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

Richard D. Lange[a,b,1], Ankani Chattoraj[a], Jeffrey M. Beck[c], Jacob L. Yates[a], Ralf M. Haefner[a,2]
[a] Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627
[b] Computer Science, University of Rochester, Rochester, NY 14627
[c] Department of Neurobiology, Duke University, Durham, NC 27708
[1] rlange@ur.rochester.edu [2] rhaefne2@ur.rochester.edu

April 22, 2019

Abstract

Human decisions are known to be systematically biased. A prominent example of such a bias occurs during the temporal integration of sensory evidence. 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 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. 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. We argue that 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. If those signals themselves reflect inferences that combine both prior expectations and sensory evidence, we suggest that this can then lead to an observable confirmation bias [30].

Formalizing this idea in the context of approximate Bayesian inference requires extending classic evidence-integration models to include an explicit intermediate sensory representation (Figure 1b). We explicitly model the inferences of the intermediate sensory representation and find that task difficulty is modulated by two distinct types of information: the information between the stimulus and sensory representation (sensory information), and the information between sensory representation and category (category information) (Figure 1b). The balance between these distinct types of information can indeed explain puzzling discrepancies in the literature with regards to how subjects weigh evidence over time across a wide range of studies. Even in tasks where all evidence is equally informative about the correct category, existing studies typically report one of three distinct motifs: some find that early evidence is weighted more strongly (a primacy effect) [23, 31] some that information is weighted equally over time

(as would be optimal) [47, 9, 37], and some find late evidence being weighted most heavily (a recency effect) [12] (Figure 1a,c). There are myriad differences between these studies such as subject species, sensory modality, stimulus parameters, and computational frameworks [23, 9, 16, 12]. However, none of these aspects can explain their different findings. We show that the differences arise naturally in a hierarchical approximate inference framework.

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 and the categorical identity of the stimulus is a binary variable $C \in \{-1, +1\}$, then evidence in favor of $C = +1$ after F independent frames is $\mathrm{p}(C = +1|e_1, \ldots, e_F) \propto \mathrm{p}(C = +1) \prod_{f=1}^{F} \mathrm{p}(e_f|C = +1)$. The ideal observer reports the most likely category, for instance by reporting the sign of $\log \mathrm{p}(C = +1|e_1, \ldots, e_F) - \log \mathrm{p}(C = -1|e_1, \ldots, e_F)$.

The ideal observer's performance is limited only by (i) the information about C available on each frame, $p(e_f|C)$, and (ii) the number of frames per trial. In the brain, however, a decision-making area computing a belief about the correct choice only has access to the sensory representation of the stimulus, which we call x, not to the outside stimulus e directly. For example, in a visual task each e_f would be the image on the screen while inferences about x_f are represented by the concurrent activity of relevant neurons in visual cortex. This implies that the information between the stimulus and category can be partitioned into the information between the stimulus and the sensory representation, and the information between sensory representation and category, which we call "sensory information" and "category information," respectively (Figure 1b). These two kinds of information span a two-dimensional space with a task being defined by a single point (Figure 1c).

To illustrate this difference, consider the classic dot motion task [29] and the Poisson clicks task [9], which occupy opposite locations in the space spanned by sensory and category information. In the classic low-coherence dot motion task, subjects view a cloud of moving dots, some percentage of which move "coherently" in one direction. Here, sensory information is low since evidence about the net motion direction at any time is weak. Category information, on the other hand, is high, since knowing the "true" motion on a single frame would be highly predictive of the correct choice (and of motion on subsequent frames). In the Poisson clicks task, 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, but category information is low since knowing the side on which a single click was presented provides only little information about the correct choice (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.

If we assume that the sensory representation – which itself is an inference about the actual stimulus – incorporates prior expectations [25, 18, 40], then, as we show below, approximate inference models predict that this will lead to a primacy effect when sensory information is low and category information is high, but not when sensory information is high and category information is low. Indeed, a qualitative placement of prior studies in the space spanned by these two kinds of information demonstrates that studies

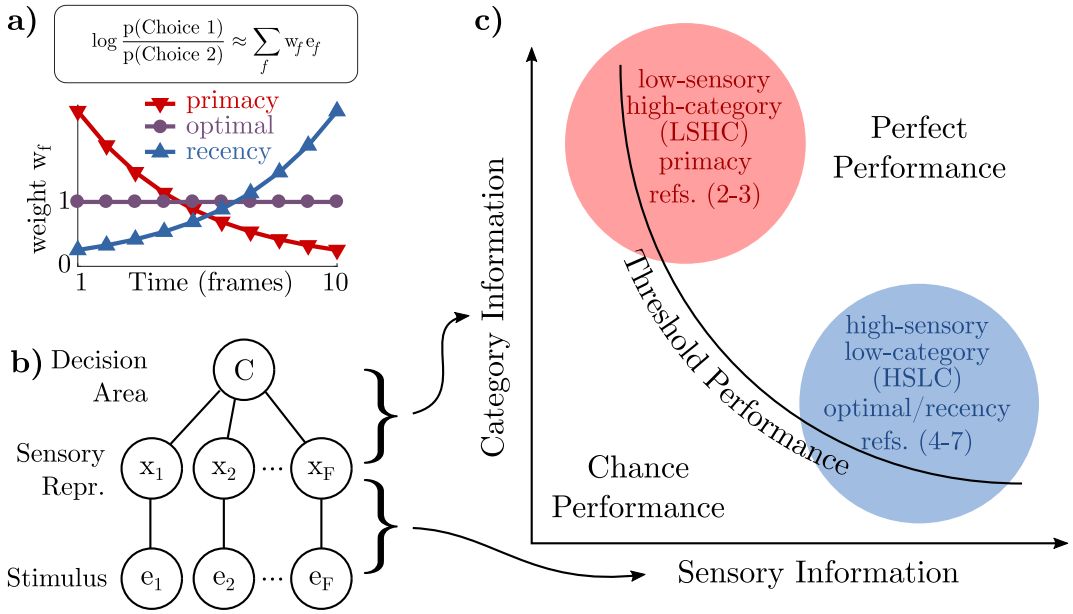

Figure 1: **a)** A subject's "temporal weighting strategy" is an estimate of how their choice is based on a weighted sum of each frame of evidence e_f. Three commonly observed motifs are decreasing weights (primacy), constant weights (optimal), or increasing weights (recency). **b)** Uncertainty in the stimulus about the category may be decomposed into uncertainty in each frame about a sensory variable ("sensory information") and uncertainty about the category given the sensory variable ("category information"). **c)** Category information and sensory information may be altered independently, creating a two-dimensional space of possible tasks, where subjects will be at threshold performance whenever the two sources of information are balanced. 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).

