A confirmation bias in perceptual decision-making due to hierarchical approximate inference

Richard D. Lange[1,2,*], Ankani Chattoraj[1], Jeffrey M. Beck[3], Jacob L. Yates[1], Ralf M. Haefner[1,*]

[1]Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.
[2]Computer Science, University of Rochester, Rochester, NY 14627, USA.
[3]Department of Neurobiology, Duke University, Durham, NC 27708, USA.
[*]Corresponding authors: rlange@ur.rochester.edu, rhaefne2@ur.rochester.edu.

January 20, 2020

Summary

Human decisions are known to be systematically biased. A prominent example of such a bias occurs when integrating a sequence of sensory evidence over time. Previous empirical studies differ in the nature of the bias they observe, ranging from favoring early evidence (primacy), to favoring late evidence (recency). Here, we present a unifying framework that explains these biases and makes novel psychophysical and neurophysiological predictions. By explicitly modeling both the approximate and the hierarchical nature of inference in the brain, we show that temporal biases depend on the balance between "sensory information" and "category information" in the stimulus. Finally, we present new data from a human psychophysics task that confirm that temporal biases can be robustly changed within subjects as predicted by our models.

Introduction

Imagine a doctor trying to infer the cause of a patient's symptoms from an x-ray image. Unsure about the evidence in the image, she asks a radiologist for a second opinion. If she tells the radiologist her suspicion, she may bias his report. If she does not, he may not detect a faint diagnostic pattern. As a result, if the evidence in the image is hard to detect or ambiguous, the radiologist's second opinion, and hence the final diagnosis, may be swayed by the doctor's initial hypothesis. The problem faced by these doctors exemplifies the difficulty of *hierarchical inference*: each doctor's suspicion both informs and is informed by their collective diagnosis. If they are not careful, their diagnosis may fall prey to circular reasoning. The brain faces a similar problem during perceptual decision-making: any decision-making area combines sequential signals from sensory brain areas, not directly from sensory input, just as the doctors' consensus is based on their individual diagnoses rather than on the evidence *per se*. If sensory signals in the brain themselves reflect inferences that combine both prior expectations and sensory evidence, we suggest that this can then lead to an observable *perceptual* confirmation bias (Nickerson, 1998).

We formalize this idea in the context of approximate Bayesian inference and classic evidence-integration tasks in which a range of biases has been observed and for which a unifying explanation is currently lacking. Evidence-integration tasks require subjects to categorize a sequence of independent and identically distributed (iid) draws of stimuli (Gold and Shadlen, 2007; Bogacz et al., 2006). Previous normative models of evidence integration hinge on two quantities: the amount of information available on a single stimulus draw and the total number of draws. One might expect, then, that temporal biases should have some canonical form in tasks where these quantities are matched. However, existing studies are heterogeneous, reporting one of three distinct motifs: some find that early evidence is weighted more strongly (a primacy effect) (Kiani et al., 2008; Nienborg and Cumming, 2009) some that information is weighted equally over time (as would be optimal) (Wyart et al., 2012; Brunton et al., 2013; Raposo et al., 2014), and some find late evidence being weighted most heavily (a recency effect) (Drugowitsch et al., 2016) (Figure 1a,c). While there are myriad differences between these studies such as subject species, sensory modality, stimulus parameters, and computational frameworks (Kiani et al., 2008; Brunton et al., 2013; Glaze et al., 2015; Drugowitsch et al., 2016), none of these aspects alone can explain their different findings.

We extend classic evidence-integration models to the *hierarchical* case by including an explicit intermediate sensory representation, analogous to modeling each doctor's individual diagnosis in addition to their consensus in the example above (Figure 1b). Taking this intermediate inference stage into account makes explicit that task difficulty is modulated by two distinct types of information exposing systematic differences between existing tasks: the information between the stimulus and sensory representation ("sensory information"), and the information between sensory representation and category ("category information") (Figure 1b). These differences alone do not entail any bias as long as inference is exact. However, inference in the brain is necessarily *approximate* and this approximation can interfere with its ability to account for its own biases. Implementing two approximate hierarchical inference algorithms, we find that they both result in biases in agreement with our data, and can indeed explain the puzzling discrepancies in the literature.

Results

"Sensory Information" vs "Category Information"

Normative models of decision-making in the brain are typically based on the idea of an *ideal observer*, who uses Bayes' rule to infer the most likely category on each trial given the stimulus. On each trial in a typical task, the stimulus consists of multiple "frames" presented in rapid succession. (By "frames" we refer to discrete independent draws of stimulus values that are not necessarily visual). If the evidence in each frame, e_f, is independent then evidence can be combined by simply multiplying the associated likelihoods. And if the categorical identity of the stimulus is a binary variable $C \in \{-1, +1\}$, then this process corresponds to the famous sequential ratio test summing the log odds implied by each piece of evidence (Wald and Wolfowitz, 1948; Bogacz et al., 2006):

$$\mathrm{p}(C=+1|e_1,\dots,e_F) \propto \mathrm{p}(C=+1)\prod_{f=1}^{F}\mathrm{p}(e_f|C=+1)$$

$$\log\frac{\mathrm{p}(C=+1|e_1,\dots,e_F)}{\mathrm{p}(C=-1|e_1,\dots,e_F)} = \log\frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)} + \sum_{f=1}^{F}\log\frac{\mathrm{p}(e_f|C=+1)}{\mathrm{p}(e_f|C=-1)}.$$

As a result, the ideal observer's performance is determined by (i) the information about C available on each frame, $\mathrm{p}(e_f|C)$, and (ii) the number of frames per trial.

However, in the brain, any decision-making area does not base its decision on the externally presented stimulus directly, but rather on an intermediate sensory representation of the stimulus.

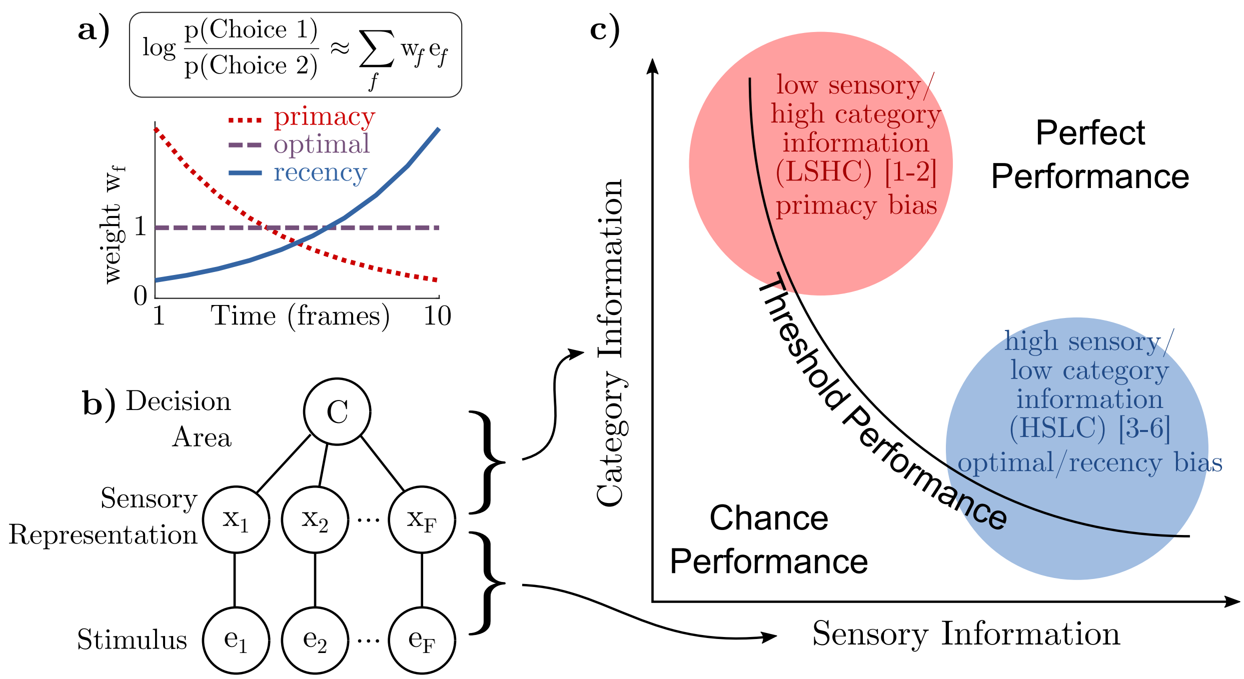

Figure 1: **a)** A subject's "temporal weighting strategy" is an estimate of how their choice is based on a weighted sum of each frame of evidence e_f. Three commonly observed motifs are decreasing weights (primacy), constant weights (optimal), or increasing weights (recency). **b)** Information in the stimulus about the category may be decomposed into information in each frame about a sensory variable ("sensory information") and information about the category given the sensory variable ("category information"). **c)** Category information and sensory information may be manipulated independently, creating a two-dimensional space of possible tasks. Any level of task performance can be the result of different combinations of sensory and category information. A qualitative placement of previous work into this space separates those that find primacy effects in the upper-left from those that find recency effects or optimal weights in the lower right (see Supplemental Text for detailed justification). Numbered references are: [1] Kiani et al., [2] Nienborg and Cumming, [3] Brunton et al., [4] Wyart et al., [5] Raposo et al., [6] Drugowitsch et al.

This intermediate representation is itself often assumed to the result an inference process in which sensory neurons compute the posterior distribution $\mathrm{p}(\mathbf{x}|e)$ (Fiser et al., 2010; Pouget et al., 2013; Gershman and Beck, 2016; Lange and Haefner, 2020) over some latent variable $\mathbf{x}$ given the external evidence e in an internal model of the world (Mumford, 1992; Lee and Mumford, 2003; Yuille and Kersten, 2006). This process is naturally formalized as hierarchical inference (Figure 1b). This implies that the information between the stimulus and category can be partitioned into the information between the stimulus and the sensory representation (e to $\mathbf{x}$), and the information between sensory representation and category ($\mathbf{x}$ to C). We call these "sensory information" and "category information," respectively (Figure 1b). These two kinds of information define a two-dimensional space in which a given task is located as a single point (Figure 1c). For example, in a visual task each e_f would be the image on the screen while x_f might be image patches that are assumed to be sparsely combined to form the image (Olshausen and Field, 1997). The posterior over the latent features x_f would be represented by the activity of relevant neurons in visual cortex.

An evidence integration task may be challenging either because each frame is perceptually unclear (low "sensory information"), or because each frame alone is insufficient to determine the category for the whole trial (low "category information"). Consider the classic dot motion task (Newsome and Pare, 1988) and the Poisson clicks task (Brunton et al., 2013), which occupy opposite locations in the space. In the classic low-coherence dot motion task, subjects view a cloud of moving dots, a small percentage of which move "coherently" in one direction. Here, sensory information is low since the percept of net motion is weak on each frame. Category information, on the other hand, is high, since knowing the true net motion on a single frame would be highly predictive of the correct choice (and of motion on subsequent frames). In the Poisson clicks task on the other hand, subjects hear a random sequence of clicks in each ear and must report the side with the higher rate. Here, sensory information is high since each click is well above sensory thresholds. Category information, however, is low, since knowing the side on which a single click was presented provides only little information about the correct choice for the trial as a whole (and the side of the other clicks). Another way to think about category information is as "temporal coherence" of the stimulus: the more each frame of evidence is predictive of the correct choice, the more the frames must be predictive of each other, whether a frame consists of visual dots or of auditory clicks. Note that our distinction between sensory and category information is different from the well-studied distinction between internal and external noise; in general, both internal and external noise will reduce the amount of sensory and category information.

Optimal inference requires accounting for all possible sources of information. Ideally, then, sensory areas would not only represent the current evidence, $\mathrm{p}(x_f|e_f)$, but should incorporate prior information based on previous frames to compute $\mathrm{p}(x_f|e_1,\dots,e_f)$. While the sensory area no longer has direct access to the earlier frames, this is mathematically equivalent to using the current belief in the category C as a prior:

$$\mathrm{p}(x_f|e_1,\dots,e_f) \propto \mathrm{p}(e_f|x_f)\sum_c \underbrace{\mathrm{p}(C=c|e_1,\dots,e_{f-1})}_{\mathrm{p}_{f-1}(C)} p(x_f|C=c). \tag{1}$$

Mechanistically, this suggests that the brain's running estimate of the category, $p_{f-1}(C)$, should be continuously fed back to sensory areas, acting as a prior that biases the representation to agree with the current belief about the category (Lee and Mumford, 2003; Haefner et al., 2016; Tajima et al., 2016; Lange and Haefner, 2020). Importantly, such a bias is optimal in the sense that it makes instantaneous sensory estimates more accurate. Despite this instantaneous sensory bias, *exact* inference in this model does *not* induce any bias in the posterior over the category C. That is, although the ideal observer's inference about x_f is biased by $e_1,\dots,e_{f-1}$, this bias is removed by precisely accounting for it in the update to $p_f(C)$ (Zylberberg et al., 2018).

