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.

November 25, 2020

Abstract

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 confirms a critical prediction of our framework showing that effective temporal integration strategies can be robustly changed within each subject, and that allows us to exclude alternate explanations through quantitative model comparison.

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; Michel and Peters, 2020).

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

Approximate hierarchical inference leads to temporal biases

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. This corresponds to the well-known process of summing the log odds implied by each piece of evidence (Wald and Wolfowitz, 1948; Bogacz et al., 2006):

$$\begin{aligned} \mathrm{p}(C=+1|e_1,\ldots,e_F) &\propto \mathrm{p}(C=+1)\prod_{f=1}^{F}\mathrm{p}(e_f|C=+1) \\ \log\mathrm{p}(C=+1|e_1,\ldots,e_F) &= \log\mathrm{p}(C=+1)+\sum_{f=1}^{F}\log\mathrm{p}(e_f|C=+1) \end{aligned} \tag{1}$$

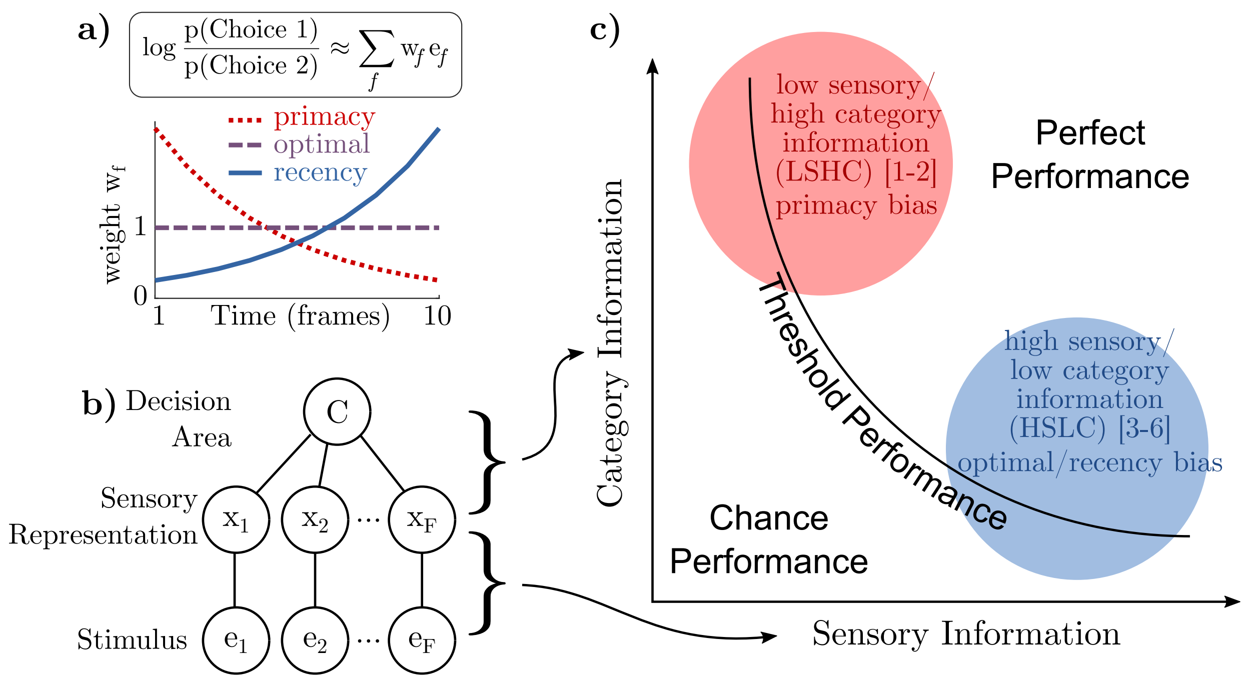

Figure 1: **a)** A subject's "temporal weighting strategy" is an estimate of how their choice is based on a weighted sum of each frame of evidence e_f. Three commonly observed motifs are decreasing weights (primacy), constant weights (optimal), or increasing weights (recency). **b)** Information in the 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.