that find early weighting are located in the upper left quadrant (low-sensory/high-category or LSHC) and studies with equal or late weighting in the lower right quadrant (high-sensory/low-category or HSLC) (Figure 1c). This suggests that the different trade-off between sensory information and category information may indeed underlie differences in temporal weighting seen in previous studies. Further, with this framework it is straightforward to predict how simple changes in stimulus statistics of previous studies should change the temporal weighting they find (Supplemental Table S1).

Visual Discrimination Task

To test this critical model prediction, 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 [1] (Figure 2b,d). On each trial, there was a single true orientation category, but individual frames might differ. 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 (Materials and Methods).

If the brain represents 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. For a ratio of $5:5$, a frame's orientation does not predict the correct choice and category information is zero. For a ratio of $10:0$, knowledge of the orientation of a single frame is sufficient to determine the correct choice and category information is high. 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 staircase lowering the sensory information while keeping category information high, or we ran a 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 defining threshold performance as 70% correct (Materials and Methods).

In agreement with our hypothesis, we find predominantly flat or decreasing temporal weights when sensory information is low and category information is high (Figure 3a). When the information is partitioned differently – into high sensory and low categorical information – we find flat or increasing weights (Figure 3b). 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 nearly all subjects (Figure 2c,d) ($p < 0.05$ for 9 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 [20, 13, 18, 34] and a parametric (mean-field) approximation [2, 36] to exact inference can explain

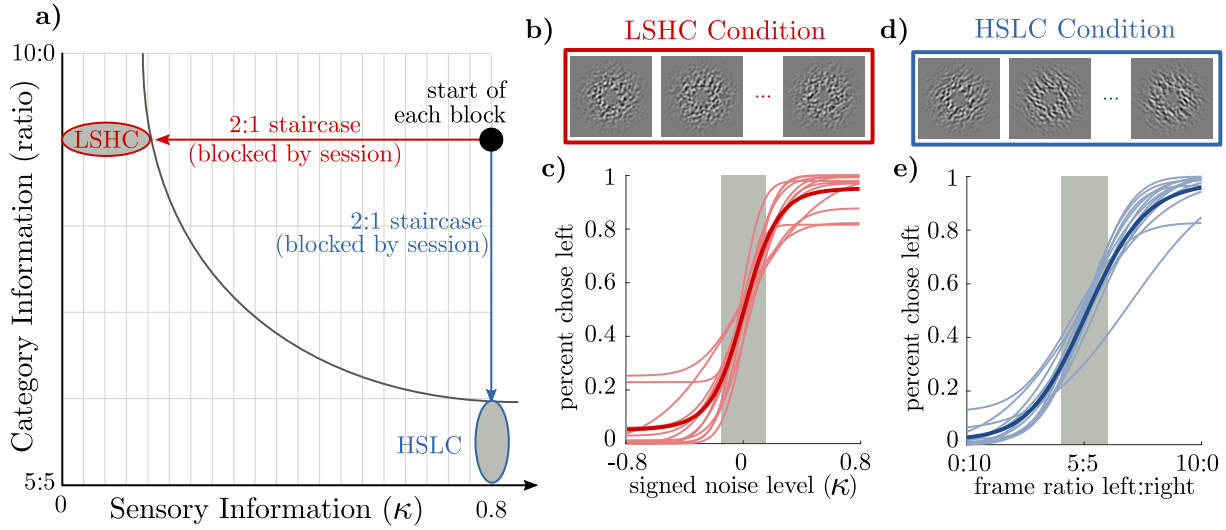

Figure 2: Summary of experiment design. **a)** In our task, category information is determined by the ratio of frame categories, and sensory information is determined by a noise parameter κ. At the start of each block, we reset the staircase to the same point, with category information at $9:1$ and κ at 0.8. We then ran a 2-to-1 staircase either on κ or on category information, always the same on a given day. The LSHC and HSLC ovals indicate sub-threshold trials; only these trials were used in 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 lines) over the κ staircase. Shaded gray area indicates the median threshold level across all subjects. **d)** Visualization of frames in the HSLC condition. Each frame the orientation is clear, but oreintations change from frame to frame. **e)** Psychometric curves over frame ratios, plotted as in (c).

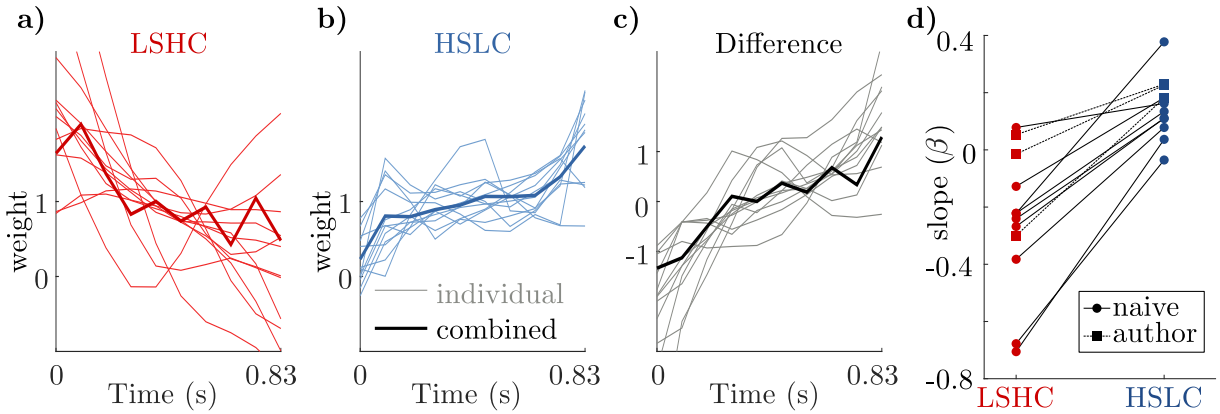

Figure 3: Regression of subjects' temporal weights. **a-b)** Temporal weights for individual subjects (thin lines) and the mean across all subjects (thick lines). Weights are always normalized to have a mean of 1. Individual subjects' curves were fit using a cross-validated smoothness term for visualization purposes only (Materials and 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.