Unlike the ideal observer, inference is in the brain is necessarily *approximate* ((Fiser et al., 2010; Pouget et al., 2013) and the implications of this fact on evidence integration has so far been unknown. Below, we consider two models, each implementing approximate hierarchical inference in one of the two major classes of approximate inference schemes known from statistics and machine learning: sampling-based and variational inference (Bishop, 2006; Murphy, 2012), both of which have been previously proposed models for neural inference (Fiser et al., 2010; Pouget et al., 2013). In both models, a confirmation bias arises as a direct consequence of the approximate nature of inference over the intermediate sensory variables in the brain. The strength of the predicted confirmation bias depends directly on the amount of category information in the stimulus, since that governs how strongly past frames inform inferences about the present frame. Our models predict an overweighting of early evidence when sensory information is low and category information is high, but not when sensory information is high and category information is low, even when performance is matched in both conditions (Fig. 1c, for model details see "Approximate inference models" section below).

Qualitatively placing prior studies in the space spanned by these two kinds of information results in two clusters: the studies that report primacy effects are located in the upper left quadrant (low-sensory/high-category or LSHC) and studies with flat weighting or recency effects are in the lower right quadrant (high-sensory/low-category or HSLC) (Figure 1c). This initially suggests that the trade-off between sensory information and category information may indeed underlie differences in temporal weighting seen in previous studies. Further, this framework allows us to make new and easily testable predictions for how simple changes in stimulus statistics of previous studies should change the temporal weighting they find (Supplemental Table S1). We next describe a novel set of visual discrimination tasks designed to directly probe this trade-off between sensory information and category information to test these predictions within individual subjects.

Visual Discrimination Task

We designed a visual discrimination task with two stimulus conditions that correspond to the two opposite sides of this task space, while keeping all other aspects of the design the same (Figure 2a). If our theory is correct, then we should be able to change individual subjects' temporal weighting strategy simply by changing the sensory-category information trade-off.

The stimulus in our task consisted of a sequence of ten visual frames (83ms each). Each frame consisted of band-pass-filtered white noise with excess orientation power either in the $-45°$ or the $+45°$ orientation (Beaudot and Mullen, 2006) (Figure 2b,d). On each trial, there was a single true orientation category, but individual frames might differ in their orientation. At the end of each trial, subjects reported whether the stimulus was oriented predominantly in the $-45°$ or the $+45°$ orientation. The stimulus was presented as an annulus around the fixation marker in order to minimize the effect of small fixational eye movements (Methods).

If the brain's intermediate sensory representation reflects the orientation in each frame, then sensory information in our task is determined by how well each frame determines the orientation of that frame (i.e. the amount of "noise" in each frame), and category information is determined by the probability that any given frame's orientation matches the trial's category. We chose to quantify both sensory information and category information, using signal detection theory, as the area under the receiver-operating-characteristic curve for e_f and x_f (sensory information), or for x_f and C (category information). Hence for a ratio of $5:5$, a frame's orientation does not predict the correct choice and category information is 0.5. For a ratio of $10:0$, knowledge of the orientation of

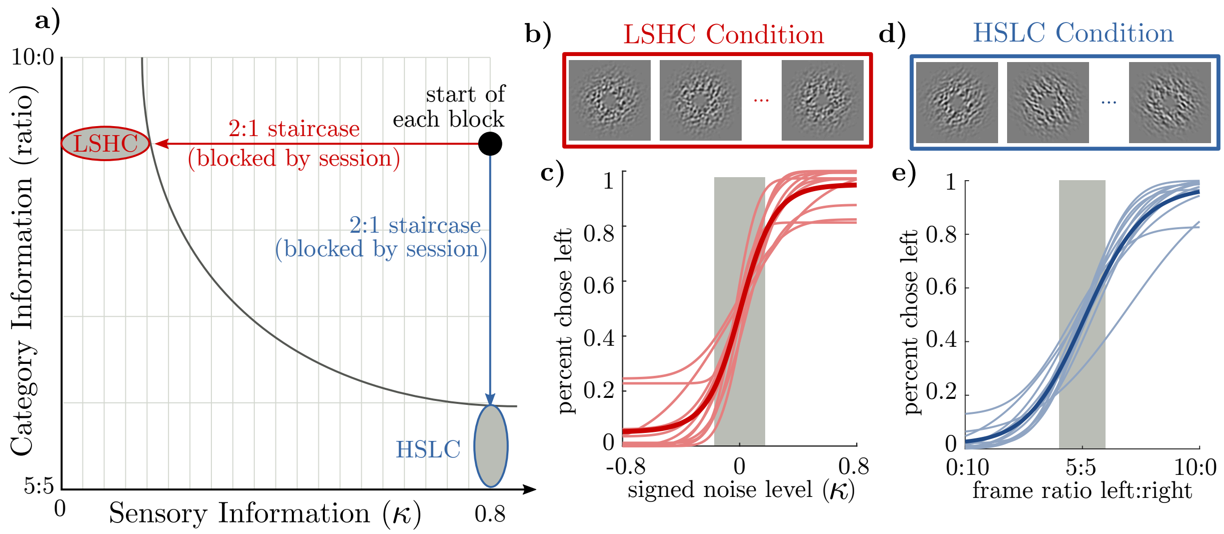

Figure 2: Summary of experiment design. **a)** Category information is determined by the expected ratio of frames in which the orientation matches the correct category, and sensory information is determined by a parameter κ determining the degree of spatial orientation coherence (Methods). At the start of each block, we reset the staircase to the same point, with category information at $9 : 1$ and κ at 0.8. We then ran a 2-to-1 staircase either on κ or on category information. The LSHC and HSLC ovals indicate sub-threshold trials; only these trials were used in the regression to infer subjects' temporal weights. **b)** Visualization of a noisy stimulus in the LSHC condition. All frames are oriented to the right. **c)** Psychometric curves for all subjects (thin lines) and averaged (thick line) over the κ staircase. Shaded gray area indicates the median threshold level across all subjects. **d)** Example frames in the HSLC condition. The orientation of each frame is clear, but orientations change from frame to frame. **e)** Psychometric curves over frame ratios, plotted as in (c).

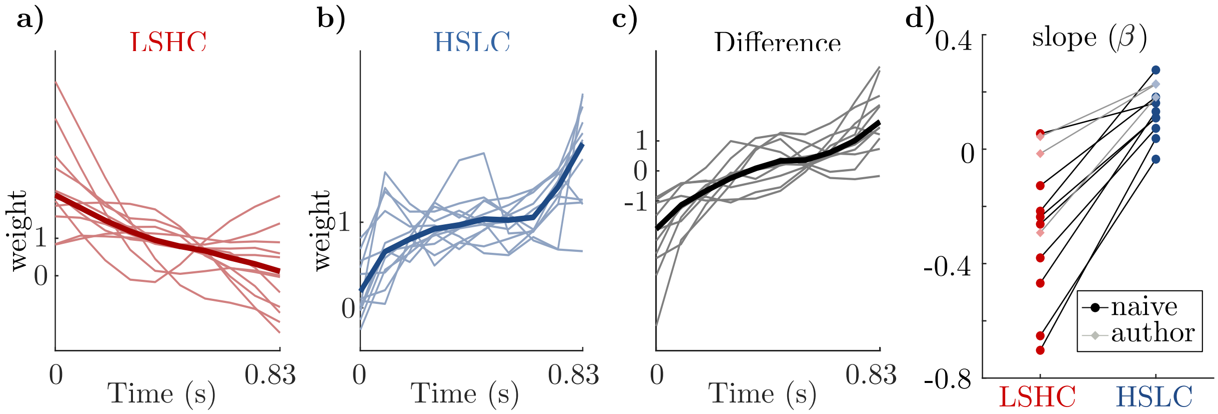

Figure 3: Subjects' temporal weights. **a-b)** Temporal weights for individual subjects (thin lines) and the mean across all subjects (thick lines). Weights are normalized to have a mean of 1 to emphasize shape rather than magnitude. Individual subjects' curves were fit using a cross-validated smoothness term for visualization purposes only (Methods). **c)** Difference of normalized weights (HSLC−LSHC). Despite variability across subjects in (a-b), each subject reliably changes in the direction of a recency effect. **d)** *Change* in slope between the two task contexts for each subject is consistently positive. Points are median slope values after bootstrap-resampling of the data. We summarize subjects' temporal weighting strategy with an exponential fit; the slope parameter $\beta > 0$ corresponds to recency and $\beta < 0$ to primacy (similar results for linear fits, see SI).

a single frame is sufficient to determine the correct choice and category information is 1. Exactly quantifying sensory information depends on individual subjects, but likewise ranges from 0.5 to 1. For a more detailed discussion, see Supplementary Text.

Using this stimulus, we tested 12 human subjects (9 naive and 3 authors) comparing two conditions intended to probe the difference between the LSHC and HSLC regimes. Starting with both high sensory and high category information, we either ran a 2:1 staircase lowering the sensory information while keeping category information high, or we ran a 2:1 staircase lowering category information while keeping sensory information high (Figure 2a). These are the LSHC and HSLC conditions, respectively (Figure 2b,d). For each condition, we used logistic regression to infer, for each subject, the influence of each frame onto their choice. Subjects' overall performance was matched in the two conditions by setting a performance threshold below which trials were included in the analysis (Methods).

In agreement with our hypothesis, we find predominantly flat or decreasing temporal weights in the LSHC condition (Figure 3a). However, when the information is partitioned differently – in the HSLC condition – we find flat or increasing weights (Figure 3b). Importantly, despite variability between subjects in each condition, a within-subject comparison revealed that the change in slope between the two conditions was as predicted for all subjects (Figure 2c,d) ($p < 0.05$ for 10 of 12 subjects, bootstrap). This demonstrates that the trade-off between sensory and category information in a task robustly changes subjects' temporal weighting strategy as we predicted, and further suggests that the sensory-category information trade-off may resolve the discrepant results in the literature.

Approximate inference models

We will now show that these significant changes in evidence weighting for different stimulus statistics arise naturally in common models of how the brain might implement approximate inference. In particular, we show that both a neural sampling-based approximation (Hoyer and Hyvärinen, 2003; Fiser et al., 2010; Haefner et al., 2016; Orbán et al., 2016) and a parametric (mean-field) approximation (Beck et al., 2013; Raju and Pitkow, 2016) can explain the observed pattern of changing temporal weights as a function of stimulus statistics.

Optimal inference in our task, as in other evidence integration tasks, requires computing the posterior over C conditioned on the evidence $e_1, \ldots, e_f$, which can be expressed as the Log Posterior Odds (LPO),

$$\underbrace{\log \frac{\mathrm{p}(C=+1|e_1,\ldots,e_f)}{\mathrm{p}(C=-1|e_1,\ldots,e_f)}}_{\mathrm{LPO}_f} = \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)} + \sum_{i=1}^{f} \underbrace{\log \frac{\mathrm{p}(e_i|C=+1)}{\mathrm{p}(e_i|C=-1)}}_{\mathrm{LLO}_i}, \tag{2}$$

where LLO_f is the log likelihood odds for frame f (Gold and Shadlen, 2007; Bogacz et al., 2006). To reflect the fact that the brain has access to only one frame of evidence at a time, this can be rewritten this as an *online* update rule, summing the previous frame's log posterior with new evidence gleaned on the current frame:

$$\mathrm{LPO}_f = \mathrm{LPO}_{f-1} + \mathrm{LLO}_f. \tag{3}$$

This expression is derived from the ideal observer and is still exact. Since the ideal observer weights all frames equally, the *online* nature of inference in the brain cannot by itself explain temporal biases. Furthermore, because performance is matched in the two conditions of our experiment, their differences cannot be explained by the total amount of information, governed by the likelihood $\mathrm{p}(e_f|C)$.