The ideal observer's performance is thus determined only by (i) the information about C available on each frame, and (ii) the number of frames per trial.

In the brain, however, a decision-making area cannot base its decision on the externally presented stimulus directly, but must rely on intermediate sensory features, which we call x_f. If sensory information is processed in a purely feedforward fashion with independent noise, then a decision-making area can simply integrate the evidence in x_f directly. This is consistent with some theories of inference in the brain in which sensory areas represent a likelihood distribution over stimuli (Ma et al., 2006; Beck et al., 2008; Pouget et al., 2013; Walker et al., 2019). However, activity in sensory areas does not rigidly track the stimulus, but is known to be influenced by past stimuli (Yates et al., 2017; Lueckmann et al., 2018), as well as by feedback from the rest of the brain (Gilbert and Li, 2013; Keller and Mrsic-Flogel, 2018). In fact, the intermediate sensory representation is itself often assumed to be the result of an inference process over latent variables 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) in which feedforward connections communicate the likelihood and feedback communicates the prior or other contextual expectations (Fiser et al., 2010; Pouget et al., 2013; Gershman and Beck, 2016; Tajima et al., 2017; Lange and Haefner, 2020).

Returning to the evidence integration problem in equation (1), accounting for intermediate sensory representations corresponds to marginalizing over the intervening x_f to compute the instantaneous evidence $\mathrm{p}(e_f|C)$ as follows:

$$\begin{aligned}\mathrm{p}(e_f|C) &= \int \mathrm{p}(e_f|x_f)\mathrm{p}(x_f|C)\mathrm{d}x_f \\ &= \int \mathrm{p}(x_f|e_f)\frac{\mathrm{p}(e_f)\mathrm{p}(x_f|C)}{\mathrm{p}(x_f)}\mathrm{d}x_f\,. \end{aligned} \tag{2}$$

The first line is simply the definition of marginalizing over x_f, and the terms in red in the second line are the result of applying Bayes' rule to the red term in the first line. The integral incorporates sensory uncertainty over x_f in the update to C, averaging over all plausible values weighted by $\mathrm{p}(x_f|e_f)$, which is the posterior distribution over sensory features.

Importantly, equation (2) is true for *any* prior over x_f, since whatever prior, $\mathrm{p}(x_f)$, is used to compute the posterior, $\mathrm{p}(x_f|e_f)$, is accounted for by dividing it out in the second term. Incorporating prior information into the sensory representation, therefore, does not introduce any bias, as long as the update to C can exactly account for (or "divide out") that prior. However, if sensory areas only approximately represent the posterior $\mathrm{p}(x_f|e_f)$, then downstream areas may only approximately be able to correct for the prior. Crucially, *approximations* to equation (2) can lead to biases.

We hypothesize that feedback of "decision-related" information to sensory areas (Nienborg et al., 2012; Cumming and Nienborg, 2016) implements a prior that reflects current beliefs about the stimulus category (Haefner et al., 2016; Tajima et al., 2016; Lange and Haefner, 2020). Such a bias is, in fact, optimal in the sense that it incorporates information from earlier frames; in a correlated world, as in our task, the first frame e_1 is informative of later sensory features x_f. Using $\mathrm{p}_{f-1}(C=c) = \mathrm{p}(C=c|e_1,\ldots,e_{f-1})$ to denote the brain's belief that the category is $C=c$ after the first $f-1$ frames, the posterior over x_f given *all* frames, $\mathrm{p}(x_f|e_1,\ldots,e_f)$, can be written as

$$\mathrm{p}(x_f|e_1,\ldots,e_f) \propto \mathrm{p}(e_f|x_f)\underbrace{\sum_c \mathrm{p}_{f-1}(C=c)\mathrm{p}(x_f|C=c)}_{\mathrm{p}_f(x_f)}\,. \tag{3}$$

In other words, sensory areas dynamically combine instantaneous evidence ($\mathrm{p}(e_f|x_f)$) with accumulated categorical beliefs ($\mathrm{p}_{f-1}(C)$) to arrive at a more precise estimate of present sensory features x_f.

