}{\sum_{j=1}^{3} e^{\mu_{t,j}^{(f)}}},$$

in the case of the perceptual model variants with the 3$^{\text{rd}}$ level of hierarchy, and

$$p_{t,i} = \frac{e^{\mu_{t,i_1}^{(e)}} + e^{\mu_{t,i_2}^{(e)}}}{\sum_{j=1}^{6} e^{\mu_{t,j}^{(e)}}},$$

in the case of the reduced perceptual model variants without the 3$^{\text{rd}}$ level (where i_1 and i_2 denote the positions of the exemplars of the corresponding ith visual feature).

Importantly, as described above, we have instructed the subjects that at the end of the experiment one of the experimental trials will be randomly selected and the subject will receive as a reward the money that they have assigned to the truly relevant visual feature. Thus, we will assume that the subject's responses depend on the subject's risk attitude. As various studies have demonstrated that humans exhibit variable risk tendencies [70–73], we will parametrize the subject's individual levels of risk aversion with an inverse risk factor θ_1. Using the formalism of the Bayesian decision theory (BDT) and under the assumption that a subject's absolute

risk aversion is inversely related to the outcome of the bet, we have derived theoretical evidence that the optimal response (for more details see S2 Text) is defined as

$$\vec{r}_t = \frac{\vec{p}_t^{\,\theta_1}}{\sum_{j=1}^{3} p_{t,j}^{\,\theta_1}}, \tag{14}$$

where the elements of the response vector $\vec{r}_t$ denote the fraction of money assigned to the corresponding visual feature. Note that the higher the θ_1 is the more money will be assigned to the visual feature with highest posterior probability $p_{t,j}$, hence the higher the θ_1 the riskier is the subject's behavior. In the limit of $\theta_1 \to 0$, the responses become independent of the posterior beliefs and the same amount of money is always assigned to all visual features, thus reflecting infinite risk aversion.

However, using the optimal response function to model subjects' behavior may be too restrictive, as the behavioral responses might deviate from the optimal responses for at least two reasons: First, the perceptual models proposed might not fully capture the hidden perceptual processes of human subject, thus there might be an unknown influences on the decision process. Second, recent findings suggest that human brain maintains only stochastic representation of posterior beliefs [51]. In other words, an exact representation of posterior expectations is not internally available to the subject. Thus, under an assumption that the posterior expectations are sampled stochastically, one expects that the deviation of the response from the optimal one is proportional to the posterior uncertainty [51].

To account for potential deviation from optimal response we will define the behavioral responses as

$$\vec{r}_t = \frac{\vec{p}_t^{\,\theta_1} e^{\vec{\xi}_t}}{\sum_{j=1}^{3} p_{t,j}^{\,\theta_1} e^{\xi_{t,j}}}, \tag{15}$$

where $\vec{\xi}_t$ denotes a vector of i.i.d. random variables representing perturbations to the optimal response. We will assume here that the perturbation term $\vec{\xi}$ has two components expressed as separate components of the covariance matrix of a zero-mean Gaussian distribution:

$$\vec{\xi}_t \sim \mathcal{N}(0, P_t);\ P_t = \theta_2 I_3 + \theta_3 \Sigma_t^{(f)}. \tag{16}$$

The first noise source represents unknown influences on the decision process, which we assume to be i.i.d. The second noise source, which represents the above stochastic sampling assumption, is proportional to the uncertainty about the expected relevance of the visual features. Note that the second component is only relevant for the probabilistic variants of the perceptual model with full hierarchical representation, as only in those cases is the posterior uncertainty about the feature relevance a dynamic quantity. Consequently, the full set of the parameters for the response model $m^{(r)}$ becomes $\theta = \{\theta_1, \theta_2, \theta_3\}$.

Finally, for the above defined response model the response likelihood is defined as the multivariate logistic-normal distribution, that is,

$$p(\vec{r}_t | b_t(\gamma), \theta) = \frac{1}{3 r_{t,1} r_{t,2} r_{t,3} \cdot Z(\vec{m}_t,\ P_t)} \mathcal{N}(clr(\vec{r}_t); \vec{m}_t, P_t).$$

Here $clr(\vec{r}_t)$ denotes the centered log-ratio transform

$$clr(\vec{r}_t) = \ln\left(\frac{\vec{r}_t}{\sqrt[3]{\prod_{i=1}^{3} r_{t,i}}}\right),$$

$Z(\vec{m}_t, P_t)$ denotes a normalization constant, and $\vec{m}_t = \theta_1 \vec{\mu}_t^{(f)}$ in the case of the full perceptual model or $\vec{m}_t = \theta_1 clr(\vec{p}_t)$ in the case of the reduced perceptual model.

The normalization constant is computed as

$$Z(\vec{m}_t, P_t) = \iiint_{\mathbb{R}} \delta(\vec{a}^T \cdot \vec{x}_t)\mathcal{N}(\vec{x}_t; \vec{m}_t,\ P_t)d\vec{x}_t = \frac{1}{\sqrt{2\pi(\vec{a}^T P_t \vec{a})}} \exp\left(-\frac{(\vec{a}^T \vec{m}_t)^2}{2(\vec{a}^T P_t \vec{a})}\right),$$

where the projection vector $\vec{a} = (1, 1, 1)^T$. The normalization constant is required because of the mapping of the space of posterior expectations $\vec{\mu}_t^{(f)} \in \mathbb{R}^3$ onto a 2D simplex, which is the space of responses $\Delta^2 = \left\{\vec{r}_t \in \mathbb{R}^3 | \sum_{i=1}^{3} r_{t,i} = 1,\ r_{t,i} \geq 0 \text{ for } \forall\, i\right\}$.

For model comparisons, we will consider two response models. For both models, all the equations in this section apply, but the critical difference is that we only allow θ_3 as a free parameter in the so-called full response model, while in the reduced response model we fix θ_3 at 0. The effect of this difference is that the reduced model assumes a constant response variability of subjects, while the full response model allows for response variability to be dependent on the internal uncertainty about feature relevance. Note that having the inverse risk factor θ_1 as a free parameter in all variants of the response model is a result of a preliminary analysis (not presented here) which showed that response model variants with fixed risk factor have substantially lower model evidence compared to the considered variants of the response model.

List of models and model evidence computation

For the model comparison, we have paired all the full variants of the Bayesian perceptual models with the two variants of the response model; the reduced variants of the Bayesian models and all the variants of the non-Bayesian perceptual models were paired only with the reduced response model, as the posterior uncertainty about the visual features $\Sigma_t^{(f)}$ is set to constant values in this cases. In addition, we have defined a simple baseline model. Hence in total we consider 17 behavioral models denoted as:

1. *BM*—Baseline model in which the beliefs and the uncertainties about the beliefs are assumed to be constant over time. Thus, all the parameters of the perceptual model are set to zero, except $\vec{\mu}_f^0$. Similarly, we fixed $\theta_1 = \theta_2 = 1$ as they are redundant for this case and leave only θ_3 as the free parameters of the response model. The role of the baseline model here is to provide for a trivial explanation to the behavioral data: subjects generated random responses around a fixed mean independent from the sensory cues.

2. $B^{f,\,r}_{rw,rd,\,d,\,w_1,\,w_2,w_3}$—Twelve different Bayesian perceptual models, where the superscript denotes the variant of the response model ($f \rightarrow \theta_2 > 0$, $r \rightarrow \theta_2 = 0$), and the subscript denotes the variants of the perceptual model ($rw \rightarrow$ reduced perceptual model with lateral inhibition, $rd \rightarrow$ reduced perceptual model without lateral inhibition, $d \rightarrow$ full perceptual model without inhibition at all levels, $w_1 \rightarrow$ full model with lateral inhibition on all levels, $w_2 \rightarrow$ full model

with lateral inhibition only at the 2nd level, $w_3 \rightarrow$ full model with lateral inhibition only at the 3rd level), see Fig 5.

3. $NB^{r}_{rw,rd,\ d,\ w_1,\ w_2,w_3}$—Six different non-Bayesian perceptual models, where the superscript denotes the only possible variant of the response model, the reduced response model, and the subscripts denote the variants of the perceptual model, with the same notation as above.