To understand how biases arise, we must examine the log likelihood odds term, LLO, in detail. In a hierarchical model, computing $\mathrm{p}(e_f|C)$ for each C requires marginalizing over the intervening x_f as follows:

$$\begin{aligned} \mathrm{p}(e_f|C) &\propto \int \mathrm{p}(e_f|x_f)\mathrm{p}(x_f|C)\mathrm{d}x_f \\ &\propto \mathbb{E}_{\mathrm{p}(x_f|e_f)}\left[\frac{\mathrm{p}(x_f|C)}{\mathrm{p}(x_f)}\right], \end{aligned} \tag{4}$$

This suggests that evidence about the current frame is formed in a two step process: first, x_f is inferred given e_f, and second an expectation is taken with respect to $\mathrm{p}(x_f|e_f)$, where the operand of the expectation depends only on the relation between x_f and C. No sub-optimalities nor biases have been introduced yet.

A key assumption in our models that gives rise to temporal biases is that sensory areas represent the approximate *posterior* belief over x_f given all available information, i.e. including the earlier frames in the trial (equation (1)). This assumption differs from some models of inference in the brain that assume populations of sensory neurons strictly encode the *likelihood* of the stimulus (or instantaneous posterior) (Ma et al., 2006; Beck et al., 2008), but is consistent with other models from both sampling and parametric families (Berkes et al., 2011; Haefner et al., 2016; Raju and Pitkow, 2016; Tajima et al., 2016).

As introduced in equation (1), representing the full posterior over x_f implies taking into account all previous frames. In other words, the brain's belief about x_f depends both on the external evidence, e_f, via the likelihood, but also on the brain's current belief about C, via the prior. If this were not the case – if sensory areas represented only the instantaneous evidence $\mathrm{p}(e_f|x_f)$ – then integrating evidence in an unbiased way would simply be a matter of applying equation (4). However, such an inference scheme comes at the expense of a worse instantaneous representation (Zylberberg et al., 2018).

There is thus tension between inferences at two timescales. Instantaneously, it seems advantageous to represent $\mathrm{p}(x_f|e_1, \ldots, e_f)$, while integrating evidence online requires an expectation taken with respect to $\mathrm{p}(x_f|e_f)$ (equation (4)). Assuming that the former is represented by sensory areas, the decision area of an approximate ideal observer now needs to correct for, or "subtract out" its influence on those sensory responses. Approximations either to the posterior $\mathrm{p}(x_f|e_1, \ldots, e_f)$ itself or to the bias-correction may underlie the observed behavioral biases. To test this, we implemented approximate hierarchical online inference (where "online" means observing a single frame at a time) for a discrimination task using two previously proposed frameworks for how inference might be implemented in neural circuits: neural sampling (Hoyer and Hyvärinen, 2003; Fiser et al., 2010; Haefner et al., 2016; Orbán et al., 2016) and mean field variational inference (Beck et al., 2013; Raju and Pitkow, 2016) (Figure 4).

Sampling model

The neural sampling hypothesis states that variable neural activity over brief time periods can be interpreted as a sequence of samples from the brain's posterior over latent variables in its internal model. In our model, samples of x_f are drawn from the full posterior having incorporated the running estimate of $\mathrm{p}_f(C)$ (equation (1), Methods), but from equation (4) we would like to use these samples to compute an expectation with respect to only the instantaneous evidence, $\mathrm{p}(x_f|e_f)$. The canonical way to compute an expectation with respect to one distribution using samples from another is "importance sampling," which weights each sample so as to adjust for the difference between the two distributions (Shi and Griffiths, 2009; Murphy, 2012, Chapter 23). In the most extreme case of continual online updates, one could imagine that the brain computes each update to $\mathrm{p}_f(C)$ after observing a single sample of x_f. In this case, no correction would be possible; a downstream area would be unable to recover the instantaneous distribution $\mathrm{p}(x_f|e_f)$ from a sample sample from the full posterior $\mathrm{p}(x_f|e_1, \ldots, e_f)$. If the brain is able to base each update on multiple samples, then the *importance weights* of each sample in the update account for the discrepancy between $\mathrm{p}(x_f|e_f)$ and $\mathrm{p}(x_f|e_1, \ldots, e_f)$ (Methods). While this approach is unbiased in the limit of infinitely many samples, it incurs a bias for a finite number – the relevant regime for the brain (Owen, 2013). The bias is *as if* the expectation in (4) is taken with respect to an intermediate distribution that lies between the fully biased one ($\mathrm{p}(x_f|e_1, \ldots, e_f)$) and the unbiased one ($\mathrm{p}(x_f|e_f)$) (Cremer et al., 2017).

Under-correcting for the prior that was fed back results in a positive feedback loop between decision-making and sensory areas which we call a "perceptual confirmation bias." Importantly, this feedback loop is strongest when category information is high, corresponding to stronger feedback, and sensory information is low, since that makes x_f less dependent on e_f. Figure 4b and Supplemental Figure S5a-c show performance for the ideal observer and for the resulting sampling-based model, respectively, across all combinations of sensory and category information. White lines show threshold performance (70% correct) as in Figure 1c.

This model reproduces the primacy effect, and how the temporal weighting changes as the stimulus information changes seen in previous studies. Importantly, it predicted the same within-

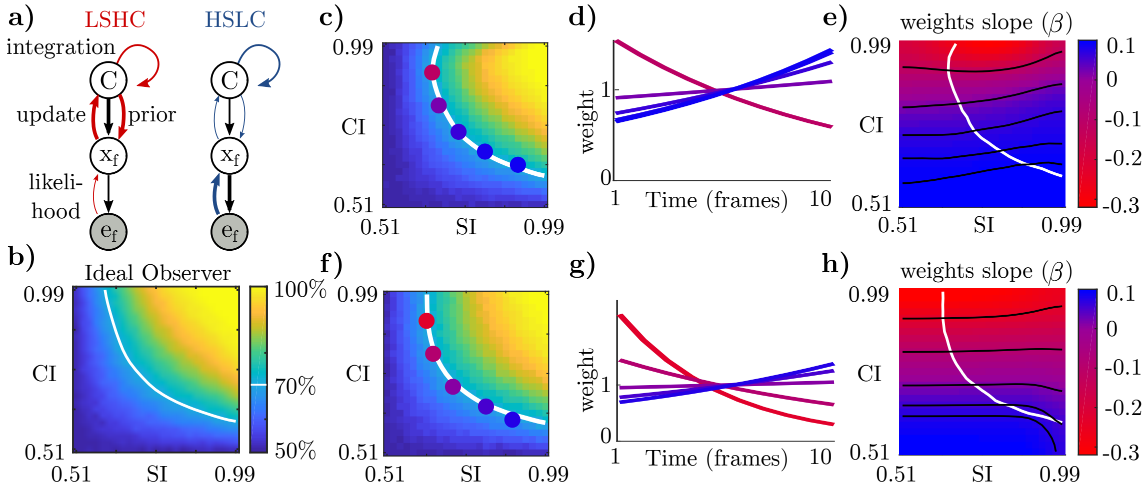

Figure 4: Approximate inference models explain results. **a)** The difference in stimulus statistics between HSLC and LSHC trade-offs implies that the relevant sensory representation is differentially influenced by the stimulus or by beliefs about the category C. A "confirmation bias" or feedback loop between x and C emerges in the LSHC condition but is mitigated in the HSLC condition. Black lines indicate the underlying generative model, and red/blue lines indicate information flow during inference. Arrow width represents coupling strength. **b)** Performance of an ideal observer reporting C given ten frames of evidence. White line shows threshold performance, defined as 70% correct. **c)** Performance of the sampling model with $\gamma = 0.1$. Colored dots correspond to lines in the next panel. **d)** Temporal weights in the model transition from recency to a strong primacy effect, all at threshold performance, as the stimulus transitions from the high-sensory/low-category to the low-sensory/high-category conditions. **e)** Using the same exponential fit as used with human subjects, visualizing how temporal biases change across the entire task space. Red corresponds to primacy, and blue to recency. White contour as in (c). Black lines are iso-contours for slopes corresponding to highlighted points in (c). **f-h)** Same as **c-d** but for the variational model with $\gamma = 0.1$.

subject change seen in our data (Haefner et al., 2016). However, double-counting the prior alone cannot explain recency effects (Supplemental Figure S5a-c,j-l).

There are two simple and biologically-plausible explanations for the observed recency effect which turn out to be mathematically equivalent. First,the brain may try to actively compensate for the prior influence on the sensory representation by subtracting out an estimate of that influence. That is, the brain could do approximate bias correction to mitigate the effect of the confirmation bias. We modeled linear bias correction by explicitly subtracting out a fraction of the running posterior odds at each step:

$$\mathrm{LPO}_f \leftarrow \mathrm{LPO}_{f-1}(1-\gamma) + \hat{\mathrm{LLO}}_f \tag{5}$$

where $0 \leq \gamma \leq 1$ and $\hat{\mathrm{LLO}}_f$ is the model's (biased) estimate of the log likelihood odds. Second, the brain may assume a non-stationary environment, i.e. C is not constant over a trial. Interestingly, Glaze et al. (2015) showed that optimal inference in this case implies equation (5), which can be interpreted as a noiseless, discrete time version of the classic drift-diffusion model (Gold and Shadlen, 2007) with γ as a leak parameter.

Incorporating equation (5) into our model reduces the primacy effect in the upper left of the task space and leads to a recency effect in the lower right (Figure 4c-e, Supplemental Figure S5), as seen in the data.

Variational model

The second major class of models for how probabilistic inference may be implemented in the brain – based on mean-field parametric representations (Ma et al., 2006; Beck et al., 2013) – behaves similarly. These models commonly assume that distributions are encoded *parametrically* in the brain, but that the brain explicitly accounts for dependencies only between subsets of variables, e.g. within the same cortical area. (Raju and Pitkow, 2016). We therefore make the assumption that the joint posterior $\mathrm{p}(x, C|e)$ is approximated in the brain by a product of parametric distributions, $\mathrm{q}(x)\mathrm{q}(C)$ (Beck et al., 2013; Raju and Pitkow, 2016). Inference proceeds by iteratively minimizing the Kullback-Leibler divergence between $\mathrm{q}(x)\mathrm{q}(C)$ and $\mathrm{p}(x, C|e)$ (Methods). As in the sampling model, the current belief about the category C acts as a prior over x. Because this model is unable to explicitly represent posterior dependencies between sensory and decision variables, both x and C being positive and both x and C being negative act as attractors of its temporal dynamics. This yields qualitatively the same behavior as the sampling model: a stronger influence of early evidence and a transition from primacy to flat weights as category information decreases. As in the sampling model, recency effects emerge only when approximate bias correction is added (Figure 4f-h, Supplemental Figure S5j-r). Whereas the limited number of samples was the key deviation from optimality in the sampling model, here it is the assumption that the brain represents its beliefs separately about x and C in a factorized form (Methods).

Optimal bias correction

A leak term implements optimal inference in a changing environment (Glaze et al., 2015), but each trial of our task is stationary. One might therefore expect that a leak term, or $\gamma > 0$, would impair the model's performance in our task. On the other hand, we motivated the leak term by suggesting that it could approximately correct for the confirmation bias. Under this second interpretation, one might instead expect performance to *improve* for some $\gamma > 0$, especially for conditions where the confirmation bias was strong.