the observed pattern of changing temporal weights as a function of stimulus statistics.

The crucial assumption in both models is that the brain computes a posterior belief over both C and x given the external evidence, i.e. $\mathrm{p}(x, C|e)$, not just over the categorical variable C. This assumption differs from some models of approximate inference in the brain that assume populations of sensory neurons strictly encode the *likelihood* of the stimulus [26], but is consistent with other models from both sampling and parametric families [4, 18, 36, 40].

In our models, the brain's belief about x depends both on the external evidence, e, via the likelihood, but also on the brain's current belief about C, via the prior. For a decision-making area in the brain to update its belief about C based on current sensory responses, it needs to account for, or "subtract out" its influence on those sensory responses. Failure to do so will result in "double-counting" evidence presented early in the trial, inducing a positive feedback loop between the sensory area and the decision making area (Figure 4a). The stronger the decision-making area's belief in a particular choice, the more likely the sensory representation of x will concur with that belief through the influence of the prior. We call this feedback loop a "perceptual confirmation bias."

Importantly, the strength of this confirmation bias depends on the relative amount of sensory and category information in the stimulus (Figure 4a). It is weakest when the posterior over x is dominated by the likelihood, a case that occurs when the category information is much weaker than the sensory information. Conversely, the feedback loop is strongest when the category information is high compared to the sensory information, as assumed in [18] who found a primacy effect in their model.

To demonstrate and quantify the intuitions laid out above, we implemented approximate 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

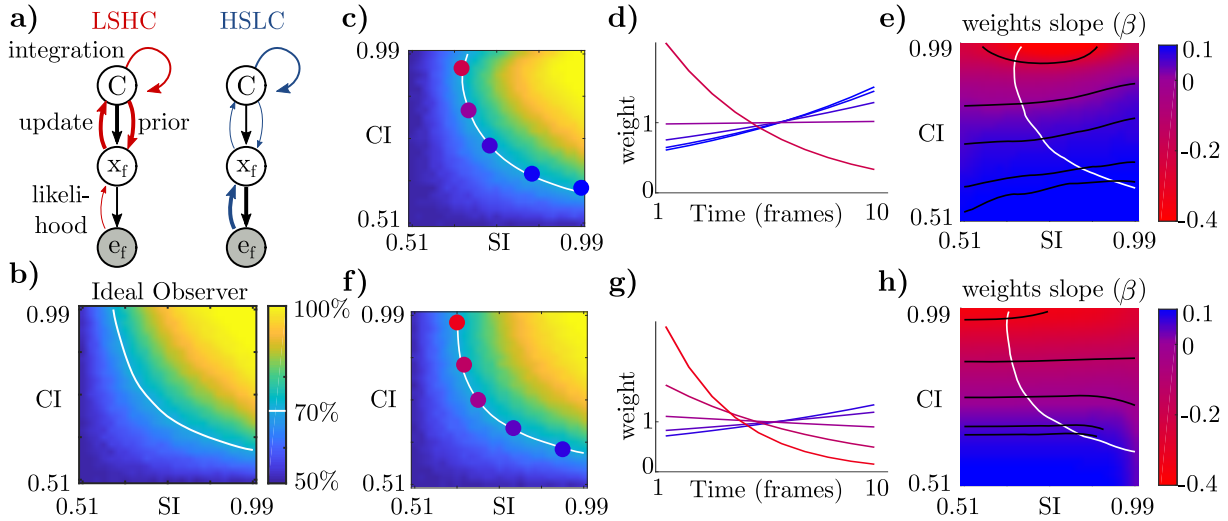

Figure 4: Approximate inference models explain results. **a)** The difference in stimulus statistics between HSLC and LSHC trade-offs implies that the relevant sensory representation is differentially influenced by the stimulus or by beliefs about the category C. A "confirmation bias" or feedback loop 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 $\gamma = 0.1$.

[20, 13, 18, 34] and mean field variational inference [2, 36] (Figure 4).

As in the Sequential Probability Ratio Test [17], both models recursively compute a running estimate of the log posterior odds (LPO) of the latent category C:

$$\begin{aligned}\mathrm{LPO}_f &\equiv \log \frac{\mathrm{p}(C=+1|e_1,\ldots,e_f)}{\mathrm{p}(C=-1|e_1,\ldots,e_f)} \\ &= \log \frac{\mathrm{p}_{f-1}(C=+1)}{\mathrm{p}_{f-1}(C=-1)} + \log \frac{\mathrm{p}(e_f|C=+1)}{\mathrm{p}(e_f|C=-1)} \\ &= \mathrm{LPO}_{f-1} + \log \frac{\mathrm{p}(e_f|C=+1)}{\mathrm{p}(e_f|C=-1)} \\ &= \mathrm{LPO}_{f-1} + \mathrm{LLO}_f \end{aligned} \tag{1}$$

where $\mathrm{LPO}_0 = \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)}$ is the log prior odds and LLO_f is the log likelihood odds for frame f, i.e. the change in belief about C implied by the new evidence e_f.