To summarize the motivation for these different variants of the perceptual model (see Methods above for details): the structure-free model variants test for the possibility that the structured representation is not required for describing the behavioral data; the model variants without the final level of the hierarchy (*rw*,*rd*) test for the possibility that the final level of hierarchy is redundant for describing the behavior; the non-Bayesian variants of the perceptual test for the possibility that the Bayesian observer assumption is not required for describing the behavior.

Each model variant is defined using a set of free parameters $\{\gamma,\theta\}$ for the perceptual and response models. To be able to define prior and posterior distributions in the same functional form of multivariate normal distributions, we transform all parameters so that they have the same domain of real numbers. Note that such a transformation does not change the value of model evidences, as to compute the model evidence one integrates over all the free parameters of a generative model. Let us denote by $\vec{\chi}$ the vector of perceptual and response parameters transformed to real space, then $\vec{\chi} = (\vartheta(\gamma), \vartheta(\theta))$, where

$$\vartheta(z) = \begin{cases} \ln(z), & if\ z \in \{\alpha, \kappa_{e,f}, q_{e,f}, w^{dist}, \sigma^{0}_{e,f}, \theta_1, \theta_2, \theta_3\} \\ \ln\left(\frac{2z}{1-2z}\right), & if\ z = \boldsymbol{\varepsilon} \\ \ln\left(\frac{2z-1}{2(z-1)}\right), & if\ z = \tau_{e,f} \\ z, & if\ z \in \{\vec{\mu}^{0}_{e}, \vec{\mu}^{0}_{f}\} \end{cases}.$$

Thus, we can define the prior distribution over model parameters as a multivariate normal distribution $\mathcal{N}(\vec{\chi}; \vec{\eta}_0, s_o I)$.

The log-joint probability distribution can then be written as

$$l(\vec{\chi}) = \sum_{k=1}^{T} \ln p\left(\vec{r}_k | b_k\left(\vec{e}_k, \vartheta^{-1}(\vec{\chi}_\gamma)\right), \vartheta^{-1}(\vec{\chi}_\theta)\right) + \ln\mathcal{N}(\vec{\chi}; \vec{\eta}_0, s_o I), \quad (17)$$

where T denotes the number of trials within a single experimental block. The Laplace approximation to the log-evidence is obtained as

$$\ln p(r_{1\ldots t} | \vec{e}_{1\ldots t}) = l(\vec{\beta}) + \frac{1}{2}\ln|2\pi S|, \quad (18)$$

where $\vec{\beta}$ denotes the mode of $l(\vec{\chi})$ and $S = -\partial_{\vec{\chi},\vec{\chi}} l(\vec{\chi})^{-1}|_{\vec{\chi}=\vec{\beta}}$, *i.e.* S is the negative inverse of the Hessian matrix at the mode $\vec{\beta}$.

To find the mode of $l(\vec{\chi})$ we applied the so-called Covariance Matrix Adaptation Evolution Strategy (CMA-ES). CMA-ES is a numerical optimization method, which has been applied successfully in various research areas [74–77] and is particularly useful for ill-conditioned and multimodal objective functions. In short, CMA-ES is a stochastic derivative-free method for numerical optimization of non-linear optimization problems [56,57]. We used a freely

available Matlab toolbox that implements the algorithm [Hansen, Nikolaus (2004). (https://www.lri.fr/~hansen/cmaes_inmatlab.html#matlab), Version 3.61].

Once the mode of the log-joint probability distribution (Eq (17)) is found, we have to estimate the curvature at the mode, that is, the Hessian matrix. We estimated the Hessian matrix by numerical differentiation [58], where we used the following toolbox [D'Errico, John (2006). (http://www.mathworks.de/matlabcentral/fileexchange/13490), MATLAB Central File Exchange. Retrieved 10. November 2013].

Because of the stochastic nature of the CMA-ES algorithm we repeated the stochastic search $N = 50$ times per experimental block for each model. For each of the N solutions we estimated the Hessian matrix and computed the Laplace approximation to the log-evidence. Finally, we kept the solution with the largest log-evidence, therefore increasing the probability of finding the maximal lower bound to the log-evidence and thus the most likely model of a subject's behavior. The numerically obtained $\vec{\beta}$ and S are used as the mean and the covariance matrix of the approximate posterior distribution $\mathcal{N}(\vec{\chi}; \vec{\beta}, S)$. Note that in this way we obtain the full covariance matrix without the need for a mean field approximation, which would neglect any existing correlations between parameters. All data processing was performed using MATLAB [version 8.1, The MathWorks Inc., Natick, Massachusetts].

Bayesian model selection

We first estimated the log model evidence of the 17 generative models described above for each experimental block. To obtain a total per-subject log-evidence for each experimental condition, we summed the estimated log-evidences over experimental blocks of a single experimental condition. This gives us the log model evidence of each generative model for each subject per experimental condition. We used the obtained log-evidences to apply the hierarchical Bayesian model selection approach described in [54,55]. By using hierarchical Bayesian model selection we assumed that the identity of the best-fitting model may vary across subjects. This requires treating the posterior model probability (the posterior belief that a given model has generated the data) as a random variable.

Thus, the two computed quantities of interest are the expected probability (EP) and the exceedance probability (XP) of each model: The EP is defined as the probability that a given model generated the behavioral data of a randomly selected subject (see [55] for a detailed mathematical description); The exceedance probability XP tells how likely it is that a given model will have the largest probability in a random sample from the posterior distribution. Importantly, the XP can be seen as a degree of confidence in the difference between posterior model probabilities [55]. Thus, when presenting the results of a model comparison we will only report the XP of the corresponding model or model family, as large XP at the same time implies significantly larger EP. Importantly, we will only consider recently proposed "protected" exceedance probability, which takes into account the null hypothesis that assumes that all the models are equally likely (see [55] for details). We will consider that the EP of a single generative model is significantly larger than the EP of other generative models, if the model's XP is above threshold value set at 0.95. Although, this threshold value was selected in the analogy to classical statistical tests that rely on p-values, its relation to the statistical power is not equivalent (see [55]).

We used the MATLAB implementation of the random-effect Bayesian model selection [(https://sites.google.com/site/jeandaunizeauswebsite/code/rfx-bms), retrieved January 2014]. In what follows we will describe the results obtained by applying the Bayesian model selection to the set of behavioral models that we used to approximate subjects' behavior in the probabilistic WCST.

Results

In Figs 6 and 7 we present the results of the random-effects Bayesian model comparison at the group-level. We have separated the model comparison between the two experimental conditions, switch and no-switch. We estimated the per-subject log-evidence for each experimental condition as the sum of log-evidences across the three corresponding experimental blocks. The top graph in both Figs 6 and 7 depicts the model attributions to the behavioral responses of each subject, that is, the posterior probability that a given model has generated the behavioral responses of each subject, for each condition separately. The bottom graphs show the corresponding XP for each of the 17 models. The direct comparison of behavioral models is