As stated above, incorporating prior information into $\mathrm{p}(x_f|e_f)$ does not necessarily lead to a bias, but *approximately* representing the posterior may lead to one. In the case where the prior contains information about earlier stimuli as in equation (3), *under*-correcting for this prior leads to earlier frames entering into the update twice, forming a positive feedback loop between estimates of x_f and the belief in C. This mechanism, which we call a "perceptual confirmation bias," leads to primacy effects. *Over*-correcting for the prior, on the other hand, leads to information from earlier frames decaying away, observable as recency effects.

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, temporal biases arise as a direct consequence of the approximate nature of inference over the intermediate sensory variables in the brain. The strength and direction of the bias (primacy or recency) depends on how how strong the prior influence of C on x_f is – when this prior influence is strong, it is under-corrected, leading to a confirmation bias and primacy effects. When the prior is weak, it is over-corrected, leading to recency effects. Importantly, the strength of the prior influence of C on x_f – and hence the predicted direction of the bias – is easily manipulated experimentally, as we describe next.

"Sensory Information" vs "Category Information"

Accounting for the intervening sensory $\mathbf{x}$ as in Figure 1b 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 the relationship between stimulus and category is ambiguous in each frame (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). When frames are sequential, 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.

Category information governs the strength of the prior fed back from C to x_f. For instance, in a task with high category information such as dot motion, 60% certainty in the stimulus category translates to 60% certainty in the net motion on the next frame. In a low category information task such as the Poisson clicks task, on the other hand, 60% certainty about the side with more clicks is only weakly predictive of where the next click will appear. In equation (3), category information corresponds to the strength of the prior $\mathrm{p}_f(x_f)$, and sensory information to the strength of the likelihood $\mathrm{p}(e_f|x_f)$. If our hypothesis is correct that temporal biases are the result of approximate hierarchical inference, then trading off between sensory information and category information should be sufficient to switch from primacy effects to recency effects, all while subjects' overall performance is kept at threshold.

Indeed, 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 provides initial empirical evidence that approximate hierarchical inference dynamics, along with the trade-off between sensory information and category information, may indeed underlie differences in temporal weighting seen in previous studies. Further, this framework predicts that simple changes in stimulus statistics should change the temporal weighting found in previous studies (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 image 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

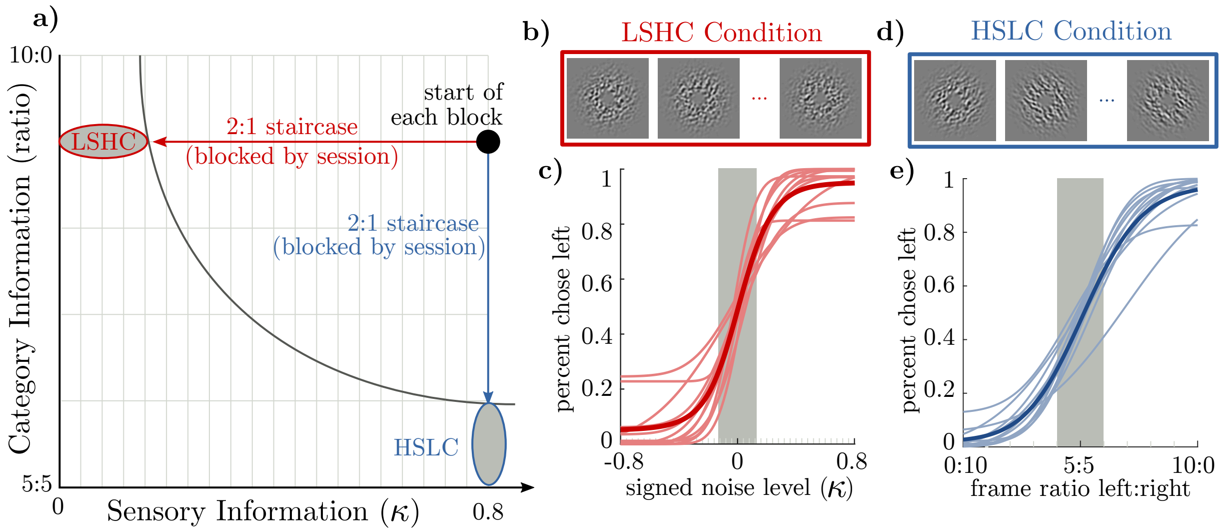

Figure 2: Summary of experiment design. **a)** Category information is determined by the expected ratio of frames in which the orientation matches the correct category, and sensory information is determined by a parameter κ determining the degree of spatial orientation coherence (Methods). At the start 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).