Equation (1) describes exact, unbiased online evidence-integration. It is "online" because it describes an *update* to a running estimate of LPO each frame. The bias in our models comes from two additional assumptions: first, that sensory areas of the brain represent the *posterior* $p(x_f|e_1,\ldots,e_f)$ over x_f given all evidence so far rather than its *likelihood* $p(e_f|x_f)$, and second that they do so only approximately.

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 x. In our model, the prior belief about C biases the distribution from which samples are generated (Materials and Methods). The canonical way to compute an expectation with respect to one distribution (the likelihood) using samples from another (the posterior) is 'importance sampling' which weights each sample so as to "subtract out" the prior as described above. 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. The bias is such that it under-corrects for the prior that has been fed back, resulting in a positive feedback loop. Figure 4b and Supplemental Figure S5a-c show performance for the ideal observer and for the resulting sampling 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-subject change seen in our data [18]. 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. 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. In particular, we modeled linear bias correction by explicitly subtracting out a fraction of the running posterior odds at each step:

$$\mathrm{LPO}_f \leftarrow \mathrm{LPO}_{f-1}(1-\gamma) + \hat{\mathrm{LLO}}_f \tag{2}$$

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 (2), which can be interpreted as a noiseless, discrete time version of the classic drift-diffusion model [17] with γ as a leak parameter.

Incorporating equation (2) 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 [26, 2] – 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. [36]. 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, $\mathrm{q}(x)\mathrm{q}(C)$ [2, 36]. Inference proceeds by iteratively minimizing the Kullback-Leibler divergence between $\mathrm{q}(x)\mathrm{q}(C)$ and $\mathrm{p}(x, C|e)$ (Materials and Methods). As in the sampling model, the running estimate the category C acts as a prior over x. Because this model is unable to explicitly represent posterior dependencies between sensory and decision variables, it is forced to commit early either to both x and C being positive or to both x and C being negative. This yields the same behavior as the sampling model: a transition from primacy to flat weights as category information decreases, with recency effects emerging 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 and that its instantaneous belief about x is unimodal.

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" [7, 10] and "differential correlations" [27] 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 [46, 18]. 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 [40, 24]. Our model is thus consistent with recent evidence that noise correlations are dominated by a task-dependent component in the f' direction [5].

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 [41]. This observation, as well as our above prediction, contrasts with classic attractor models which posit a recurrent feedback loop *within* a decision making area [45, 46].

Discussion

We have shown that online inference in a hierarchical model can result in characteristic task-dependent temporal biases, and further that such biases are unavoidable in two specific families of biologically-plausible approximate inference algorithms. Explicitly modeling the mediating sensory representation allowed 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. We explicitly probed this space using a standard visual discrimination task and show that individual subjects' temporal biases change as predicted by our model. We argue that the discrepancy in biases reported by previous studies is resolved by considering how their tasks trade off sensory and category information.

The "confirmation bias" emerges in our model as the result of two key assumptions: (1) intermediary sensory areas represent posterior beliefs and as a result incorporate expectations fed back from decision areas, and (2) those expectations cannot be completely accounted for in subsequent updates to the decision-variable.

Our work is in line with converging evidence that populations of sensory neurons encode posterior distributions of corresponding sensory variables [25, 49, 4, 2, 40, 41, 34, 18] incorporating dynamic prior beliefs via feedback connections [25, 49, 2, 33, 40, 41, 18, 24], which contrasts with other probabilistic theories in which only the likelihood is represented

[26, 3, 35, 44].

Optimal inference often requires computing the likelihood of each data point independently. However, in the streaming setting – when data are encountered sequentially – an agent must simultaneously make accurate judgments of current data (based on the current posterior) and make adjustments to a model of long-term trends (based on all likelihoods). Basing model updates on posteriors rather than likelihoods will further entrench existing biases [15, 50].

One recently proposed solution to the general problem of online updating is to make two separate inferences about each data point: one using the current prior for decision-making, and one using a "counterfactual" flat prior for model updates (i.e. using the likelihood) [50]. Since the likelihood can in principle be recovered by dividing the posterior by the prior (or "subtracting out" the log prior), our models can be seen as approximations to recovering the likelihood without explicitly making a second "counterfactual" inference.

In line with our model, it has been proposed that post-decision feedback biases subsequent perceptual estimations [39, 42]. 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 [23]. 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 [11]. 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 [46, 48]. For an analogous mechanism to explain recency effects, however, the sensory population would need to become *less* adapted over time in our HSLC condition. While these circuit dynamics of sensory neurons could in principle explain our behavioral results, this would make different predictions both for the task-dependence of the dynamics of sensory populations [41] and for the origin and prevalence of differential correlations [5], both of which are consistent with our model, as described above.

Models of "leaky" evidence accumulation are known to result in recency effects [43, 23, 9, 16]. Interestingly, leaky evidence accumulation has also been shown to be optimal in non-stationary environments [16] 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. While our data are well-explained by a consistent leak parameter γ across conditions (Figure 4), it remains to be seen whether subjects' apparent evidence-weighting strategies would approach the optimal strategy given further experience in each condition (i.e. γ may be learned slowly).

In the brain, decisions are not based directly on external evidence but on intermediate representations. If those intermediate representations themselves in part reflect prior beliefs, and if inference in the brain is approximate, then this is likely to result in a bias. The nature of this bias is directly related to the integration of internal "top-down" beliefs and external "bottom-up" evidence previously implicated in clinical dysfunctions of perception [21]. Importantly, we have shown how the strength of this effect depends on the nature of the information in the task in a way

that may generalize to cognitive contexts where the confirmation bias is typically studied. For instance, our model makes predictions about when beliefs will be updated in line with the presented evidence, and when it will paradoxically be updated in contradiction to the presented evidence [22]. Finally, 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 [14].

Materials and 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 [1, 32] 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 [6]. 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 [1, 32, 5] (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. Psychometric curves were fit to the concatenation of all trials from all sessions using the Psignifit Matlab package [38], and temporal weights were fit to all trials below each subject's threshold, separately in each condition.

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).