We investigated the relationship between the leak (γ) and model performance. First, we simulated the importance sampling model with $\gamma = 0.1$ and $\gamma = 0.5$ and compared its performance

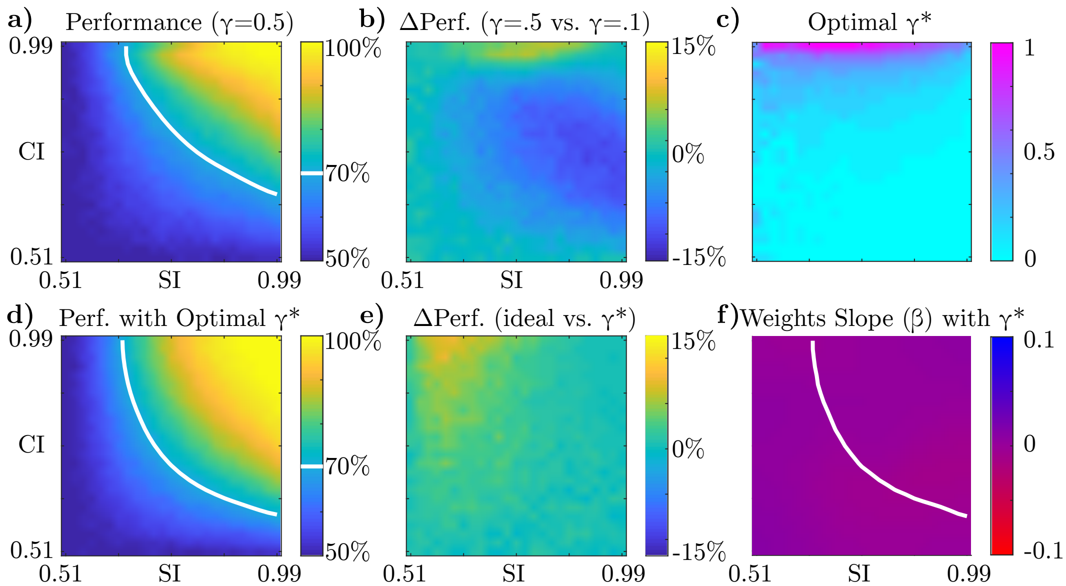

Figure 5: Optimizing performance with respect to γ (see also Supplemental Figure S6). **a)** Model performance across task space with $\gamma = 0.5$ (compare with Figure 4c in which $\gamma = 0.1$). **b)** Difference in performance for $\gamma = 0.5$ versus $\gamma = 0.1$. Higher γ improves performance in the upper part of the space where the confirmation bias is strongest. **c)** Optimizing for performance, the optimal γ^* depends on the task. Where the confirmation bias had been strongest, optimal performance is achieved with a stronger leak term. **d)** Model performance when the optimal γ^* from (c) is used in each task. **e)** Comparing the ideal observer to (d), the ideal observer still outperforms the model but only in the upper part of the space. **f)** Temporal weight slopes when using the optimal γ^* are flat everywhere. The models reproduce the change in slopes seen in the data only when γ is fixed across tasks (compare Figure S5).

across the space of category and sensory information (Figure 5a-b). We found that in the LSHC regime where the confirmation bias had been strongest, the larger value of γ counteracts the bias and leads to better performance, but in the HSLC regime where there had been no confirmation bias, the optimal γ is zero (Figure 5c). We thus see that the optimal value of γ depends on the task statistics, i.e. the balance of sensory information and category information: the stronger the primacy effect or confirmation bias measured above, the higher γ must be to correct for it (Figure 5d). Analogous results were found for the variational model (Supplemental Figure S6).

We next asked what the effect would be on the model's temporal weights if it could utilize the best γ for each task. We found that the $\gamma-$optimized model displayed near-flat weights across the entire space of tasks (Figure 5e). Our data therefore imply that either the brain does not optimize its leak to the statistics of the current task, or that it does so on a timescale that is slower than a single experimental session (roughly 1hr, Methods).

Predictions for Neurophysiology

Both the sampling and variational models induce a confirmation bias by creating an "attractor" dynamic between different levels of the cortical hierarchy – the decision-making area and the relevant sensory areas. Our model therefore makes a number of novel and testable neurophysiological predictions.

First, our model predicts that both "choice probabilities" (Britten et al., 1996; Cumming and Nienborg, 2016) and "differential correlations" (Moreno-Bote et al., 2014) in populations of task-relevant sensory neurons will be stronger in contexts where category information is high and sensory information is low, i.e. when subjects exhibit primacy effects (Wimmer et al., 2015; Haefner et al., 2016). This is because the feedback from the decision-making to sensory areas in our model explicitly biases the sensory representation *in the direction that encodes the stimulus strength*, which is the $f'-$direction (Tajima et al., 2016; Lange and Haefner, 2020). Our model is thus consistent with recent evidence that noise correlations contain a task-dependent component in the f' direction (Bondy et al., 2018).

Second, our model predicts that apparent attractor-dynamics measured in both sensory and decision-making areas are in fact driven by inter- rather than within-area dynamics, and will depend on the decision-making context. In particular, categorization tasks should induce a stronger confirmation bias, and hence stronger attractor-like dynamics, than equivalent estimation tasks, as was recently reported (Tajima et al., 2017). This observation, as well as our above prediction, contrasts with classic attractor models which posit a recurrent feedback loop *within* a decision making area (Wang, 2008; Wimmer et al., 2015).

Discussion

Our work makes three main contributions. First, we show that online inference in a hierarchical model can result in characteristic task-dependent temporal biases, and further that such biases naturally arise in two specific families of biologically-plausible approximate inference algorithms. Second, explicitly modeling the mediating sensory representation allows us to partition the information in the stimulus about the category into two parts – "sensory information" and "category information" – defining a novel two-dimensional space of possible tasks. Third, we collect new data confirming a critical prediction of our theory, namely that individual subjects' temporal biases change depending on the nature of the information in the stimulus. These results strongly suggest that the discrepancy in temporal biases reported by previous studies is resolved by considering how their tasks trade off sensory and category information.

The "confirmation bias" emerges in our models as the result of four key assumptions. Our first assumption is that inference in evidence integration tasks is hierarchical, and that the brain approximates the posterior distribution over both the category, C, and intermediate sensory variables, x. This is in line with converging evidence that populations of sensory neurons encode posterior distributions of corresponding sensory variables (Lee and Mumford, 2003; Yuille and Kersten, 2006; Berkes et al., 2011; Beck et al., 2013) incorporating dynamic prior beliefs via feedback connections (Lee and Mumford, 2003; Yuille and Kersten, 2006; Beck et al., 2013; Nienborg and Roelfsema, 2015; Tajima et al., 2016, 2017; Orbán et al., 2016; Haefner et al., 2016; Lange and Haefner, 2020), which contrasts with other probabilistic theories in which only the likelihood is represented in sensory areas (Ma et al., 2006; Beck et al., 2008; Orhan and Ma, 2017; Walker et al., 2019).

Our second key assumption is that evidence is accumulated online. In our models, the belief over C is updated based only on the posterior from the previous step and the current posterior over x. This can be thought of as an assumption that the brain does not have a mechanism to store and retrieve earlier frames veridically, but must make use of currently available summary statistics. This is consistent with drift-diffusion models of decision-making (Gold and Shadlen, 2007). As mentioned in the main text, the assumptions until now – hierarchical inference with online updates – do not entail any temporal biases for an ideal observer.

Third, we implemented hierarchical online inference making specific assumptions about the limited representational power of sensory areas. In the sampling model, we assumed that the brain can draw a limited number of independent samples of x per update to C. Interestingly, we found that in the small sample regime, the models is inherently unable to account for the prior bias of C on x in its updates to C. Existing neural models of sampling typically assume that samples are distributed temporally (Hoyer and Hyvärinen, 2003; Fiser et al., 2010), but it has also been proposed that the brain could run multiple sampling "chains" distributed spatially (Savin and Denève, 2014). The relevant quantity for our model is the total *effective* number of independent samples that can be generated, stored, and evaluated in a batch to compute each update. The more samples, the smaller the bias predicted by this model.

We similarly limited the representational capacity of the variational model by enforcing that the posterior over x is unimodal, and that there is no explicit representation of dependencies between x and C. Importantly, this does not imply that x and C do not influence each other. Rather, the Variational Bayes algorithm expresses these dependencies in the *dynamics* between the two areas: each update that makes $C = +1$ more likely pushes the distribution over x further towards $+1$, and vice versa. Because the number of dependencies between variables grows exponentially, such approximates are necessary in variational inference with many variables (Fiser et al., 2010). The Mean Field Variational Bayes algorithm algorithm that we use here has been previously proposed as a candidate algorithm for neural inference (Raju and Pitkow, 2016).

The assumptions up to now predict a primacy effect but cannot account for the observed recency effects. When we incorporate a leak term in our models, they reproduce the observed range of biases from primacy to recency. The existence of such a leak term is supported by previous literature (Usher and McClelland, 2001; Bogacz et al., 2006). Further, it is normative in our framework in the sense that reducing the bias in the above models improves performance (Figure 5). The optimal amount of bias correction depends on the task statistics: in the LSHC regime where the confirmation bias is strongest, a higher γ is needed to correct for it. While it is conceivable that the brain would optimize this leak term to the task (Brunton et al., 2013; Piet et al., 2018), our data suggest the leak term is stable across our LSHC and HSLC conditions, or adapted slowly.

It has been proposed that post-decision feedback biases subsequent perceptual estimations (Stocker and Simoncelli, 2007; Talluri et al., 2018). While in spirit similar to our confirmation bias model, there are two conceptual differences between these models and our own: First, the feedback from decision area to sensory area in our model is both continuous and online, rather than conditioned on a single choice after a decision is made. Second, our models are derived from an ideal observer and only incur bias due to approximations, while previously proposed "self-consistency" biases are not normative and require separate justification.

Alternative models have been previously proposed to explain primacy and recency effects in evidence accumulation. Kiani et al. (2008) suggested that an integration-to-bound process is more likely to ignore later evidence even when task-relevant stimuli are of a fixed duration (Kiani et al., 2008). Deneve (2012) showed that simultaneous inference about stimulus strength and choice and in tasks with trials of variable difficulty can lead to either a primacy or a recency effect (Deneve, 2012). However, both models of evidence integration are based entirely on total information per frame (i.e. $p(C|e_f)$) and hence cannot explain the difference between the data for the LSHC and the HSLC conditions since both conditions are matched in terms of total information. In general, *any* model based only on $p(C|e_f)$ cannot explain the pattern in our data. While such a model can coexist with the confirmation bias dynamic proposed by our model, it is not sufficient to explain the pattern in our data for which the trade-off between sensory- and category-information is crucial.

It has also been proposed that primacy effects could be the result of near-perfect integration of an adapting sensory population (Wimmer et al., 2015; Yates et al., 2017). For this mechanism to explain our full results, however, the sensory population would need to become *less* adapted over frames in our HSLC condition, while at the same time *more* adapted in the LSHC condition. We are unaware of such an adaptation mechanism in the literature. Further, although the circuit dynamics of sensory populations could in principle explain our behavioral results, this would not predict top-down neural effects such as the task-dependence of the dynamics of sensory populations (Tajima et al., 2017) nor the origin and prevalence of differential correlations (Bondy et al., 2018), both of which are consistent with our model, as described above.

Models of "leaky" evidence accumulation are known to result in recency effects (Usher and McClelland, 2001; Kiani et al., 2008; Brunton et al., 2013; Glaze et al., 2015). Interestingly, leaky evidence accumulation has also been shown to be optimal in non-stationary environments (Glaze et al., 2015) and could thus in principle indicate that subjects assume such non-stationarity in our HSLC condition. However, this explanation alone cannot explain the presence of primacy effects in the LSHC condition. In sum, while there are numerous existing models that can explain either primacy or recency effects with dedicated mechanisms, ours is the first model to predict the full range of biases and how they may be controlled by the stimulus statistics. Further, because our approximate inference models compute log posterior odds, previously proposed mechanisms like integration to bound are complementary and could be incorporated into our framework.

While our focus is on the perceptual domain in which subjects integrate evidence over a timescale on the order of tens or hundreds of milliseconds, analogous principles hold in the cognitive domain over longer timescales. The crucial computational motif underlying our model of the confirmation bias is hierarchical inference over multiple timescales. An agent in such a setting must simultaneously make accurate judgments of current data (based on the current posterior) and track long-term trends (based on all likelihoods). For instance, Zylberberg et al. (2018) identified an analogous challenge when subjects must simultaneously make categorical decisions each trial (their "fast" timescale) while tracking the stationary statistics of a block of trials (their "slow" timescale), analogous to our LSHC condition. As the authors describe, if subjects base model updates on posteriors rather than likelihoods, they will further entrench existing beliefs (Zylberberg et al., 2018). However, the authors did not investigate order effects; our confirmation bias would predict that subjects' estimates of block statistics is biased towards earlier trials in the block (primacy). Schustek et al. (2018) likewise asked subjects to track information across trials in a cognitive task more analogous to our HSLC condition, and report close to flat weighting of evidence across trials Schustek and

Moreno-bote (2018).

The strength of the perceptual confirmation bias is directly related to the integration of internal "top-down" beliefs and external "bottom-up" evidence previously implicated in clinical dysfunctions of perception (Jardri and Denéve, 2013). Therefore, the differential effect of sensory and category information may be useful in diagnosing clinical conditions that have been hypothesized to be related to abnormal integration of sensory information with internal expectations (Fletcher and Frith, 2009).

Hierarchical (approximate) inference on multiple timescales is a common motif across perception, cognition, and machine learning. We suspect that all of these areas will benefit from the insights on the causes of the confirmation bias mechanism that we have described here and how they depend on the statistics of the inputs in a task.

Methods

Visual Discrimination Task

We recruited students at the University of Rochester as subjects in our study. All were compensated for their time, and methods were approved by the Research Subjects Review Board. We found no difference between naive subjects and authors, so all main-text analyses are combined, with data points belonging to authors and naive subjects indicated in Figure 3d.