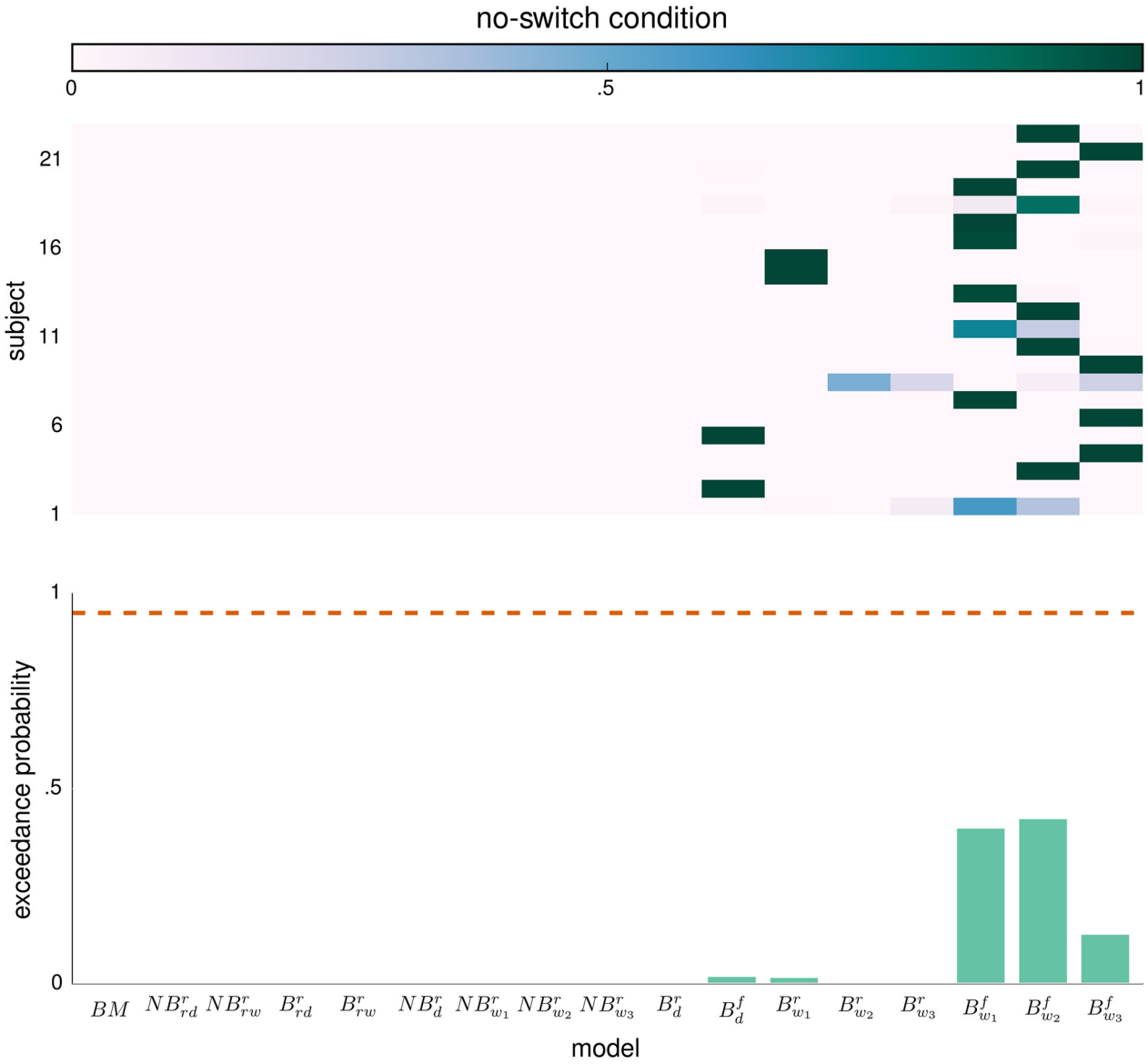

Fig 6. Random-effects model comparison for the no-switch condition. (top) Posterior model probability (see color bar) for each subject. For an exact description of each of the 17 models see main text. (bottom) Exceedance probability (XP) that a given model is more likely to generate the data than any other model. The dashed orange line denotes the confidence threshold level set at 0.95.

doi:10.1371/journal.pcbi.1004558.g006

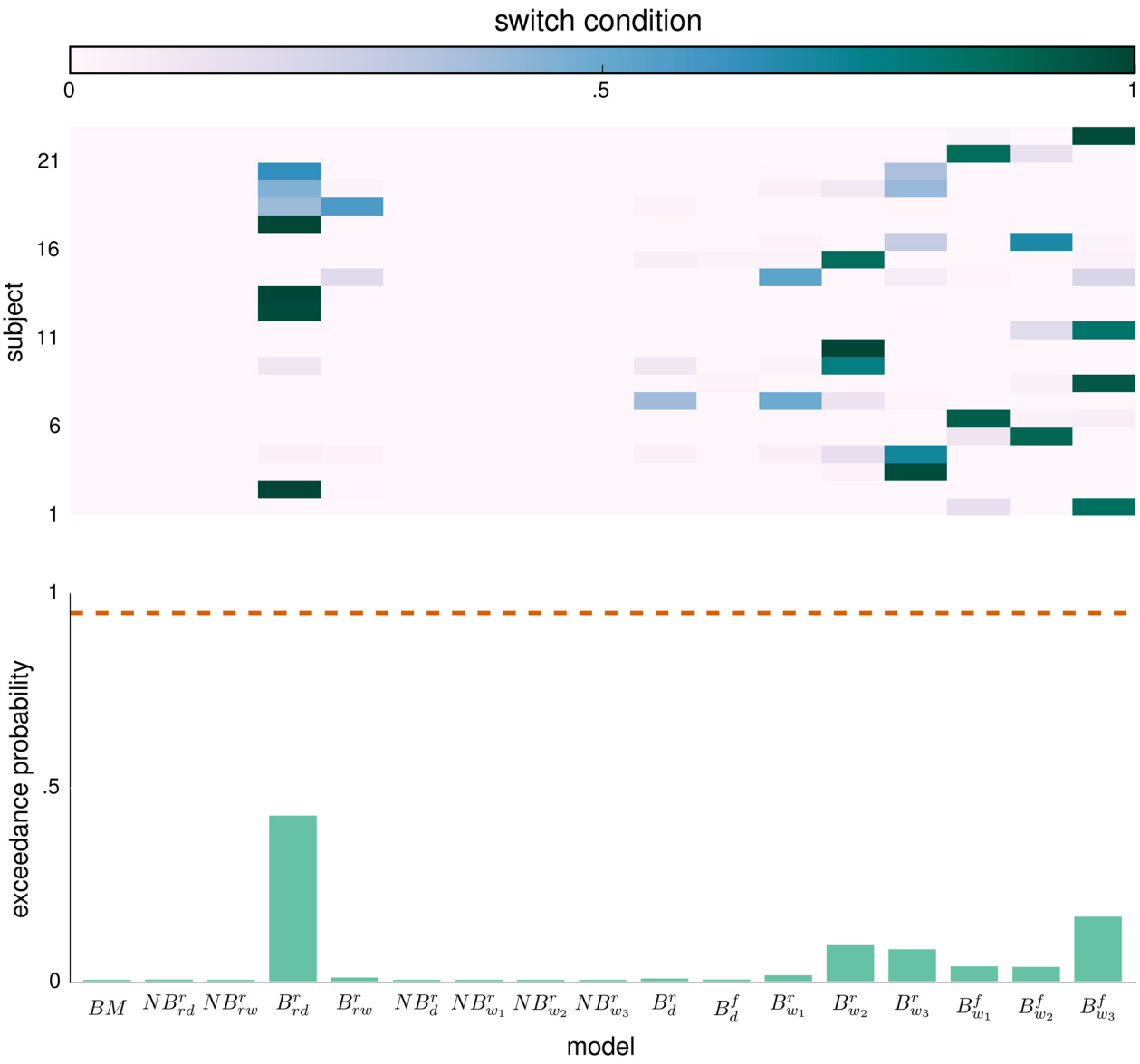

Fig 7. Random-effects model comparison for the switch condition. (top) Posterior model probability (see color bar) for each subject. For the exact description of each of the 17 models see main text. (bottom) Exceedance probability (XP) that a given model is more likely to generate the data than any other model. The dashed orange line denotes the confidence threshold level set at 0.95.

doi:10.1371/journal.pcbi.1004558.g007

inconclusive, as the highest XP is in both cases below the threshold value. Note that this is a typical issue when the model comparison set contains groups of closely related models [78].