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 dis-

crete 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

$$\mathbf{w}_f^{\text{exponential}} = \alpha \, \exp(\beta f) \tag{3}$$

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:

$$\mathbf{w}_f^{\text{linear}} = a + \mathit{slope} \times f \tag{4}$$

where $\mathit{slope} > 0$ corresponds to recency and $\mathit{slope} < 0$ 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 $\mathbf{w}/\text{mean}(\mathbf{w})$, 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 [18]. 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:

$$x_f^{(gen)} \sim \begin{cases} \mathcal{N}(+C, \sigma_x^2) \text{ with prob. equal to cat. info.} \\ \mathcal{N}(-C, \sigma_x^2) \text{ otherwise} \end{cases} \tag{5}$$

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

$$e_f^{(gen)} \sim \mathcal{N}(x_f, \sigma_e^2) \tag{6}$$

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

$$\text{sensory info.} = \Phi(\sqrt{2}/\sigma_e) \tag{7}$$

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 (3)), 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:

$$x_f^{(s)} \sim Q(x) \propto \mathrm{p}(e_f|x_f) \sum_c \mathrm{p}(x_f|C=c)\mathrm{p}_{f-1}(C=c) \tag{8}$$

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

$$\hat{w}^{(s)} = \left(\sum_c \mathrm{p}(x_f^{(s)}|C=c)\mathrm{p}_{f-1}(C=c) \right)^{-1} \tag{9}$$

This yields the following estimate for the log-likelihood ratio needed for the belief update rule in equation (2):

$$\hat{\mathrm{LLO}}_f = \log \frac{\sum_{s=1}^{S} \mathrm{p}(x_f^{(s)}|C=+1)w^{(s)}}{\sum_{s=1}^{S} \mathrm{p}(x_f^{(s)}|C=-1)w^{(s)}} \tag{10}$$

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 [28].

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" [8]. 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 fundamentally 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:

$$z_f = \begin{cases} +1 & \text{with probability equal to category info} \\ -1 & \text{otherwise} \end{cases} \tag{11}$$

such that

$$x_f \sim \mathcal{N}(z_f C, \sigma_x^2). \tag{12}$$

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 [28]:

$$\begin{aligned} \log \mathrm{q}(x_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \\ \log \mathrm{q}(z_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(x_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \\ \log \mathrm{q}(C) &\leftarrow \mathbf{E}_{\mathrm{q}(x_f)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + const \end{aligned} \tag{13}$$

After simplifying, the new $\mathrm{q}(x_f)$ term is a Gaussian with mean given by equation (14) and constant variance

$$\mu_{x_f} \leftarrow \frac{\sigma_e^2 \mu_C \mu_{z_f} + \sigma_x^2 e_f}{\sigma_e^2 + \sigma_x^2} \tag{14}$$

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:

$$\mathrm{LPO}_{z_f} \leftarrow \log \frac{\mathrm{p}(z_f = +1)}{\mathrm{p}(z_f = -1)} + 2\frac{\mu_{x_f}\mu_C}{\sigma_e^2 + \sigma_x^2}. \tag{15}$$

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

$$\mathrm{LPO}_C \leftarrow \log \frac{\mathrm{p}(C = +1)}{\mathrm{p}(C = -1)} + 2\frac{\mu_{x_f}\mu_{z_f}}{\sigma_x^2} \tag{16}$$

Note that the first term in equation (16) – the log prior – will be replaced with the log posterior estimate from the previous frame (see Supplemental Algorithm S2). Comparing equations (16) and (1), we see that in the variational model, the log likelihood odds estimate is given by

$$\hat{\mathrm{LLO}}_f = 2\frac{\mu_{x_f}\mu_{z_f}}{\sigma_x^2} \tag{17}$$

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 (17) n_{U} times results in a stronger primacy effect than observed in the data. Allowing for an additional parameter η scaling this update (corresponding to the step size in Stochastic Variational Inference [19]) seems biologically plausible because it simply corresponds to a coupling strength in the feedforward 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

RL and RH wrote the manuscript. RL coded the experiment and data analysis. JY and AC helped design the experiment and temporal weighting analysis. AC recruited and ran subjects. RL, AC, and RH created the sampling model. RL and JB created the variational model.

References

[1] Beaudot, W. H. A., and Mullen, K. T. Orientation discrimination in human vision: Psychophysics and modeling. Vision Research 46 (2006), 26–46.

[2] Beck, J., Heller, K., and Pouget, A. Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models. Advances in Neural Infromation Processing Systems 25 (2013), 3068–3076.

[3] Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. Probabilistic Population Codes for Bayesian Decision Making. Neuron 60, 6 (2008), 1142–1152.

[4] Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. Science 331, January (2011), 83–87.

[5] Bondy, A. G., Haefner, R. M., and Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. Nature Neuroscience 21, 4 (2018), 598–606.

[6] Brainard, D. H. The psychophysics toolbox. Spatial Vision 10 (1997), 433–436.

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

[8] Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. Streaming variational bayes. Advances in Neural Information Processing Systems 26 (2013), 1727–1735.

[9] Brunton, B. W., Botvinick, M. M., and Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. Science 340, 6128 (2013), 95–8.

[10] Cumming, B. G., and Nienborg, H. Feedforward and feedback sources of choice probability in neural population responses. Current Opinion in Neurobiology 37 (2016), 126–132.

[11] Deneve, S. Making Decisions with Unknown Sensory Reliability. Frontiers in Neuroscience 6, June (2012), 1–13.

[12] Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. Neuron 92, 6 (2016), 1398–1411.

[13] Fiser, J. J., Berkes, P., Orbán, G., and Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. Trends in Cognitive Sciences 14, 3 (2010), 119–30.

[14] Fletcher, P. C., and Frith, C. D. Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. Nature Reviews Neuroscience 10 (2009), 48–58.

[15] Fryer, R. G., Harms, P., and Jackson, M. O. Updating Beliefs When Evidence Is Open to Interpretation: Implications for Bias and Polarization. 2018.