Our stimulus consisted of ten frames of band-pass filtered noise (Beaudot and Mullen, 2006; Nienborg and Cumming, 2014) masked by a soft-edged annulus, leaving a "hole" in the center for a small cross on which subjects fixated. The stimulus subtended 2.6 degrees of visual angle around fixation. Stimuli were presented using Matlab and Psychtoolbox on a 1920x1080px 120 Hz monitor with gamma-corrected luminance (Brainard, 1997). Subjects kept a constant viewing distance of 36 inches using a chin-rest. Each trial began with a 200ms "start" cue consisting of a black ring around the location of the upcoming stimulus. Each frame lasted 83.3ms (12 frames per second). The last frame was followed by a single double-contrast noise mask with no orientation energy. Subjects then had a maximum of 1s to respond, or the trial was discarded (Supplemental Figure S1). The stimulus was designed to minimize the effects of small fixational eye movements: (i) small eye movements do not provide more information about either orientation, and (ii) each 83ms frame was too fast for subjects to make multiple fixations on a single frame.

The stimulus was constructed from white noise that was then masked by a kernel in the Fourier domain to include energy at a range of orientations and spatial frequencies but random phases (Beaudot and Mullen, 2006; Nienborg and Cumming, 2014; Bondy et al., 2018) (a complete description and parameters can be found in the Supplemental Text). We manipulated sensory information by broadening or narrowing the distribution of orientations present in each frame, centered on either $+45°$ or $-45°$ depending on the chosen orientation of each frame. We manipulated category information by changing the proportion of frames that matched the orientation chosen for that trial. The range of spatial frequencies was kept constant for all subjects and in all conditions.

Trials were presented in blocks of 100, with typically 8 blocks per session (about 1 hour). Each session consisted of blocks of only HSLC or only LSHC trials (Figure 2). Subjects completed between 1500 and 4400 trials in the LSHC condition, and between 1500 and 3200 trials in the HSLC condition. After each block, subjects were given an optional break and the staircase was reset to $\kappa = 0.8$ and $p_{\text{match}} = 0.9$. p_{match} is defined as the probability that a single frame matched the category for a given trial. In each condition, psychometric curves were fit to the concatenation of all trials from all sessions using the Psignifit Matlab package (Schütt et al., 2016), and temporal weights were fit to all trials below each subject's threshold.

Low Sensory-, High Category-Information (LSHC) Condition

In the LSHC condition, a continuous 2-to-1 staircase on κ was used to keep subjects near threshold (κ was incremented after each incorrect response, and decremented after two correct responses in a row). p_{match} was fixed to 0.9. On average, subjects had a threshold (defined as 70% correct) of $\kappa = 0.17 \pm 0.07$ (1 standard deviation). Regression of temporal weights was done on all sub-threshold trials, defined per-subject.

High Sensory-, Low Category-Information (HSLC) Condition

In the HSLC condition, the staircase acted on p_{match} while keeping κ fixed at 0.8. Although p_{match} is a continuous parameter, subjects always saw 10 discrete frames, hence the true ratio of frames ranged from 5:5 to 10:0 on any given trial. Subjects were on average $69.5\% \pm 4.7\%$ (1 standard deviation) correct when the ratio of frame types was 6:4, after adjusting for individual biases in the 5:5 case. Regression of temporal weights was done on all 6:4 and 5:5 ratio trials for all subjects.

Logistic Regression of Temporal Weights

We constructed a matrix of per-frame signal strengths $\mathbf{S}$ on sub-threshold trials by measuring the empirical signal level in each frame. This was done by taking the dot product of the Fourier-domain energy of each frame as it was displayed on the screen (that is, including the annulus mask applied in pixel space) with a difference of Fourier-domain kernels at $+45°$ and $-45°$. This gives a scalar value per frame that is positive when the stimulus contained more $+45°$ energy and negative when it contained more $-45°$ energy. Signals were z-scored before performing logistic regression, and weights were normalized to have a mean of 1 after fitting.

Temporal weights were first fit using (regularized) logistic regression with different types of regularization. The first regularization method consisted of an AR0 (ridge) prior, and an AR2 (curvature penalty) prior. We did not use an AR1 prior to avoid any bias in the slopes, which is central to our analysis.

To visualize regularized weights in Figure 3, the ridge and AR2 hyperparameters were chosen using 10-fold cross-validation for each subject, then averaging the optimal hyperparameters across subjects for each task condition. This cross validation procedure was used only for display purposes for individual subjects in Figure 3a-c of the main text, while the linear and exponential fits (described below) were used for statistical comparisons. Supplemental Figure S4 shows individual subjects' weights with no regularization.

We used two methods to quantify the shape (or slope) of $\mathbf{w}$: by constraining $\mathbf{w}$ to be either an exponential or linear function of time, but otherwise optimizing the same maximum-likelihood objective as logistic regression. Cross-validation suggests that both of these methods perform similarly to either unregularized or the regularized logistic regression defined above, with insignificant differences (Supplemental Figure S3). The exponential is defined as

$$\mathbf{w}_f^{\text{exponential}} = \alpha \exp(\beta f) \tag{6}$$

where f refers to the frame number. β gives an estimate of the shape of the weights $\mathbf{w}$ over time, while α controls the overall magnitude. $\beta > 0$ corresponds to recency and $\beta < 0$ to primacy. The β parameter is reported for human subjects in Figure 3d, and for the models in Figure 4e,h.

The second method to quantify slope was to constrain the weights to be a linear function in time:

$$\mathbf{w}_f^{\text{linear}} = a + slope \times f \tag{7}$$

where *slope* > 0 corresponds to recency and *slope* < 0 to primacy.

Figure 3d shows the median exponential shape parameter (β) after bootstrapped resampling of trials 500 times for each subject. Both the exponential and linear weights give comparable results (Supplemental Figure S2).

To compute the combined temporal weights across all subjects (in Figure 3a-c), we first estimated the mean and variance of the weights for each subject by bootstrap-resampling of the data 500 times without regularization. The combined weights were computed as a weighted average across subjects at each frame, weighted by the inverse variance estimated by bootstrapping.

Because we are not explicitly interested in the magnitude of $\mathbf{w}$ but rather its *shape* over stimulus frames, we always plot a "normalized" weight, $\mathbf{w}/\mathrm{mean}(\mathbf{w})$, both for our experimental results (Figure 3a-c) and for the model (Figure 4d,g).

Approximate inference models

We model evidence integration as Bayesian inference in a three-variable generative model (Figure 4a) that distills the key features of online evidence integration in a hierarchical model (Haefner et al., 2016). The variables in the model are mapped onto the sensory periphery (e), sensory cortex (x), and a decision-making area (C) in the brain.

In the generative direction, on each trial, the binary value of the correct choice $C \in \{-1, +1\}$ is drawn from a 50/50 prior. x_f is then drawn from a mixture of two Gaussians:

$$x_f^{(gen)} \sim \begin{cases} \mathcal{N}(+C, \sigma_x^2) \text{ with prob. equal to category info.} \\ \mathcal{N}(-C, \sigma_x^2) \text{ otherwise} \end{cases} \tag{8}$$

Finally, each e_f is drawn from a Gaussian around x_f:

$$e_f^{(gen)} \sim \mathcal{N}(x_f, \sigma_e^2) \tag{9}$$

When we model inference in this model, we assume that the subject has learned the correct model parameters, even as parameters change between the two different conditions. This is why we ran our subjects in blocks of only LSHC or HSLC trials on a given day.

Category information in this model can be quantified by the probability that $x_f^{(gen)}$ is drawn from the mode that matches C. We quantify sensory information as the probability with which an ideal observer can recover the sign of x_f. That is, in our model sensory information is equivalent to the area under the ROC curve for two univariate Gaussian distributions separated by a distance of 2, which is given by

$$\text{sensory info.} = \Phi(\sqrt{2}/\sigma_e) \tag{10}$$

where Φ is the inverse cumulative normal distribution.

Because the effective time per update in the brain is likely faster than our 83ms stimulus frames, we included an additional parameter n_{U} for the number of online belief updates per stimulus frame. In the sampling model described below, we amortize the per-frame updates over n_{U} steps, updating n_{U} times per frame using $\frac{1}{n_{\mathrm{U}}}\hat{\mathrm{LLO}}_f$. In the variational model, we interpret n_{U} as the number of coordinate ascent steps.

Simulations of both models were done with 10000 trials per task type and 10 frames per trial. To quantify the evidence-weighting of each model, we used the same logistic regression procedure that was used to analyze human subjects' behavior. In particular, temporal weights in the model are best described by the exponential weights (equation (6)), so we use β to characterize the model's biases.

Sampling model

The sampling model estimates $\mathrm{p}(e_f|C)$ using importance sampling of x, where each sample is drawn from a pseudo-posterior using the current running estimate of $p_{f-1}(C) \equiv p(C|e_1,..,e_{f-1})$ as a marginal prior:

$$x_f^{(s)} \sim Q(x) \propto \mathrm{p}(e_f|x_f)\sum_c \mathrm{p}(x_f|C=c)\mathrm{p}_{f-1}(C=c) \tag{11}$$

Using this distribution, we obtain the following unnormalized importance weights.

$$\hat{w}^{(s)} = \left(\sum_c \mathrm{p}(x_f^{(s)}|C=c)\mathrm{p}_{f-1}(C=c)\right)^{-1} \tag{12}$$

In the self-normalized importance sampling algorithm these weights are then normalized as follows,

$$\hat{w}^{(s)} = \frac{w^{(s)}}{\sum_i w^{(i)}},$$

though we found that this had no qualitative effect on the model's ability to reproduce the trends in the data. The above equations yield the following estimate for the log-likelihood ratio needed for the belief update rule in equation (5):

$$\hat{\mathrm{LLO}}_f = \log \frac{\sum_{s=1}^{S} \mathrm{p}(x_f^{(s)}|C=+1)w^{(s)}}{\sum_{s=1}^{S} \mathrm{p}(x_f^{(s)}|C=-1)w^{(s)}} \tag{13}$$

In the case of infinitely many samples, these importance weights exactly counteract the bias introduced by sampling from the posterior rather than likelihood, thereby avoiding any double-counting of the prior, and hence, any confirmation bias. However, in the case of finite samples, S, biased evidence integration is unavoidable.

The full sampling model is given in Supplemental Algorithm S1. Simulations in the main text were done with $S = 5$, $n_{\mathrm{U}} = 5$, normalized importance weights, and $\gamma = 0$ or $\gamma = 0.1$.

Variational model

The core assumption of the variational model is that while a decision area approximates the posterior over C and a sensory area approximates the posterior over x, no brain area explicitly represents posterior dependencies between them. That is, we assume the brain employs a *mean field approximation* to the joint posterior by factorizing $\mathrm{p}(C, x_1, \ldots, x_F|e_1, \ldots, e_F)$ into a product of approximate marginal distributions $\mathrm{q}(C)\prod_{f=1}^{F}\mathrm{q}(x_f)$ and minimizes the Kullback-Leibler divergence between q and p using a process that can be modeled by the Mean-Field Variational Bayes algorithm (Murphy, 2012).

By restricting the updates to be online (one frame at a time, in order), this model can be seen as an instance of "Streaming Variational Bayes" (Broderick et al., 2013). That is, the model computes a sequence of approximate posteriors over C using the same update rule for each frame. We thus only need to derive the update rules for a single frame and a given prior over C; this is extended to multiple frames by re-using the posterior from frame $f-1$ as the prior on frame f.

As in the sampling model, this model is unable to completely discount the added prior over x. Intuitively, since the mean-field assumption removes explicit correlations between x and C, the model is forced to commit to a marginal posterior in favor of $C = +1$ or $C = -1$ and $x > 0$ or $x < 0$ after each update, which then biases subsequent judgments of each.

To keep conditional distributions in the exponential family (which is only a matter of mathematical convenience and has no effect on the ideal observer), we introduce an auxiliary variable $z_f \in \{-1, +1\}$ that selects which of the two modes x_f is in:

$$z_f = \begin{cases} +1 & \text{with probability equal to category info} \\ -1 & \text{otherwise} \end{cases} \tag{14}$$

such that

$$x_f \sim \mathcal{N}(z_f C, \sigma_x^2). \tag{15}$$

We then optimize $\mathrm{q}(C) \prod_{f=1}^{F} \mathrm{q}(x_f)\mathrm{q}(z_f)$.