The solution here is that instead of trying to answer which of the models provides the best description of behavioral data, we should ask which of the features of the perceptual and the response model are the most relevant for generating the data [78]. Note that in both figures we observe clustering of high model probabilities (top graphs) within closely related perceptual models (e.g. $B^f_{w_1,w_2,w_3}$) which only differ in the type of the connectivity matrix (see subsection Structured models in Methods). Thus, to determine which of the features of the perceptual and the response model are the most relevant for generating the behavioral data, we have performed four so-called family-wise model comparisons [78]. To test whether non-Bayesian or

Bayesian model variants better describe the behavioral data, we grouped all models into baseline (BM = $\{BM\}$), non-Bayesian $\left(\text{NB} = \left\{NB^{r}_{rw,rd,d,w_1,w_2,w_3}\right\}\right)$ and Bayesian $\left(\text{B} = \left\{B^{f,r}_{rw,rd,d,w_1,w_2,w_3}\right\}\right)$ model families. Similarly, to test whether a hierarchical representation of feature relevance is truly necessary we have grouped the models into BM, reduced perceptual $\left(\text{RP} = \left\{NB^{r}_{rw,rd}, B^{r}_{rw,rd}\right\}\right)$, and full perceptual $\left(\text{FP} = \left\{NB^{r}_{d,w_1,w_2,w_3}, B^{f,r}_{d,w_1,w_2,w_3}\right\}\right)$ model families. Finally, to test whether the attractor dynamics contributes to an explanation of the behavioral data, we have grouped models into the BM, structure-free $\left(\text{SFM} = \left\{NB^{r}_{rd,d}, B^{r}_{rd}, B^{f,r}_{d}\right\}\right)$, and structured $(\text{SM} = \{NB^{r}_{rw,w_1,w_2,w_3}, B^{r}_{rw}, B^{f,r}_{w_1,w_2,w_3}\})$ model families. In addition to separating behavioral models based on the features of perceptual model, we have grouped them based on the features of the response model, for which we considered only two model families, a model family with the reduced response model $\left(\text{RR} = \left\{BM, NB^{r}_{rw,rd,d,w_1,w_2,w_3}, B^{r}_{rw,rd,d,w_1,w_2,w_3}\right\}\right)$ and a family with the full response model $\left(\text{FR} = \left\{B^{f}_{rw,rd,d,w_1,w_2,w_3}\right\}\right)$.

From the results of the four family-wise model comparisons, shown in Fig 8, we can conclude with high confidence (XP above the threshold level of 0.95) that the Bayesian formulation of the perceptual model is essential for generating behavioral data in both experimental conditions (see Fig 8A and 8B). To understand the difference between NB and B model families in their ability to predict subjects' behavior we tested how well the behavioral models within each of these families predict subjects' performance. We computed the mean model performance by

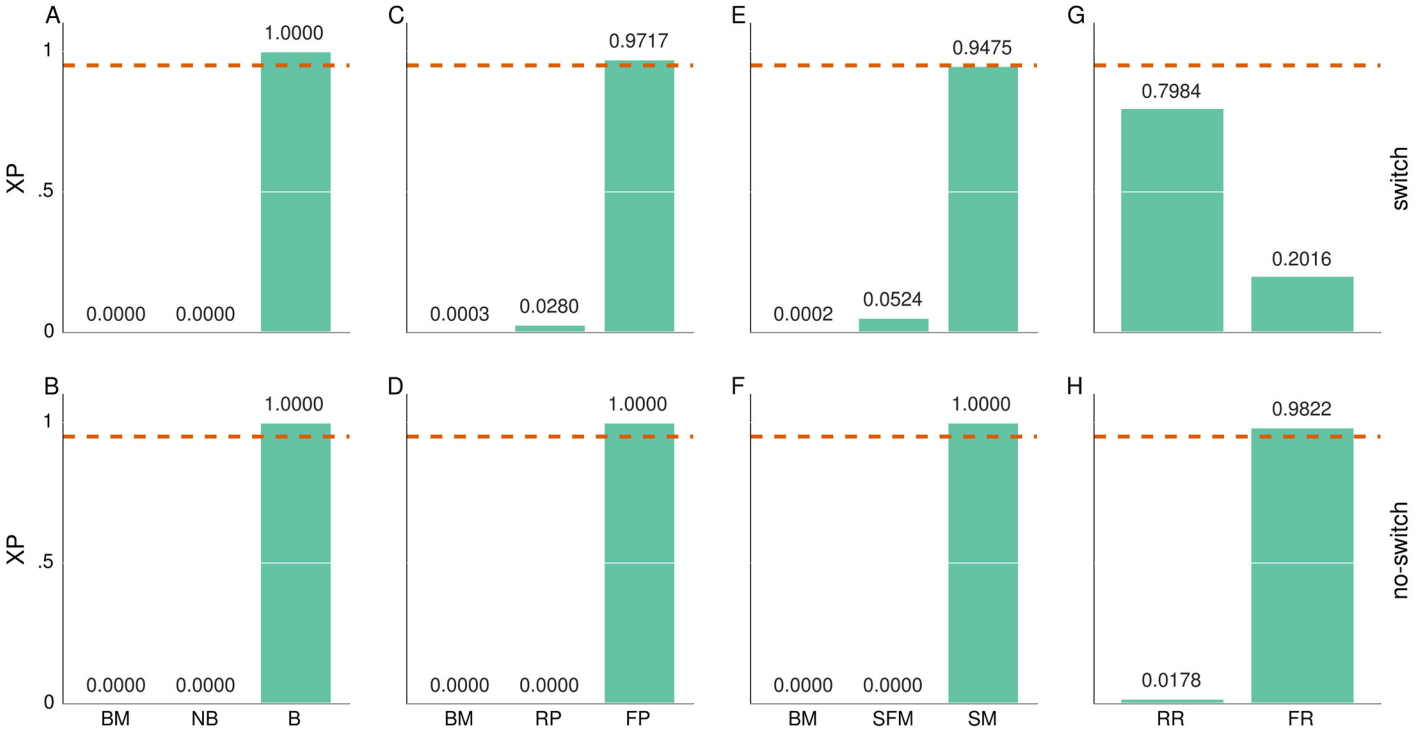

Fig 8. Family-wise model comparisons. (A-B) Exceedance probability (XP) of the baseline model (BM) non-Bayesian (NB) and Bayesian (B) model families. (C-D) XP of the BM, reduced perceptual (RP) and full perceptual (FP) model families. (E-F) XP of the BM, structure-free (SFM) and structured (SM) model families. (G-H) XP of the reduced response (RR) and full response (RR) model families. The top graphs (A,C,E,G) show the exceedance probability of model families for the switch condition, whereas the bottom graphs (B, D, F, H) show the exceedance probability of the model families for the no-switch condition. The dashed orange lines denote the confidence threshold level set at 0.95.

doi:10.1371/journal.pcbi.1004558.g008

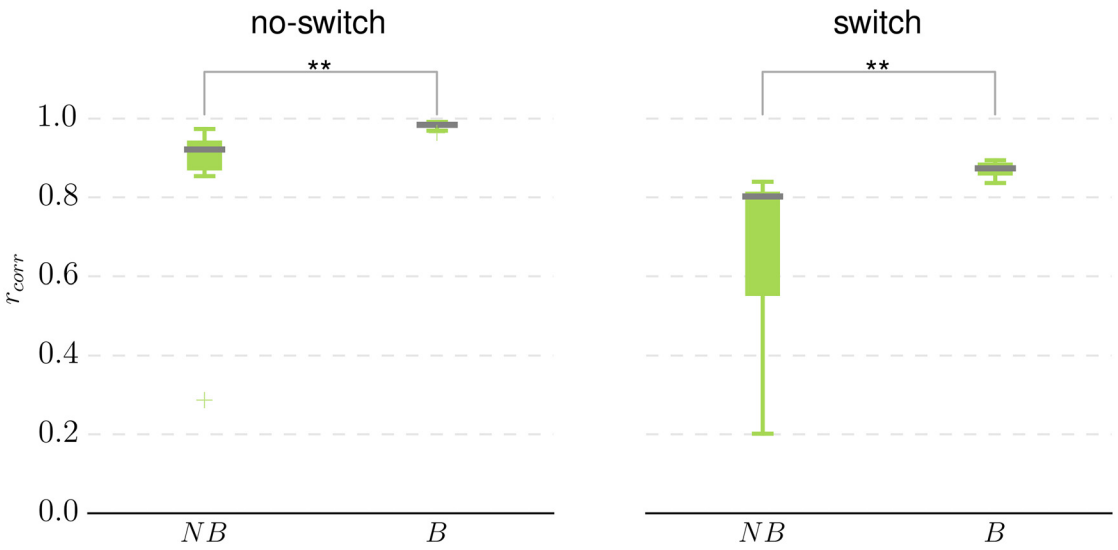

Fig 9. Distribution of the correlations between the mean model performance and the mean subjects' performance across two model families. Boxplot of the Pearson correlation coefficient r_{corr} estimated for each model within the non-Bayesian (NB) and the Bayesian (B) model families in the no-switch and switch condition. For each model within each family we have computed the Pearson correlation coefficient between the mean model performance and mean subjects' performance. In both conditions the non-Bayesian model family has a significantly lower median correlation (denoted by a dark horizontal line within the boxes) with $p<0.005$ (Kruskal-Wallis test).