[16] Glaze, C. M., Kable, J. W., and Gold, J. I. Normative evidence accumulation in unpredictable environments. eLife 4 (2015), 1–27.

[17] Gold, J. I., and Shadlen, M. N. The neural basis of decision making. Annual review of neuroscience 30, 30 (2007), 535–574.

[18] Haefner, R. M., Berkes, P., and Fiser, J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron 90, 3 (2016), 649–660.

[19] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. Journal of Machine Learning Research 14 (2013), 1303–1347.

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

[21] Jardri, R., and Deneve, S. Circular inferences in schizophrenia. Brain 136, 11 (2013), 3227–3241.

[22] Jern, A., Chang, K.-M. K., and Kemp, C. Belief polarization is not always irrational. Psychological review 121, 2 (2014), 206–224.

[23] Kiani, R., Hanks, T. D., and Shadlen, M. N. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. The Journal of Neuroscience 28, 12 (2008), 3017–3029.

[24] Lange, R. D., and Haefner, R. M. Inferring the brain's internal model from sensory responses in a probabilistic inference framework. bioRxiv (2016).

[25] Lee, T. S., and Mumford, D. Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A 20, 7 (2003), 1434–1448.

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

[27] Moreno-Bote, R., Beck, J. M., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. Information-limiting correlations. Nature Neuroscience 17, 10 (2014), 1410–1417.

[28] Murphy, K. P. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

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

[30] Nickerson, R. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology 2, 2 (1998), 175–220.

[31] Nienborg, H., and Cumming, B. G. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459, 7243 (2009), 89–92.

[32] Nienborg, H., and Cumming, B. G. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. The Journal of Neuroscience 34, 10 (2014), 3579–85.

[33] Nienborg, H., and Roelfsema, P. R. Belief states as a framework to explain extra-retinal influences in visual cortex. Current opinion in neurobiology 32 (2015), 45–52.

[34] Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. Neuron 92, 2 (2016), 530–543.

[35] Orhan, A. E., and Ma, W. J. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nature Communications 8, 138 (2017).

[36] Raju, R. V., and Pitkow, X. Inference by Reparameterization in Neural Population Codes. Advances in Neural Information Processing Systems 30 (2016).

[37] Raposo, D., Kaufman, M. T., and Churchland, A. K. A category-free neural population supports evolving demands during decision-making. Nature Neuroscience 17, 12 (2014), 1784–1792.

[38] Schütt, H. H., Harmeling, S., Macke, J. H., and Wichmann, F. A. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. Vision Research 122 (2016), 105–123.

[39] Stocker, A. A., and Simoncelli, E. P. A Bayesian Model of Conditioned Perception. Advances in Neural Infromation Processing Systems 2007 (2007), 1409–1416.

[40] Tajima, C. I., Tajima, S., Koida, K., Komatsu, H., Aihara, K., and Suzuki, H. Population code dynamics in categorical perception. Nature Scientific Reports 5, August 2015 (2016), 1–13.

[41] Tajima, S., Koida, K., Tajima, C. I., Suzuki, H., Aihara, K., and Komatsu, H. Task-dependent recurrent dynamics in visual cortex. eLife 6 (2017), 1–27.

[42] Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., Donner, T. H., Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., and Donner, T. H. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence Report Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. Current Biology (2018), 1–8.

[43] Usher, M., and McClelland, J. L. The Time Course of Perceptual Choice: The Leaky, Competing Accumulator Model. Psychological Review 108, 2 (2001), 550–592.

[44] Walker, E. Y., Cotton, R. J., Ma, W. J., and Tolias, A. S. A neural basis of probabilistic computation in visual cortex. bioRxiv (2018).

[45] Wang, X. J. Decision Making in Recurrent Neuronal Circuits. Neuron 60, 2 (2008), 215–234.

[46] Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A., and Rocha, J. D. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. Nature Communications 6, 6177 (2015), 1–13.

[47] Wyart, V., Gardelle, V. D., Scholl, J., and Summerfield, C. Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. Neuron 76, 4 (2012), 847–858.

[48] Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., and Huk, A. C. Functional dissection of signal and noise in MT and LIP during decision-making. Nature neuroscience 20, 9 (2017), 1285–1292.

[49] Yuille, A., and Kersten, D. Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Sciences 10, 7 (2006), 301–308.

[50] Zylberberg, A., Wolpert, D. M., and Shadlen, M. N. Counterfactual reasoning underlies the learning of priors in decision making. Neuron 99 (2018), 1–15.

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

Richard D. Lange, Ankani Chattoraj, Jeffrey M. Beck, Jacob L. Yates, Ralf M. Haefner

April 22, 2019

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 1. 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 [6, 7]. 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 [6].

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 [8].

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 [10, 5].

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 [4].

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 [1, 9, 2]. 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 [3]. 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 2.

Algorithms

Algorithm S1 Importance Sampling (IS) model for evidence integration

$\mathrm{LPO} \leftarrow \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)}$ ▷ 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
    $w \leftarrow w / \sum_{s'} w^{(s')}$ ▷ (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 $C = -1$
    $\mathrm{L\hat{L}O}_f \leftarrow \log p_+^{tot} - \log p_-^{tot}$
    $\mathrm{LPO} \leftarrow \mathrm{LPO}(1 - \gamma/n_{\mathrm{U}}) + \mathrm{L\hat{L}O}_f/n_{\mathrm{U}}$ ▷ equations (10,2) amortized for n_{U} updates

Algorithm S2 Variational Bayes (VB) model for evidence integration

$\mathrm{LPO} \leftarrow \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)}$ ▷ 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 (14)
    $\mathrm{LPO}_{z_f} \leftarrow \log \frac{\mathrm{p}(z_f=+1)}{\mathrm{p}(z_f=-1)} + 2\frac{\mu_{x_f}\mu_C}{\sigma_x^2 + \sigma_e^2}$ ▷ equation (15)
    $\mu_{z_f} \leftarrow 2(1 + \exp(-\mathrm{LPO}_{z_f})^{-1} - 1$ ▷ convert log-odds to mean of z_f
    $\mathrm{L\hat{L}O}_f \leftarrow \frac{2\mu_{x_f}\mu_{z_f}}{\sigma_x^2}$ ▷ Equation (17)
    $\mathrm{LPO} \leftarrow \mathrm{LPO}(1 - \gamma/n_{\mathrm{U}}) + \eta \mathrm{L\hat{L}O}_f/n_{\mathrm{U}}$ ▷ Equations (2) and (16) amortized for n_{U} updates with update strength η

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 1: 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 2: Stimulus parameters.