Mean-Field Variational Bayes is a coordinate ascent algorithm on the parameters of each approximate marginal distribution. To derive the update equations for each step, we begin with the following (Murphy, 2012):

$$\begin{aligned} \log \mathrm{q}(x_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \\ \log \mathrm{q}(z_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(x_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \\ \log \mathrm{q}(C) &\leftarrow \mathbf{E}_{\mathrm{q}(x_f)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \end{aligned} \tag{16}$$

After simplifying, the new $\mathrm{q}(x_f)$ term is a Gaussian with mean given by equation (17) and constant variance

$$\mu_{x_f} \leftarrow \frac{\sigma_e^2 \mu_C \mu_{z_f} + \sigma_x^2 e_f}{\sigma_e^2 + \sigma_x^2} \tag{17}$$

where μ_C and μ_z are the means of the current estimates of $\mathrm{q}(C)$ and $\mathrm{q}(z)$.

For the update to $\mathrm{q}(z_f)$ in terms of log odds of z_f we obtain:

$$\mathrm{LPO}_{z_f} \leftarrow \log \frac{\mathrm{p}(z_f = +1)}{\mathrm{p}(z_f = -1)} + 2 \frac{\mu_{x_f} \mu_C}{\sigma_e^2 + \sigma_x^2}. \tag{18}$$

Similarly, the update to $\mathrm{q}(C)$ is given by:

$$\mathrm{LPO}_C \leftarrow \log \frac{\mathrm{p}(C = +1)}{\mathrm{p}(C = -1)} + 2 \frac{\mu_{x_f} \mu_{z_f}}{\sigma_x^2} \tag{19}$$

Note that the first term in equation (19) – the log prior – will be replaced with the log posterior estimate from the previous frame (see Supplemental Algorithm S2). Comparing equations (19) and (3), we see that in the variational model, the log likelihood odds estimate is given by

$$\mathrm{L\hat{L}O}_f = 2 \frac{\mu_{x_f} \mu_{z_f}}{\sigma_x^2} \tag{20}$$

Analogously to the sampling model we assume a number of updates n_{U} reflecting the speed of relevant computations in the brain relative to how quickly stimulus frames are presented. Unlike for the sampling model, naively amortizing the updates implied by equation (20) n_{U} times results in a stronger primacy effect than observed in the data, since the Variational Bayes algorithm naturally has attractor dynamics built in. Allowing for an additional parameter η scaling this update (corresponding to the step size in Stochastic Variational Inference (Hoffman et al., 2013)) seems biologically plausible because it simply corresponds to a coupling strength in the feed-forward direction. Decreasing η both reduces the primacy effect and improves the model's performance. Here we used $\eta = 0.05$ in all simulations based on a qualitative match with the data. The full variational model is given in Algorithm S2.

Acknowledgements

This work was supported by NEI/NIH awards R01 EY028811-01 (RMH) and T32 EY007125 (RDL, JLY), as well as an NSF/NRT graduate training grant NSF-1449828 (RDL).

Author Contributions

Author contributions are shown in the following table, where black = significant contribution, gray = partial contribution, and white = zero or minimal contribution.

	RL	AC	JB	JY	RH
Experiment Design					
Experiment Code					
Data Collection					
Data Analysis					
Sampling Model					
Variational Model					
Writing					

References

William H A Beaudot and Kathy T. Mullen. Orientation discrimination in human vision: Psychophysics and modeling. *Vision Research*, 46:26–46, 2006.

Jeff Beck, Katherine Heller, and Alexandre Pouget. Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models. *Advances in Neural Infromation Processing Systems*, 25:3068–3076, 2013.

Jeffrey M. Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152, 2008.

Pietro Berkes, Gergo Orbán, Máté Lengyel, and Jósef Fiser. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, 331(January):83–87, 2011.

C.M. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer (New York), 2006.

Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700–765, 2006.

Adrian G. Bondy, Ralf M. Haefner, and Bruce G. Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature Neuroscience*, 21(4):598–606, 2018.

D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

K H Britten, W T Newsome, M N Shadlen, S Celebrini, and J A Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci*, 13(1):87–100, 1996.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. *Advances in Neural Information Processing Systems*, 26:1727–1735, 2013.

Bingni W Brunton, Matthew M. Botvinick, and Carlos D Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–8, 2013.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. *arXiv*, pages 1–6, 2017.

Bruce G. Cumming and Hendrikje Nienborg. Feedforward and feedback sources of choice probability in neural population responses. *Current Opinion in Neurobiology*, 37:126–132, 2016.

Sophie Deneve. Making Decisions with Unknown Sensory Reliability. *Frontiers in Neuroscience*, 6 (June):1–13, 2012.

Jan Drugowitsch, Valentin Wyart, Anne-Dominique Devauchelle, and Etienne Koechlin. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6):1398–1411, 2016.

József Jósef Fiser, Pietro Berkes, Gergo Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–30, 2010.

Paul C. Fletcher and Chris D. Frith. Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10:48–58, 2009.

Samuel J Gershman and Jeffrey M. Beck. Complex Probabilistic Inference: From Cognition to Neural Computation. In Ahmed Moustafa, editor, *Computational Models of Brain and Behavior*, chapter Complex Pr, pages 1–17. Wiley-Blackwell, 2016.

Christopher M. Glaze, Joseph W. Kable, and Joshua I. Gold. Normative evidence accumulation in unpredictable environments. *eLife*, 4:1–27, 2015.

Joshua I Gold and Michael N. Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30(30):535–574, 2007.

Ralf M. Haefner, Pietro Berkes, and Jozsef Fiser. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron*, 90(3):649–660, 2016.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

P. O. Hoyer and A. Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. *Advances in Neural Information Processing Systems*, 17(1):293–300, 2003.

Renaud Jardri and Sophie Denéve. Circular inferences in schizophrenia. *Brain*, 136(11):3227–3241, 2013.

Roozbeh Kiani, Timothy D Hanks, and Michael N. Shadlen. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of Neuroscience*, 28(12):3017–3029, 2008.

Richard D Lange and Ralf M Haefner. Task-induced neural covariability as a signature of Bayesian learning and inference. *bioRxiv*, 2020.

Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Wei Ji Ma, Jeffrey M. Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

Rubén Moreno-Bote, Jeffrey M. Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, 2014.

David Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 251: 241–251, 1992.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

W. T. Newsome and E. B. Pare. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *The Journal of Neuroscience*, 8(6):2201–2211, 1988.

Rs Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Hendrikje Nienborg and Bruce G. Cumming. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, 459(7243):89–92, 2009.

Hendrikje Nienborg and Bruce G Cumming. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *The Journal of Neuroscience*, 34(10): 3579–85, 2014.

Hendrikje Nienborg and Pieter R. Roelfsema. Belief states as a framework to explain extra-retinal influences in visual cortex. *Current opinion in neurobiology*, 32:45–52, 2015.

Bruno a Olshausen and D J Field. Sparse coding with an incomplete basis set: a strategy employed by \protect{V1}, 1997.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, 92(2):530–543, 2016.

A. Emin Orhan and Wei Ji Ma. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, 8(138), 2017.

Art B. Owen. Importance Sampling. In *Monte Carlo theory, methods and examples*, chapter 9. 2013.

Alex T Piet, Ahmed El Hady, and Carlos D. Brody. Rats adopt the optimal timescale for evidence integration in a dynamic environment. *Nature Communications*, 9, 2018.

Alexandre Pouget, Jeffrey M. Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–8, 2013.

Rajkumar Vasudeva Raju and Xaq Pitkow. Inference by Reparameterization in Neural Population Codes. *Advances in Neural Information Processing Systems*, 30, 2016.

David Raposo, Matthew T Kaufman, and Anne K Churchland. A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, 17(12):1784–1792, 2014.

Cristina Savin and Sophie Denève. Spatio-temporal representations of uncertainty in spiking neural networks. *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

Philipp Schustek and Rubén Moreno-bote. Human confidence judgments reflect reliability-based hierarchical integration of contextual information. *bioRxiv*, 2018.

Heiko H. Schütt, Stefan Harmeling, Jakob H. Macke, and Felix A. Wichmann. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, 122:105–123, 2016.

L Shi and Tl Griffiths. Neural implementation of hierarchical Bayesian inference by importance sampling. *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

Alan A Stocker and Eero P Simoncelli. A Bayesian Model of Conditioned Perception. *Advances in Neural Infromation Processing Systems*, 2007:1409–1416, 2007.

Chihiro I. Tajima, Satohiro Tajima, Kowa Koida, Hidehiko Komatsu, Kazuyuki Aihara, and Hideyuki Suzuki. Population code dynamics in categorical perception. *Nature Scientific Reports*, 5(August 2015):1–13, 2016.

Satohiro Tajima, Kowa Koida, Chihiro I. Tajima, Hideyuki Suzuki, Kazuyuki Aihara, and Hidehiko Komatsu. Task-dependent recurrent dynamics in visual cortex. *eLife*, 6:1–27, 2017.

Bharath Chandra Talluri, Anne E Urai, Konstantinos Tsetsos, Marius Usher, Tobias H Donner, Bharath Chandra Talluri, Anne E Urai, Konstantinos Tsetsos, Marius Usher, and Tobias H Donner. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence Report Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, pages 1–8, 2018.

Marius Usher and James L. McClelland. The Time Course of Perceptual Choice: The Leaky, Competing Accumulator Model. *Psychological Review*, 108(2):550–592, 2001.

A. Wald and J. Wolfowitz. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948.

Edgar Y Walker, R. James Cotton, Wei Ji Ma, and Andreas S Tolias. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23:122–129, 2019.

Xiao Jing Wang. Decision Making in Recurrent Neuronal Circuits. *Neuron*, 60(2):215–234, 2008.

Klaus Wimmer, Albert Compte, Alex Roxin, Diogo Peixoto, Alfonso Renart, and Jaime De Rocha. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications*, 6(6177):1–13, 2015.

Valentin Wyart, Vincent De Gardelle, Jacqueline Scholl, and Christopher Summerfield. Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron*, 76(4):847–858, 2012.

Jacob L. Yates, Il Memming Park, Leor N. Katz, Jonathan W. Pillow, and Alexander C. Huk. Functional dissection of signal and noise in MT and LIP during decision-making. *Nature neuroscience*, 20(9):1285–1292, 2017.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Ariel Zylberberg, Daniel M Wolpert, and Michael N Shadlen. Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, 99:1–15, 2018.

Supplemental Information: A confirmation bias in perceptual decision-making due to hierarchical approximate inference

Richard D. Lange[1,2,*], Ankani Chattoraj[1],
Jeffrey M. Beck[3], Jacob L. Yates[1], Ralf M. Haefner[1,*]

[1]Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

[2]Computer Science, University of Rochester, Rochester, NY 14627, USA.

[3]Department of Neurobiology, Duke University, Durham, NC 27708, USA.

[*]Corresponding authors: rlange@ur.rochester.edu, rhaefne2@ur.rochester.edu.

January 20, 2020

Sensory Information and Category Information in Previous Literature

In this section we justify our categorization of previous studies' stimuli into the low-sensory/high-category information (LSHC) or high-sensory/low-category information (HSLC) regime in relation to Figure 1 and Table S1. While category information and sensory information are well defined in our model, in the brain they will depend on the nature of the intermediate variable x relative to e and C, and those relationships depend on the sensory system under consideration. For instance, a high spatial frequency grating may contain high sensory information to a primate, but low sensory information to a species with lower acuity. Similarly, when "frames" are presented quickly, they may be temporally integrated with the effect of both reducing sensory information and increasing category information. Therefore, the placement of each study in the sensory vs category information space is our best estimate, and we generally only distinguish between high and low along each dimension. Note that for the orientation discrimination task that we designed, we report the *within*-subject *change* in weights from one task condition to the other, which overcomes the difficulties described above: while we cannot estimate the absolute values of sensory and category information due to our limited knowledge about the nature of the human sensory system's representation even in our task, our two-staircase task design acting on the two kinds of information separately guarantees that there will be a change in both sensory information and category information between the LSHC and HSLC conditions while performance is kept constant.

Studies finding a primacy effect

Kiani et al. (2008) studied the classic motion direction discrimination task in which a monkey views a dynamic random dot motion stimulus with a certain percentage of "coherent" dots moving together and the rest moving randomly (Kiani et al., 2008; Newsome and Pare, 1988). Monkeys were trained to categorize the direction of motion as predominantly leftward or rightward. Since the direction of the coherently moving dots (the signal) does not change over time within a trial, this stimulus contains high category information. Since the motion direction is difficult to perceive for any motion frame, it contains low sensory information (Kiani et al., 2008).