doi:10.1371/journal.pcbi.1004558.g009

first estimating the expected performance per trial. To do this, we fixed model parameters to the mode $\vec{\beta}$ of the posterior parameter distribution and computed the expected model response; hence the expected performance per trial corresponds to the mean fraction of money assigned to the truly relevant visual feature at that trial. We averaged the per-trial expected model performance over a whole experimental block to obtain the mean model performance per experimental block. We then estimated the Pearson correlation coefficient between the mean model performance and mean subjects' performance across blocks and both experimental conditions. In Fig 9 we illustrate, with a box plot, the distribution of the estimated correlation within NB and B model families. The correlation coefficient shows that, on average, the NB model family has significantly lower correlation with subjects' performance, or in other words, the NB model family provides a worse fit to subjects' behavior compared to the Bayesian model family. Interestingly, within the NB family the models with consistently low correlation, in both conditions, are the structure-free model variants NB^r_d and NB^r_{rd} (see S1 Fig), whose update equation correspond to what is typically used in classical reinforcement learning models. On the other hand, the non-Bayesian model variants with attractor dynamics, namely NB^r_{rw,w_1,w_2,w_3}, show consistently high correlation with subjects' performance in both conditions (with one exception being model $NB^r_{w_2}$). This indicates that even only within the NB model family the attentional focus mechanism plays a critical role in replicating subjects' behavior.

Importantly, from the results of the family-wise model comparison we can also conclude with high confidence that the full variant of the perceptual model (including both the 2nd and 3rd level of the hierarchy, see Reduced structured and structure-free models for details) is an essential feature in both experimental conditions (see Fig 8C and 8D). The structured family of the perceptual model shows an XP above the threshold level only in the no-switch condition (Fig 8F), whereas in the switch condition the XP is slightly below the confidence threshold level (Fig 8E), but still high enough to be considered a trend. One possible explanation for the slightly reduced confidence in the structured model family (Fig 8E) is that in the switch

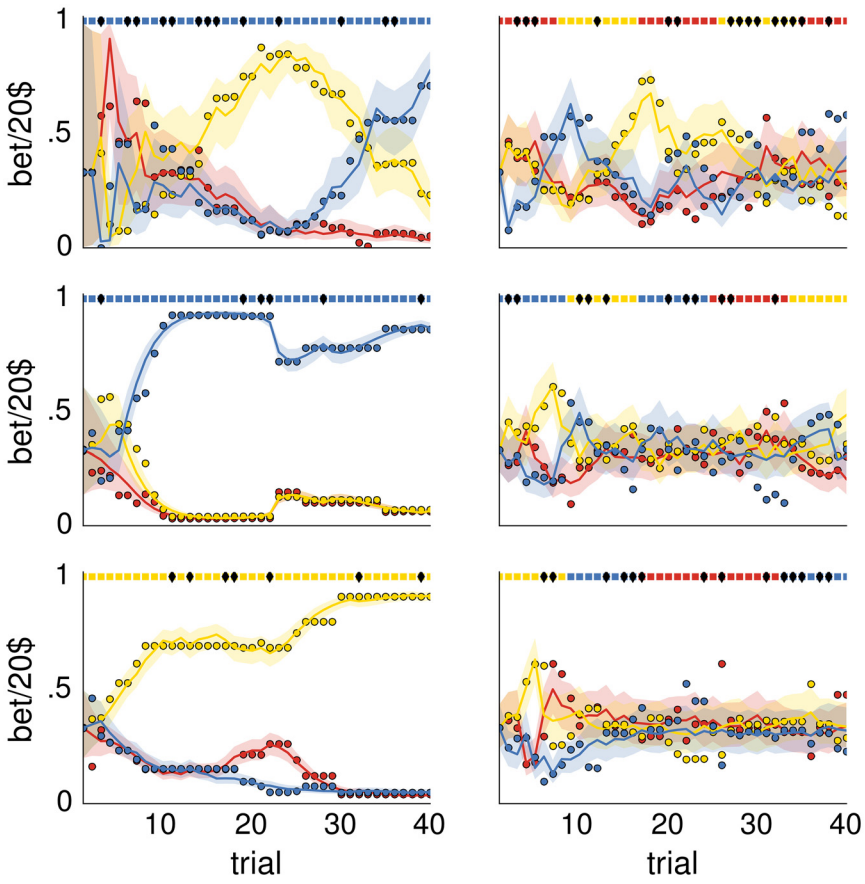

Fig 10. Behavioral responses and modeled responses for a representative single subject. (left) The three no-switch blocks, and (right) the three switch blocks. Colored circles denote the behavioral responses of subject #9 obtained as the fraction of money assigned to each of the three visual features on single trials. Solid lines denote the expected model response computed at the mode $\vec{\beta}$ of the posterior distribution over model parameters and averaged over posterior model probabilities for subject #9 (see Figs 6 and 7). The shaded area corresponds to the 95% probability interval. Each color corresponds to one of the three visual features (red—color, yellow—motion, blue—shape). The dotted colored line at the top of each plot denotes the relevant visual feature during each experimental trial, where the black diamond marks on the dotted line denote trials in which the experimenter selected the wrong card.

doi:10.1371/journal.pcbi.1004558.g010

condition one expects high levels of posterior uncertainty about the relevance of visual features. This is due to an increased difficulty in assigning contradicting evidence either to an experimenter's error or a change in the selection rule. Thus, in such an environment one does not expect that a subject can form strong beliefs about the relevance of each visual feature. Hence the attractor dynamics would not show strong advantages in generating the data, when compared to the structure-free model family.

Finally, when comparing model families with the full against the reduced variant of the response model we get mixed results across conditions. The full response model seems to be relevant for generating behavioral data only in the no-switch condition (Fig 8H), whereas in the switch condition the evidence is inconclusive (Fig 8G). This discrepancy between the confidence levels in the two experimental conditions may be caused by the increased difficulty of the switch task, which effectively introduced a higher variability in subjects' responses. Most of this variability may be explained simply by a high but constant level of response noise as formulated in the reduced response model.

To illustrate the dynamics encountered under the most likely types of behavioral model ($B^f_{w_1,w_2,w_3}$ in the no-switch condition and $B^r_{w_1,w_2,w_3}$ in the switch condition) we have plotted the measured and modeled responses of a representative subject (#9), see Fig 10. The modeled response was averaged over posterior model probability (see top graphs of Figs 6 and 7). Note that for the selected subject only the $B^f_{w_3}$ (in the no-switch condition) and $B^r_{w_2}$ (in the switch condition) have posterior model probabilities close to one and therefore contributed to the shown modeled responses. Importantly, one can see that the expected model responses appropriately track the subject's responses in all six experimental blocks, and that the deviations of the subject's responses from the expected response are mostly explained by the response variability, as indicated by the shaded area.

Discussion

We have used a probabilistic variant of the Wisconsin card sorting task (WCST) with belief solicitation to show that, in a rather complex environment, update of beliefs is modulated by an attentional focus mechanism. We analyzed behavioral data of 22 subjects using a meta-Bayesian framework [49,50]. This framework allowed us to compare multiple behavioral models, each implementing different assumptions about the underlying mechanisms that govern update of beliefs. We found evidence that incorporating an attentional focus mechanism within the behavioral model is the essential feature for modeling behavior. Specifically, we demonstrated that the attentional focus mechanism modulates subjects' expectations about the relevance of each visual feature and consequently influences the update of beliefs when new visual evidence is provided. In addition, we found that introducing a deviation from optimal responses (as predicted by Bayesian decision theory), during belief solicitation, further increased model evidence in one experimental condition.

WCST and belief solicitation

The variant of the WCST used here can be seen as a simple but representative task to which humans are often exposed, namely making decisions in situations where the relevant features of the environment are not obvious but need to be inferred first. What makes the WCST simpler when compared to natural environment is the reduced number of possible pre-learned hypotheses. However, the dynamic complexity is comparable to real world situations: (i) the rules of the environment can change, and (ii) in the specific WCST used here the experimenter occasionally 'makes a mistake' just as in the natural environment one often cannot know something with certainty. For the WCST task, these two naturally occurring sources of uncertainties make the necessary inference sufficiently complex to compute the subject's uncertainty about the relevance of visual features. To better infer the hidden internal beliefs and uncertainties of subjects, we used belief solicitation in a form of a betting assignment, which reflect a subject's hidden beliefs over the space of possible hypotheses. To our knowledge, such belief solicitation was not previously used in a WCST task, although similar experimental designs were used for simpler tasks [11,79].