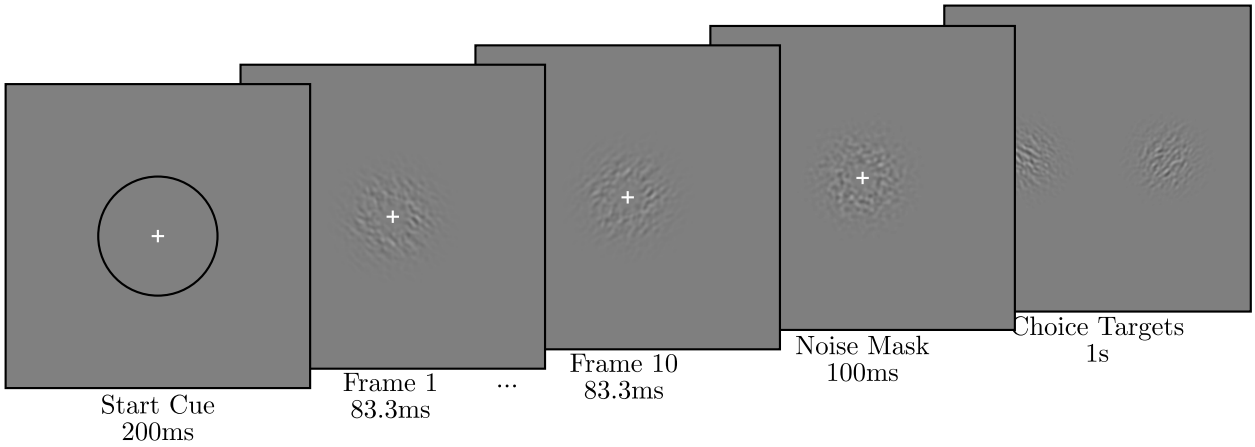

Figure 1: Stimulus timing for each trial in our visual discrimination task

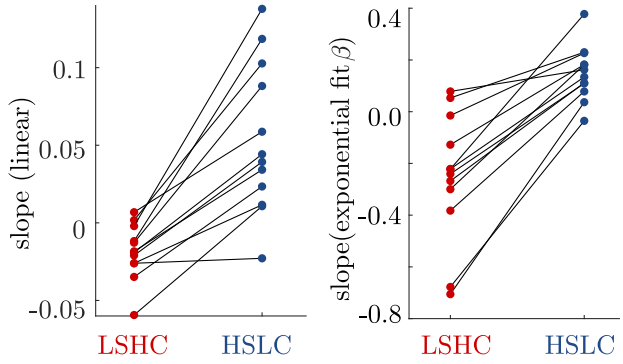

Figure 2: Same as Figure 3d in the main text, comparing slope of $\mathbf{w}$ using a linear fit (left) or an exponential fit (right). Using the linear fit, 11 of 12 subjects individually have a significant increase in slope ($p < 0.05$). Using the exponential fit, 9 of 12 subjects individually have a significant increase in slope ($p < 0.05$).

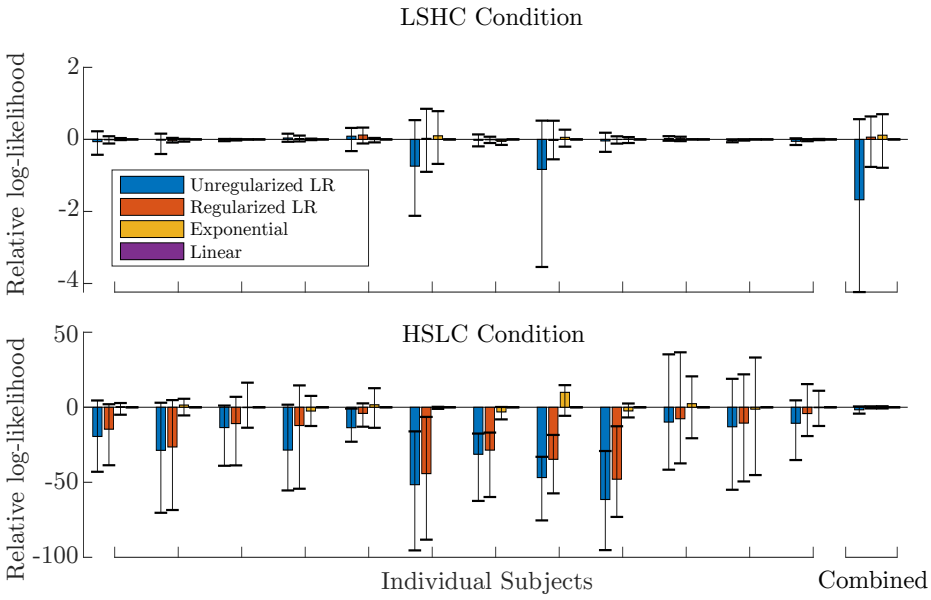

Figure 3: Cross-validation selects linear or exponential shapes for temporal weights, compared to both un-regularized and AR2-regularized logistic regression. Panels show 20-fold cross-validation performance of four methods to fit evidence-weighting profiles, separated by task type and by subject. Magnitudes 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 (3) plus a bias term). Similarly, the "Linear" model constrains the weights to be a linear function of time as in equation (4), plus a bias term.