Nienborg et al. (2009) developed a task in which subjects viewed a disc with varying binocular disparity. The disc moved back and forth relative to a reference plane (the surrounding ring), changing every 10ms, at a rate too high for the macaques' (and humans') binocular system to resolve, resulting in a percept of a jittering cloud of dots which was located slightly in front of or behind the surrounding ring and blurred in depth (Nienborg – private communication). After 200 frames presented over 2 seconds, subjects judged whether the center disc was in front or behind the reference plane. Since the location of the perceived dot cloud is relatively stable, but itself uncertain with respect to the reference, this stimulus contains high category and low sensory information (Nienborg and Cumming, 2009).

Studies finding a recency effect or flat weighting

In two similar studies by Wyart et al. (2012) and by Drugowitsch et al. (2016), human participants viewed a sequence of eight clearly visible oriented gratings presented for at least 250ms each. Participants reported whether, on average, the tilt of the eight elements fell closer to the cardinal or diagonal axes. These tasks contain high sensory information since for a subject there is little uncertainty about the orientation of any one grating. However they contain low category information since the orientation of any one grating provides only little information about the correct choice (Wyart et al., 2012; Drugowitsch et al., 2016).

Brunton et al. (2013) studied both a visual task and an auditory task where subjects were trained to indicate whether they saw/heard more flashes/clicks on the left or right side of the midline. These task stimuli contain high sensory information since each flash/click is high contrast/loud – well above subjects' detection thresholds. However, they contain low category information since each flash/click contains only little information about the correct choice (Brunton et al., 2013).

Stimulus details

The stimulus was constructed from white noise that was then masked by a kernel in the Fourier domain to include energy at a range of orientations and spatial frequencies but random phases (Beaudot and Mullen, 2006; Nienborg and Cumming, 2014; Bondy et al., 2018). The Fourier-domain kernel consisted of a product of two probability density functions (PDFs): a von Mises PDF over orientation, and a Rician PDF over spatial frequency. This is best expressed using polar coordinates in the Fourier domain:

$$K_{\rho\theta} = \text{vonMises}(\theta; \mu_\theta, \kappa)\text{Rician}(\rho; \mu_\rho, \sigma_\rho)$$

where θ is the angular coordinate and ρ is the spatial frequency coordinate. After transforming back from the Fourier domain to an image, we applied a soft circular aperture with a hole cut out in the center for the fixation cross. The full pixel-space mask is defined by the equation

$$M = \underbrace{\exp(-4\hat{\rho}^2)}_{\text{Gaussian aperture}} \times \underbrace{(1 + \text{erf}(10 \times (\hat{\rho} - \tau_{\text{ap}}/w_{\text{im}})))}_{\text{Center cutout for fixation cross}}$$

where $\hat{\rho}$ is the normalized Euclidean distance to the center of the image ($\hat{\rho} = 0$ at the center, and $\hat{\rho} = \sqrt{2}$ at the corners), and erf is the Error Function. τ_{ap} controlled the width of the central cutout, and w_{im} is the total width of the stimulus. To summarize, each stimulus frame, I, was generated according to

$$\text{I} = M \otimes \mathcal{F}^{-1}\left[\mathcal{F}[\mathcal{W}] \otimes K_{\rho\theta}\right]$$

where $\mathcal{F}$ is the 2D discrete Fourier transform, $\otimes$ is element-wise multiplication of each pixel, and $\mathcal{W}$ is white noise. Images were displayed using Psychtoolbox on a 1920x1080px 120 Hz monitor with gamma-corrected luminance (Brainard, 1997). Using an 8-bit luminance range (0 to 255), each frame was normalized to $127 \pm c$ where c is a contrast parameter. All stimulus parameters are summarized in table S2.

Algorithms

Algorithm S1 Importance Sampling (IS) model for evidence integration

LPO $\leftarrow \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)}$ ▷ initialize log posterior odds to log prior odds
for $f = 1$ to F **do**
  for $n = 1$ to n_{U} **do**
    $p_C \leftarrow (1 + \exp(-LPO))^{-1}$ ▷ current posterior that $C = +1$
    $\hat{p}(x) \leftarrow p_C \mathcal{N}(+1, \sigma_x^2) + (1 - p_C)\mathcal{N}(-1, \sigma_x^2)$ ▷ Mixture of Gaussians prior on x
    $Q(x) \leftarrow \hat{p}(x)p(e_f|x)$
    for $s = 1 \ldots S$ **do**
      $x^{(s)} \sim Q(x)$ ▷ sensory sample from current posterior
      $p_+^{(s)} \leftarrow p(x^{(s)}|C = +1)$ ▷ contribution of each sample to $C = +1$ pool
      $p_-^{(s)} \leftarrow p(x^{(s)}|C = -1)$ ▷ contribution of each sample to $C = -1$ pool
      $w^{(s)} \leftarrow \left(\sum_c \mathrm{p}(x^{(s)}|C = c)\mathrm{p}_{f-1}(C = c)\right)^{-1}$ ▷ (unnormalized) weight of each sample
    end for
    $w \leftarrow w / \sum_{s'} w^{(s')}$ ▷ (optionally) normalize weights
    $p_+^{tot} \leftarrow \sum_s p_+^{(s)} w^{(s)}$ ▷ aggregate evidence for $C = +1$
    $p_-^{tot} \leftarrow \sum_s p_-^{(s)} w^{(s)}$ ▷ aggregate evidence for $C = -1$
    $\hat{\mathrm{LLO}}_f \leftarrow \log p_+^{tot} - \log p_-^{tot}$
    LPO $\leftarrow$ LPO$(1 - \gamma/n_{\mathrm{U}}) + \hat{\mathrm{LLO}}_f/n_{\mathrm{U}}$ ▷ equations (13,5) amortized for n_{U} updates
  end for
end for

Algorithm S2 Variational Bayes (VB) model for evidence integration

LPO $\leftarrow \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)}$ ▷ initialize to log prior odds
for $f = 1$ to F **do**
  $\mu_{z_f} \leftarrow 2\mathrm{p}(z_f = +1) - 1$ ▷ initialize μ_{z_f} to the prior
  for $n = 1$ to n_{U} **do**
    $\mu_C \leftarrow 2(1 + \exp(-\mathrm{LPO}_C))^{-1} - 1$ ▷ convert log-odds to mean of C
    $\mu_{x_f} \leftarrow \frac{\sigma_e^2 \mu_C \mu_{z_f} + \sigma_x^2 e_f}{\sigma_e^2 + \sigma_x^2}$ ▷ equation (17)
    $\mathrm{LPO}_{z_f} \leftarrow \log \frac{\mathrm{p}(z_f=+1)}{\mathrm{p}(z_f=-1)} + 2\frac{\mu_{x_f}\mu_C}{\sigma_x^2+\sigma_e^2}$ ▷ equation (18)
    $\mu_{z_f} \leftarrow 2(1 + \exp(-\mathrm{LPO}_{z_f})^{-1} - 1$ ▷ convert log-odds to mean of z_f
    $\hat{\mathrm{LLO}}_f \leftarrow \frac{2\mu_{x_f}\mu_{z_f}}{\sigma_x^2}$ ▷ Equation (20)
    LPO $\leftarrow$ LPO$(1 - \gamma/n_{\mathrm{U}}) + \eta\hat{\mathrm{LLO}}_f/n_{\mathrm{U}}$ ▷ Equations (5) and (19) amortized for n_{U} updates with update strength η
  end for
end for

Example study	Justification for placement in task space (Figure 1, color-coded)	Suggested stimulus manipulation to change weighting (color-coded)
Brunton et al. (2013), Raposo et al. (2014)	Each click is perceptually clear but only weakly predictive of which side has the higher rate.	Make clicks softer or embed them in noise and increase difference in rates between left and right side.
Wyart et al. (2012), Drugowitsch et al. (2016)	Orientation of each frame is clear but only weakly predictive of which "deck" the orientations were drawn from.	Decrease contrast of each frame or increase pixel noise and reduce variance of orientations within each deck.
Kiani et al. (2008)	Net motion is weak (low coherence) and constant over a trial.	Increase motion coherence but vary net motion direction across stimulus frames within a trial.
Nienborg et al. (2009)	Percept is of a jittering cloud of dots whose depth is close to fixation point.	Increase the distance between cloud and fixation point in depth; vary distance across stimulus frames at a rate resolvable by depth perception

Table S1: Justification of placement of example prior studies in Figure 1c and description of stimulus manipulations that will move it to the opposite side of the category–sensory–information space. Each manipulation corresponds to a prediction about how temporal weighting of evidence should change from primacy (red) to flat/recency (blue), or vice versa, as a result.

Parameter	**Description**	**Values (Units)**
μ_ρ	mean spatial frequency	6.90 (cycles per degree)
σ_ρ	spread of spatial frequency	3.45 (cycles per degree)
κ	(inverse) spread of orientation energy	$0 \leq \kappa \leq 0.8$
c	image contrast	22
τ_{ap}	width of central annulus cutout	25 (pixels) or 0.43 (°)
w_{im}	full image width & height	120 (pixels) or 2.08 (°)

Table S2: Stimulus parameters.

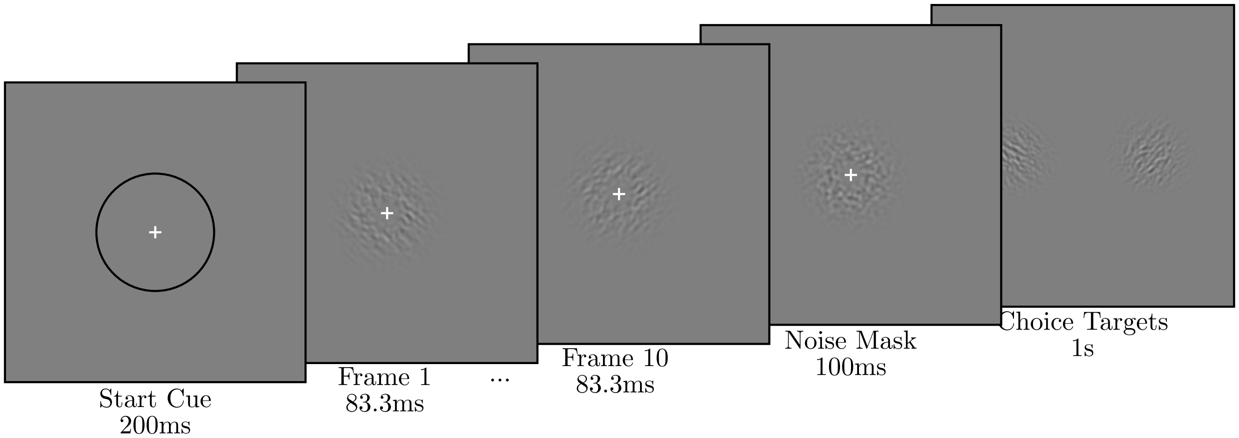

Figure S1: Stimulus timing for each trial in our visual discrimination task

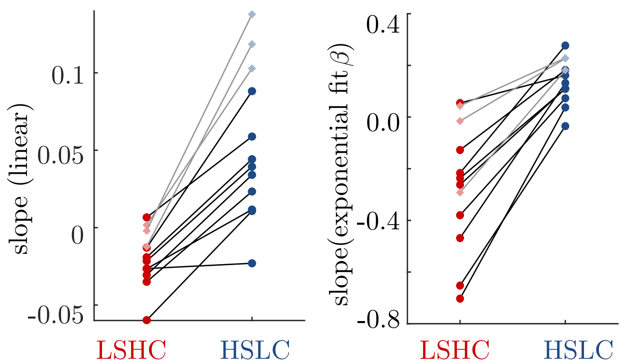

Figure S2: Same as Figure 3d in the main text, comparing slope of $\mathbf{w}$ using a linear fit (left) or an exponential fit (right). Using the linear fit, 11 of 12 subjects individually have a significant increase in slope ($p < 0.05$). Using the exponential fit, 10 of 12 subjects individually have a significant increase in slope ($p < 0.05$).