Modeling effects of attention on evolution of beliefs

To incorporate attentional-focus within the perceptual part of the behavioral model we modeled the dynamics of the hidden states of a probabilistic generative model with a winner-take-all (WTA) dynamics. This is a well-known type of dynamics applied to artificial neural networks [37–40,80–82] and used as a part of connectionist models of decision making and

planning [19,25]. In addition, WTA network dynamics have been reported to capture a wide range of experimental findings [48,83–86].

For our purposes, the WTA neuronal network implemented a dynamic and self-regulated attention formation at the top level of a hierarchical representation of environmental features.

In comparison to the classical connectionist approach, e.g. [25], the main advantage of using the WTA dynamics within a Bayesian framework is that the adaptive coupling between the intrinsic network dynamics and external input (see Eqs (11) and (12)) is derived automatically as part of the update equations. These update equations provide Bayes-optimal behavior of the model by setting the connection weights to their optimal value. Although the optimization technique used by the brain may be different, such weight optimization may be assumed as a guiding computational principle of information processing in the brain.

Our finding—that competitive inhibitory WTA dynamics as a model of attentional focus is required for describing the hidden update process of subjects' beliefs—is in agreement with previous findings of Wilson and Niv [1]. This suggests that in a WCST task humans actively track only the evidence corresponding to features they pay attention to, that is, the ones they found potentially relevant for the current task. Importantly, as a safe-guard against over-fitting the data with a complex WTA dynamics, we employed simpler (with a reduced number of free parameters) variants of the perceptual model. The fact that the less complex behavioral models have lower model evidence suggests that the WTA dynamics has indeed adequate complexity to describe the behavioral data.

Predicting effects on behavior

The WTA dynamics introduces the following features in the evolution of beliefs: (i) faster convergence of beliefs to the working hypothesis; (ii) the beliefs are more inert to frequent changes in the environment, that is, to switch between the hypotheses sufficient amount of contradicting evidence has to accumulate. (iii) The beliefs change faster if the changes in the environment are rare, as after the fixed point is reached beliefs do not evolve further. In contrast, the diffusive dynamics of the SFM variants of the perceptual model is not bounded within finite volume of the belief space. Hence, as the posterior beliefs about a hypothesis' relevance can be strongly separated if the environment is stable for a long period of time and, once the switch occurs it would take a very long time to adjust the beliefs as nothing constrains the separation of the posterior expectations.

Consequently, as our results suggest, the proposed attractor dynamics modulate expectations. This would predict the following effects on behavior: (i) Even small amount of evidence can have a big impact on beliefs, (ii) if changes in the environment are too frequent they will have smaller impact on beliefs than expected from the diffusive dynamics, and (iii) if changes in the environment are rare it will take less contradicting evidence to change the working hypothesis than predicted by the diffusive and unconstrained dynamics.

Sub-optimality in human behavior

Although various studies have demonstrated that human behavior can approximate a Bayesian observer [26–28,60–62,87], human subjects can also behave sub-optimally when exposed to sufficiently complex tasks [28].

In recent work Acerbi et al. [51] have demonstrated that the response variability (deviation from expected response) is proportional to posterior uncertainty. Such a deviation from optimal responses can be explained if one assumes a stochastic representation of the posterior beliefs by the human brain [52,53].

Thus, to account for potential dependence of response variability on posterior uncertainty we considered two variants of the response model. In the first variant we assume that the response variability is constant over an experimental block. In the second variant we additionally allow for the variability of the modeled responses proportional to the posterior uncertainty (see Eqs (15) and (16)), which accounts for the potential stochastic representation of posterior beliefs.

Depending on the experimental condition both variants of the response model provide good accounts for the deviation of subjects' responses from the optimal response. In the no-switch condition (the relevance of visual feature is unchanged during the block, see Fig 8H) we found that the response variability is indeed proportional to the posterior uncertainty; in the switch condition (Fig 8G) the evidence is inconclusive although in favor of the assumption that the response variability is fixed and independent of the posterior uncertainty. A reason for this inconclusive result may be the increased difficulty of the experimental task in the switch condition. An increased difficulty makes the behavioral responses noisier (responses deviate more from the optimal response compared to the no-switch condition, see Fig 10). As the average response variability increases, there is less information about the dependency of response variability on experimental trials. Hence, most of this additional variability may be explained simply by a rather high but constant level of response noise as formulated in the reduced response model.

Related work on the computational role of attentional processes

Earlier work on the computational role of attention in the processing of sensory information suggested that attention can be understood as prior expectations about the sensory stimuli [88,89]. This rather simple view of attention as a prior has recently been extended to account for both selective and integrative attentional phenomena [34–36]. This extended view suggests that due to the computational complexity of the exact probabilistic inference and the limited amount of available cognitive resources, the human brain has to rely on approximations to efficiently solve perceptual tasks. In other words, the role of attention is to assign limited cognitive resources to the relevant part of the sensory stimuli, which provides local refinement of the internal representation of the hidden states of the environment.

However, this view on attention as an approximation to the exact Bayesian inference has been recently challenged. Under the free-energy principle [90]—which suggests that perception, attention, and action are all aimed toward suppressing the perceptual surprise about future sensory stimuli—attention is viewed as a sampling of only those parts of sensory stimuli that have high-precision in relation to the predictions of the internal model of the world [33]. Importantly, if the model of the world also predicts the precision of different parts of sensory stimuli, then that prediction is what Friston and colleagues propose to be associated with attention.

Our work presented here can be related to both assumptions about the computational role of attention, and as such cannot reconcile this dispute. Note, that the competitive attractor dynamics can be seen both as an approximation to the exact inference (the attractor dynamics regulates the update of beliefs by assigning the computational resources only to the most relevant hypothesis) and as a suppressor of the perceptual surprise (the attractor dynamics actively reduces the uncertainty about future sensory stimuli by predicting both the future expectation and precision of a categorical probability of hypothesis relevance; see Eq (9)).

Potential limitations of the experimental design

We believe that the probabilistic WCST provides a promising experimental paradigm for investigating complex behavioral models. However, one can probably improve on the current

design using two changes. Firstly, in spite of the initial training, several subjects exhibit rather poor performance in the no-switch condition (see Fig 2). Ten out of twenty two subjects show poor performance in at least one experimental block of the no-switch condition. Importantly, we have included these subjects in our analysis, because the model comparison did not show any correlation between subjects' performance and the best fitting behavioral model. Also note that a key strength of the proposed model is that it can explain this poor performance well, see for example Fig 10; insofar a potentially suboptimal performance does not pose a limitation to the proposed modelling approach. However, the obtained results may be even more compelling if subjects practiced the task until a stable performance is reached for both conditions. Secondly, as mentioned in the Methods section, the error rate ε was set to values that induced the most distinct behavioral responses between two experimental conditions, while rendering the switch condition informative enough to induce betting responses in subjects. However, these led to a partially imbalanced manipulation between conditions. Thus, a potential improvement would be to introduce a fractal design, such that both the error rate and the switch probability are incrementally increased. Such a fractal design would provide further insights into how each environmental parameter influences behavior and what effects, if any, each parameter might have on the model comparison.