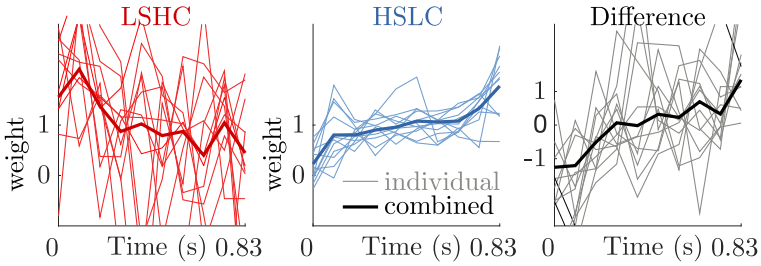

Figure 4: Same as Figure 3a-c in the main text, but with no regularization applied to logistic regression for individual subjects. Both here and in the main text, the "combined" weights are computed using the un-regularized individual weights.

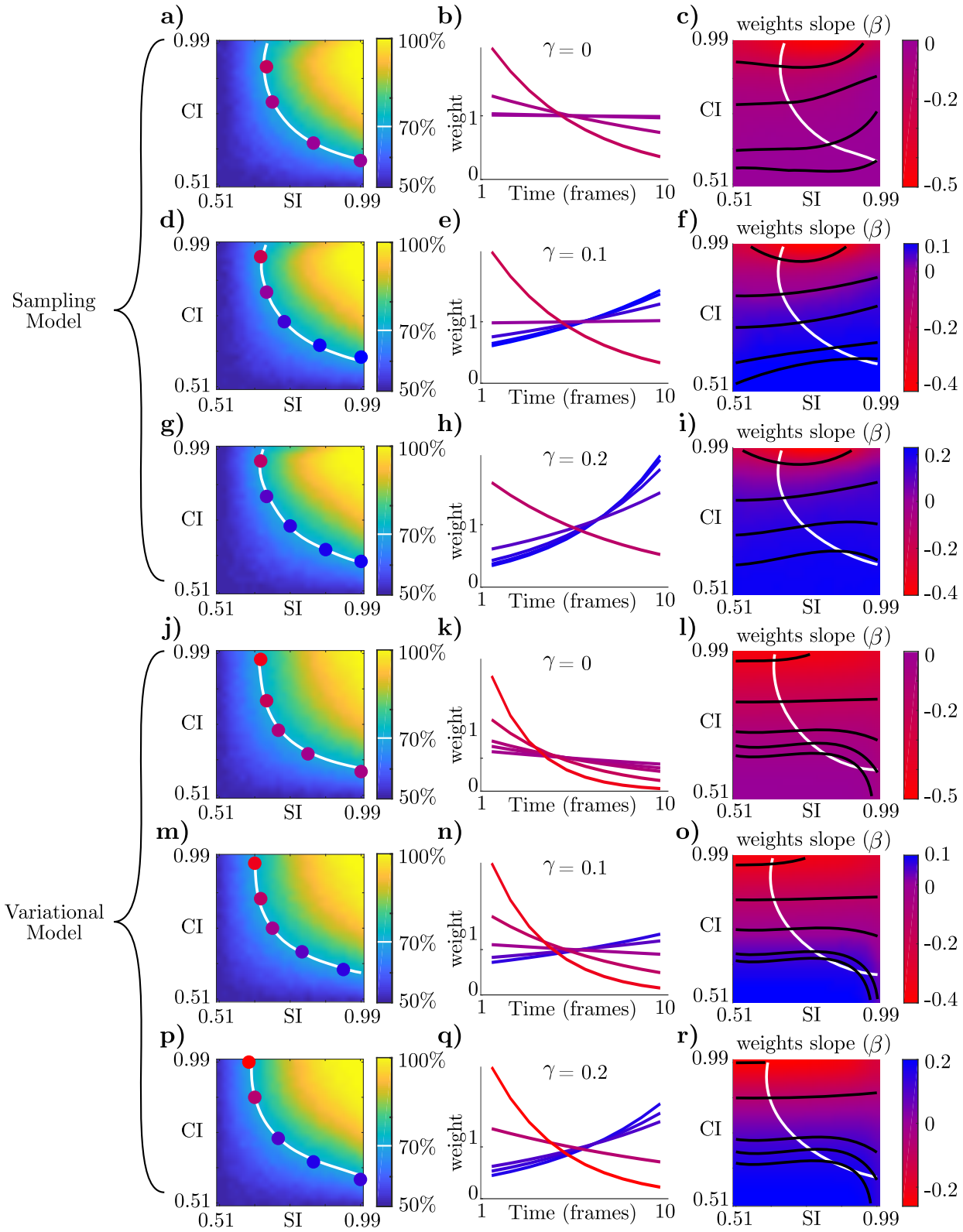

Figure 5: In both models, larger γ increases the prevalence of recency effects across the entire task space. Panels are as in Figure 4 in the main text. **a-c** sampling model with $\gamma = 0$. **d-f** sampling model with $\gamma = 0.1$. **g-i** sampling model with $\gamma = 0.2$. **j-l** variational model with $\gamma = 0$. **m-o** variational model with $\gamma = 0.1$. **p-r** variational model with $\gamma = 0.2$.

References

[1] Beaudot, W. H. A., and Mullen, K. T. Orientation discrimination in human vision: Psychophysics and modeling. *Vision Research 46* (2006), 26–46.

[2] Bondy, A. G., Haefner, R. M., and Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. *Nature Neuroscience 21*, 4 (2018), 598–606.

[3] Brainard, D. H. The psychophysics toolbox. *Spatial Vision 10* (1997), 433–436.

[4] Brunton, B. W., Botvinick, M. M., and Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. *Science 340*, 6128 (2013), 95–8.

[5] Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron 92*, 6 (2016), 1398–1411.

[6] Kiani, R., Hanks, T. D., and Shadlen, M. N. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of Neuroscience 28*, 12 (2008), 3017–3029.

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

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

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

[10] Wyart, V., Gardelle, V. D., Scholl, J., and Summerfield, C. Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron 76*, 4 (2012), 847–858.