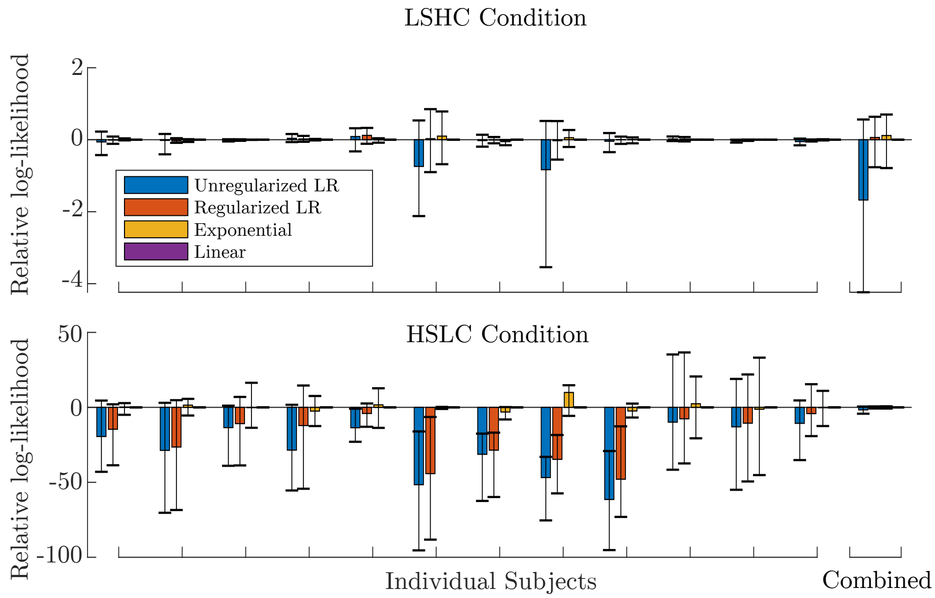

Figure S3: Cross-validation selects linear or exponential shapes for temporal weights, compared to both unregularized and AR2-regularized logistic regression. Panels show 20-fold cross-validation performance of four methods to fit evidence-weighting profiles, separated by task type and by subject. Magnitudes are always relative to the mean log-likelihood of the linear model. Error bars show 50% confidence intervals across folds of shuffled data. "Unregularized LR" refers to standard logistic regression with no regularization. "Regularized LR" refers to the ridge- and AR2-regularized logistic regression objective, where the hyperparameters were chosen to maximize cross-validated fitting performance for each subject. "Exponential" is is the 3-parameter model where weights are an exponential function of time (equation (6) plus a bias term). Similarly, the "Linear" model constrains the weights to be a linear function of time as in equation (7), plus a bias term.

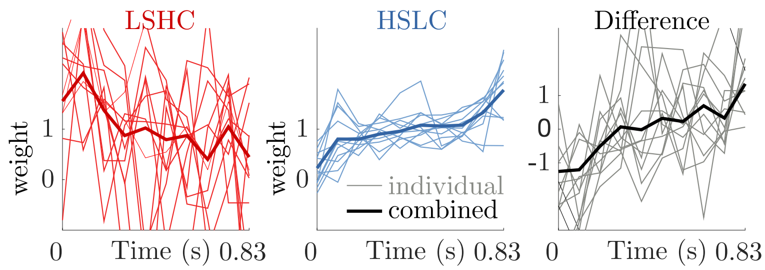

Figure S4: Same as Figure 3a-c in the main text, but with no regularization applied to logistic regression for individual subjects. Both here and in the main text, the “combined” weights are computed using the un-regularized individual weights.

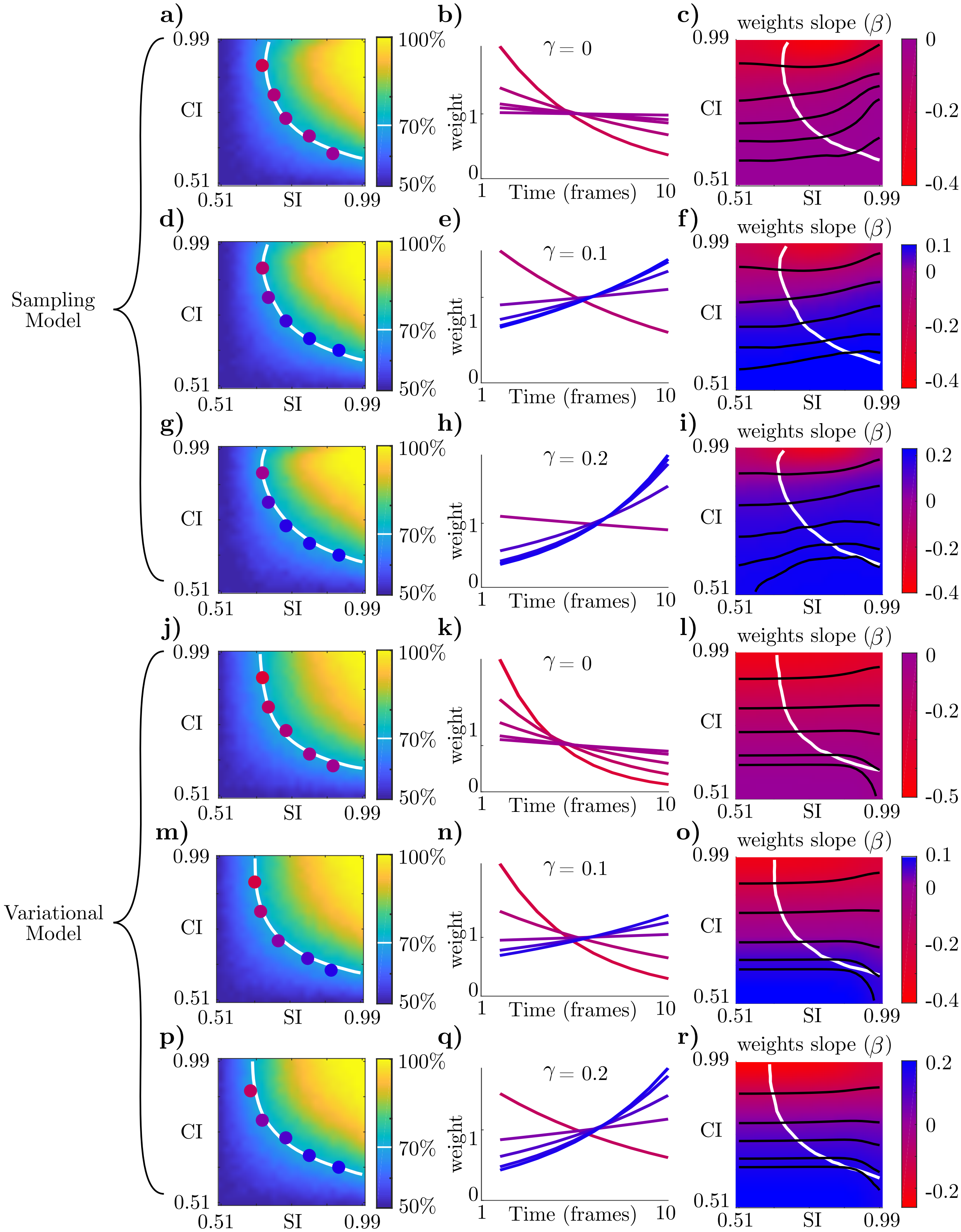

Figure S5: In both models, larger γ increases the prevalence of recency effects across the entire task space. Panels are as in Figure 4 in the main text. **a-c** sampling model with $\gamma = 0$. **d-f** sampling model with $\gamma = 0.1$. **g-i** sampling model with $\gamma = 0.2$. **j-l** variational model with $\gamma = 0$. **m-o** variational model with $\gamma = 0.1$. **p-r** variational model with $\gamma = 0.2$.

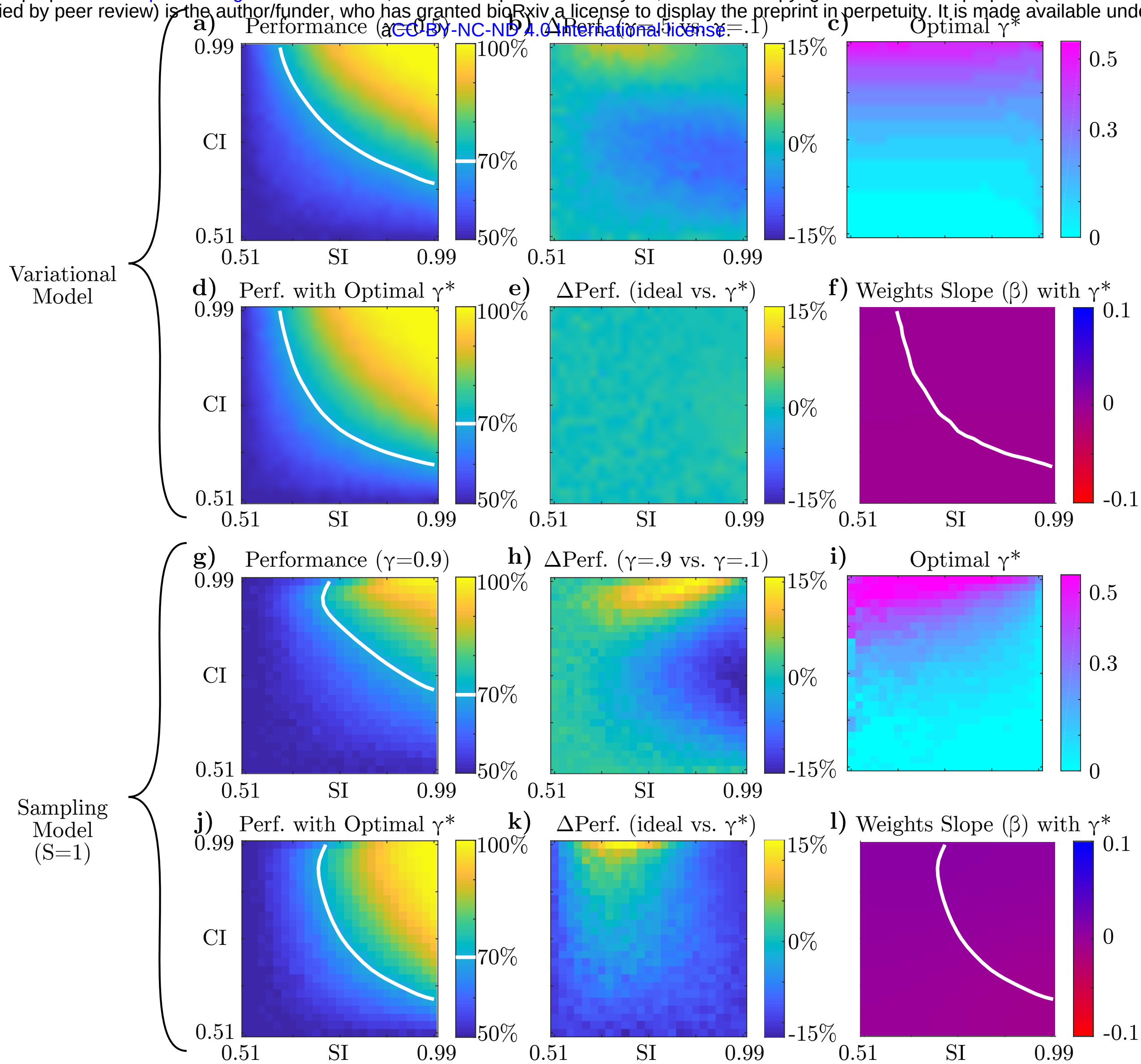

Figure S6: Simulation results for optimal leak (γ) for two further model variations, panels as in Figure 5 in the main text. **a-f** Variational model results. As in the sampling model, we see that the optimal value of $\gamma*$ increases with category information, or with the strength of the confirmation bias. **h-l** Sampling model results with $S = 1$ (in the main text we used $S = 5$). Since the sampling model without a leak term approaches the ideal observer in the limit of $S \to \infty$, the optimal γ^* was close to 0 for much of the space in the main text figure. Here, by comparison, $\gamma^* > 0$ is more common because the $S = 1$ model is more biased.

References

William H A Beaudot and Kathy T. Mullen. Orientation discrimination in human vision: Psychophysics and modeling. *Vision Research*, 46:26–46, 2006.

Adrian G. Bondy, Ralf M. Haefner, and Bruce G. Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature Neuroscience*, 21(4): 598–606, 2018.

D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

Bingni W Brunton, Matthew M. Botvinick, and Carlos D Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–8, 2013.

Jan Drugowitsch, Valentin Wyart, Anne-Dominique Devauchelle, and Etienne Koechlin. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6):1398–1411, 2016.

Roozbeh Kiani, Timothy D Hanks, and Michael N. Shadlen. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of Neuroscience*, 28(12):3017–3029, 2008.

W. T. Newsome and E. B. Pare. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *The Journal of Neuroscience*, 8(6): 2201–2211, 1988.

Hendrikje Nienborg and Bruce G. Cumming. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, 459(7243):89–92, 2009.

Hendrikje Nienborg and Bruce G Cumming. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *The Journal of Neuroscience*, 34 (10):3579–85, 2014.

Valentin Wyart, Vincent De Gardelle, Jacqueline Scholl, and Christopher Summerfield. Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron*, 76(4):847–858, 2012.