Limitations of the analytical method

Similar to the experimental design, the analytical approach presented here may also be potentially improved upon. Firstly, as mentioned in the Methods section, the behavioral model proposed here is not the only possible formulation. Depending on how one defines the observation likelihood (Eq (4)) and the parametrization of the hypothesis probability (Eq (5)), one can obtain different variants of the perceptual model. Although we have tested a couple of them (one additional, alternative formulation is described in S1 Text), there is a large number of possible perceptual models. We anticipate that more studies are required to come to a general conclusion which of the models or model families is the most useful for describing behavioral data of studies similar to the one presented here. Secondly, the model comparison presented here relies solely on the Bayesian model selection that is useful for inferring which of the given models is most likely to generate the data. However, it cannot be directly used to answer the question whether a given model is a good predictor of behavior. To address this question one has to rely on cross-validation strategies, that is, on model testing [91]. Still, one important prior assumption of model testing is that the behavior can be described by parameters which are stable over blocks. We do not assume that this is the case for our experimental data as subjects were not over-trained which would motivate the assumption that subjects performed the task in some stable parameter regime. Thus, it is plausible that the experience in previous experimental blocks influences, at least slightly, the behavior in subsequent blocks. For this reason model testing may not be usefully applicable to our study. Nevertheless, for future studies changes to the training procedure may stabilize behavior across experimental blocks and would allow one to also apply model testing methods to predict behavior.

Neuroimaging application

Although the presented analysis has been applied to behavioral data only, it would be potentially useful and feasible to extend the behavioral analysis to the investigation of neuroimaging data. The inferred belief trajectories would be used as regressors [13], and thus can provide insights into the functional aspects of specific brain areas involved in the decision making process during the ongoing task.

Conclusion

We found strong evidence that an attention-like mechanism modulates the update of beliefs in human subjects who had to infer the relevance of various features in a dynamic and noisy environment. Effectively, this attentional focus facilitates the increase of expectations about the relevant feature and inhibits the expectations about irrelevant features. Subsequently, these modulated expectations affect update of beliefs. We expect that the same computational mechanism can be applied to modelling other complex tasks that impose high cognitive load on subjects, thus require the attentional focus strategies for decision making.

Supporting Information

S1 Text. An alternative formulation of the perceptual model. Contains derivations of an alternative perceptual model (and reduced model variants) and also contains the results of the model comparison.
(PDF)

S2 Text. Response model derivation. Contains detailed derivation of the response model within the framework of Bayesian decision theory.
(PDF)

S1 Data. Collection of data files. Contains behavioral data, posterior and prior expectation (and covariance matrix) of the free model parameters, estimated log-model evidence for each behavioral model, and the model comparison results.
(GZ)

S1 Fig. Correlations between the expected model performance and the measured subjects' performance. Pearson correlation coefficient r_{corr} between the mean subject performance and the mean model performance, for each behavioral model in the switch (top) and no-switch condition (bottom).
(TIFF)

Acknowledgments

We thank Sebastian Bitzer and Daniel McNamee for helpful discussions and comments on earlier versions of the manuscript. We also thank the Center of Information Services and High Performance Computing (ZIH) at Technische Universität Dresden for providing the computer resources.

Author Contributions

Conceived and designed the experiments: JG JO PB. Performed the experiments: JG. Analyzed the data: DM SJK. Wrote the paper: DM SJK.

References

1. Wilson RC, Niv Y (2011) Inferring relevance in a changing world. Frontiers in human neuroscience 5.
2. Roberts AC, Robbins TW, Weiskrantz LE (1998) The prefrontal cortex: Executive and cognitive functions: Oxford University Press.
3. Milner B (1963) Effects of different brain lesions on card sorting: The role of the frontal lobes. Archives of Neurology 9: 90–00.
4. Drewe E (1974) The effect of type and area of brain lesion on Wisconsin Card Sorting Test performance. Cortex 10: 159–170. PMID: 4844468

5. Nelson HE (1976) A modified card sorting test sensitive to frontal lobe defects. Cortex 12: 313–324. PMID: 1009768
6. Robinson AL, Heaton RK, Lehman RA, Stilson DW (1980) The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. Journal of consulting and clinical psychology 48: 605. PMID: 7410659
7. Robbins TW, Weinberger D, Taylor J, Morris R (1996) Dissociating executive functions of the prefrontal cortex [and discussion]. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 351: 1463–1471.
8. Rougier NP, O'Reilly RC (2002) Learning representations in a gated prefrontal cortex model of dynamic task switching. Cognitive Science 26: 503–520.
9. Kaplan GB, Şengör NS, Gürvit H, Genç İ, Güzeliş C (2006) A composite neural network model for perseveration and distractibility in the Wisconsin card sorting test. Neural Networks 19: 375–387. PMID: 16343846
10. Bishara AJ, Kruschke JK, Stout JC, Bechara A, McCabe DP, et al. (2010) Sequential learning models for the Wisconsin card sort task: Assessing processes in substance dependent individuals. Journal of mathematical psychology 54: 5–13. PMID: 20495607
11. Stern ER, Gonzalez R, Welsh RC, Taylor SF (2010) Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. The Journal of neuroscience 30: 8032–8041. doi: 10.1523/JNEUROSCI.4729-09.2010 PMID: 20534851
12. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. Nat Neurosci 10: 1214–1221. PMID: 17676057
13. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66: 585–595. doi: 10.1016/j.neuron.2010.04.016 PMID: 20510862
14. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, et al. (2013) Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron 80: 519–530. doi: 10.1016/j.neuron.2013.09.009 PMID: 24139048
15. Dehaene S, Changeux J-P (1991) The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. Cerebral cortex 1: 62–79. PMID: 1822726
16. Berdia S, Metz J (1998) An artificial neural network stimulating performance of normal subjects and schizophrenics on the Wisconsin card sorting test. Artificial intelligence in medicine 13: 123–138. PMID: 9654382
17. Morton JB, Munakata Y (2002) Active versus latent representations: A neural network model of perseveration, dissociation, and decalage. Developmental psychobiology 40: 255–265. PMID: 11891637
18. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. Proceedings of the National Academy of Sciences of the United States of America 102: 7338–7343. PMID: 15883365
19. Stemme A, Deco G, Busch A, Schneider WX (2005) Neurons and the synaptic basis of the fMRI signal associated with cognitive flexibility. Neuroimage 26: 454–470. PMID: 15907303
20. Guigon E, Dorizzi B, Burnod Y, Schultz W (1995) Neural correlates of learning in the prefrontal cortex of the monkey: a predictive model. Cerebral Cortex 5: 135–147. PMID: 7620290
21. Dehaene S, Changeux JP (1995) Neuronal models of prefrontal cortical functions. Annals of the New York Academy of Sciences 769: 305–320. PMID: 8595034
22. Houghton G (2005) Connectionist models in cognitive psychology: Psychology Press.
23. Thomas MS, McClelland JL (2008) Connectionist models of cognition. The Cambridge handbook of computational psychology: 23–58.
24. O'Reilly RC, Herd SA, Pauli WM (2010) Computational models of cognitive control. Current opinion in neurobiology 20: 257–261. doi: 10.1016/j.conb.2010.01.008 PMID: 20185294
25. Dehaene S, Changeux J-P (1997) A hierarchical neuronal network for planning behavior. Proceedings of the National Academy of Sciences 94: 13293–13298.
26. Weiss Y, Simoncelli EP, Adelson EH (2002) Motion illusions as optimal percepts. Nat Neurosci 5: 598–604. PMID: 12021763
27. Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27: 712–719. PMID: 15541511
28. Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends in cognitive sciences 10: 319–326. PMID: 16807063
29. Norris D (2006) The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. Psychological Review 113: 327. PMID: 16637764

30. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, et al. (2007) Causal inference in multisensory perception. PLoS one 2: e943. PMID: 17895984

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

32. Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, et al. (2013) Spatial attention, precision, and bayesian inference: A study of saccadic response speed. Cerebral Cortex: bhs418.

33. Feldman H, Friston KJ (2010) Attention, uncertainty, and free-energy. Frontiers in human neuroscience 4.

34. Whiteley L, Sahani M (2012) Attention in a Bayesian framework. Frontiers in human neuroscience 6.

35. Koechlin E (2014) An evolutionary computational theory of prefrontal executive function in decision-making. Philosophical Transactions of the Royal Society of London B: Biological Sciences 369: 20130474. doi: 10.1098/rstb.2013.0474 PMID: 25267817

36. Chikkerur S, Serre T, Tan C, Poggio T (2010) What and where: A Bayesian inference theory of attention. Vision research 50: 2233–2247. doi: 10.1016/j.visres.2010.05.013 PMID: 20493206

37. Fang Y, Cohen MA, Kincaid TG (1996) Dynamics of a winner-take-all neural network. Neural Networks 9: 1141–1154. PMID: 12662589

38. Gros C (2009) Cognitive computation with autonomously active neural networks: an emerging field. Cognitive Computation 1: 77–90.

39. Kaski S, Kohonen T (1994) Winner-take-all networks for physiological models of competitive learning. Neural Networks 7: 973–984.

40. Maass W (2000) On the computational power of winner-take-all. Neural computation 12: 2519–2535. PMID: 11110125

41. Bitzer S, Yildiz IB, Kiebel SJ (2012) Online Discrimination of Nonlinear Dynamics with Switching Differential Equations. arXiv preprint arXiv:12110947.

42. Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychological review 108: 550. PMID: 11488378

43. Hopfield JJ, Tank DW (1985) "Neural" computation of decisions in optimization problems. Biological cybernetics 52: 141–152. PMID: 4027280

44. Summerfield C, Behrens TE, Koechlin E (2011) Perceptual classification in a rapidly changing environment. Neuron 71: 725–736. doi: 10.1016/j.neuron.2011.06.022 PMID: 21867887

45. Payzan-LeNestour E (2010) Bayesian learning in unstable settings: Experimental evidence based on the bandit problem. Swiss Finance Institute Research Paper: 1–41.

46. Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for individual learning under uncertainty. Frontiers in Human Neuroscience 5.

47. Mathys C, Daunizeau J, Iglesias S, Diaconescu AO, Weber LAE, et al. (2012) Computational modeling of perceptual inference: A hierarchical Bayesian approach that allows for individual and contextual differences in weighting of input. Int J Psychophysiol 85: 317–318.

48. Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS computational biology 4: e1000209. doi: 10.1371/journal.pcbi.1000209 PMID: 19008936

49. Daunizeau J, Den Ouden HE, Pessiglione M, Kiebel SJ, Stephan KE, et al. (2010) Observing the observer (I): meta-Bayesian models of learning and decision-making. PLoS One 5: e15554. doi: 10.1371/journal.pone.0015554 PMID: 21179480

50. Daunizeau J, Den Ouden HE, Pessiglione M, Kiebel SJ, Friston KJ, et al. (2010) Observing the observer (II): deciding when to decide. PLoS one 5: e15555. doi: 10.1371/journal.pone.0015555 PMID: 21179484

51. Acerbi L, Vijayakumar S, Wolpert DM (2014) On the Origins of Suboptimality in Human Probabilistic Inference. PLOS Computational Biology 10: e1003661. doi: 10.1371/journal.pcbi.1003661 PMID: 24945142

52. Vul E, Goodman ND, Griffiths TL, Tenenbaum JB. One and done? Optimal decisions from very few samples; 2009. pp. 66–72.

53. Vul E, Pashler H (2008) Measuring the crowd within probabilistic representations within individuals. Psychological Science 19: 645–647. doi: 10.1111/j.1467-9280.2008.02136.x PMID: 18727777

54. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46: 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025 PMID: 19306932

55. Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies—revisited. Neuroimage 84: 971–985. PMID: 24018303

56. Hansen N, Müller SD, Koumoutsakos P (2003) Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary Computation 11: 1–18. PMID: 12804094

57. Hansen N, Kern S. Evaluating the CMA evolution strategy on multimodal test functions; 2004. Springer. pp. 282–291.

58. Lyness J, Moler C (1966) van der Monde systems and numerical differentiation. Numerische Mathematik 8: 458–464.

59. Friel N, Wyse J (2012) Estimating the evidence–a review. Statistica Neerlandica 66: 288–308.

60. Geisler WS, Kersten D (2002) Illusions, perception and Bayes. Nat Neurosci 5: 508–510. PMID: 12037517

61. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55: 271–304. PMID: 14744217

62. Knill DC, Richards W (1996) Perception as Bayesian inference: Cambridge University Press.

63. Durbin J, Koopman SJ (2012) Time series analysis by state space methods: Oxford University Press.

64. Changeux J-P, Dehaene S (2000) Hierarchical neuronal modeling of cognitive functions: from synaptic transmission to the Tower of London. Int J Psychophysiol 35: 179–187. PMID: 10677646

65. Goela V, Pullara SD, Grafman J (2001) A computational model of frontal lobe dysfunction: Working memory and the Tower of Hanoi task. Cognitive Science 25: 287–313.

66. Standage DI, Trappenberg TP, Klein RM (2005) Modelling divided visual attention with a winner-take-all network. Neural networks 18: 620–627. PMID: 16087317

67. Beal MJ (2003) Variational algorithms for approximate Bayesian inference: University of London.

68. Miller RR, Barnet RC, Grahame NJ (1995) Assessment of the Rescorla-Wagner model. Psychological bulletin 117: 363. PMID: 7777644

69. Siegel S, Allan LG (1996) The widespread influence of the Rescorla-Wagner model. Psychonomic Bulletin & Review 3: 314–321.

70. Bland AR, Schaefer A (2012) Different varieties of uncertainty in human decision-making. Frontiers in neuroscience 6.

71. De Palma A, Ben-Akiva M, Brownstone D, Holt C, Magnac T, et al. (2008) Risk, uncertainty and discrete choice models. Marketing Letters 19: 269–285.

72. Kolling N, Wittmann M, Rushworth MF (2014) Multiple neural mechanisms of decision making and their competition under changing risk pressure. Neuron 81: 1190–1202. doi: 10.1016/j.neuron.2014.01.033 PMID: 24607236

73. Platt ML, Huettel SA (2008) Risky business: the neuroeconomics of decision making under uncertainty. Nat Neurosci 11: 398–403. doi: 10.1038/nn2062 PMID: 18368046

74. Heidrich-Meisner V, Igel C (2008) Evolution strategies for direct policy search. Parallel Problem Solving from Nature–PPSN X: Springer. pp. 428–437.

75. Heidrich-Meisner V, Igel C (2009) Neuroevolution strategies for episodic reinforcement learning. Journal of Algorithms 64: 152–168.

76. Hou S, Li Y (2009) Short-term fault prediction based on support vector machines with parameter optimization by evolution strategy. Expert Systems with Applications 36: 12383–12391.

77. Meng Y, Zhang Y, Jin Y (2011) Autonomous self-reconfiguration of modular robots by evolving a hierarchical mechanochemical model. Computational Intelligence Magazine, IEEE 6: 43–54.

78. Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, et al. (2010) Comparing families of dynamic causal models. PLoS computational biology 6: e1000709. doi: 10.1371/journal.pcbi.1000709 PMID: 20300649

79. Kepecs A, Mainen ZF (2012) A computational framework for the study of confidence in humans and animals. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 1322–1337.

80. Choi J, Sheu BJ (1993) A high-precision VLSI winner-take-all circuit for self-organizing neural networks. Solid-State Circuits, IEEE Journal of 28: 576–584.

81. Coultrip R, Granger R, Lynch G (1992) A cortical model of winner-take-all competition via lateral inhibition. Neural networks 5: 47–54.

82. Ermentrout B (1992) Complex dynamics in winner-take-all neural nets with slow inhibition. Neural networks 5: 415–431.

83. Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. Matters of Intelligence: Springer. pp. 115–141.

84. Lee DK, Itti L, Koch C, Braun J (1999) Attention activates winner-take-all competition among visual filters. Nat Neurosci 2: 375–381. PMID: 10204546

85. Bodegård A, Geyer S, Grefkes C, Zilles K, Roland PE (2001) Hierarchical processing of tactile shape in the human brain. Neuron 31: 317–328. PMID: 11502261

86. Wessinger C, VanMeter J, Tian B, Van Lare J, Pekar J, et al. (2001) Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. Journal of Cognitive Neuroscience 13: 1–7. PMID: 11224904

87. Friston KJ, Daunizeau J, Kilner J, Kiebel SJ (2010) Action and behavior: a free-energy formulation. Biological cybernetics 102: 227–260. doi: 10.1007/s00422-010-0364-z PMID: 20148260

88. Angela JY, Dayan P. Inference, attention, and decision in a Bayesian neural architecture; 2004. pp. 1577–1584.

89. Rao RP (2005) Bayesian inference and attentional modulation in the visual cortex. Neuroreport 16: 1843–1848. PMID: 16237339

90. Friston K (2010) The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11: 127–138. doi: 10.1038/nrn2787 PMID: 20068583

91. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Statistics surveys 4: 40–79.