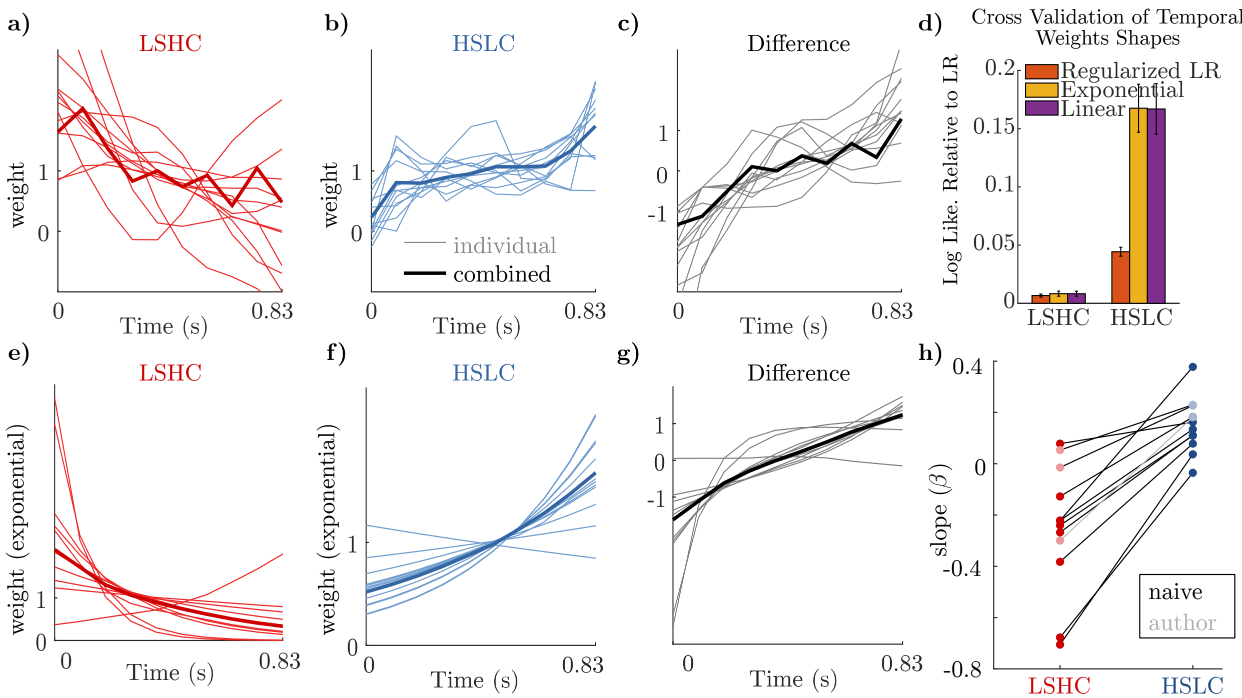

Figure 3: Subjects' temporal weights robustly change with stimulus statistics. **a-b)** Temporal weights from logistic regression for individual subjects (thin lines) and the mean across all subjects (thick lines). Weights are normalized to have a mean of 1 to emphasize shape rather than magnitude. **c)** 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)** Average log-likelihood difference from logistic regression for three regularized weight functions: logistic regression with a smoothness prior, and with weights constrained to be linear or exponential functions of time. Cross-validation indicates that constraining weights to be linear or exponential functions of time is best. **e-g)** Individual and average temporal weights, plotted as in (a-c), now using weights constrained to be exponential functions of time. Weights in (e) and (f) are normalized to have mean 1 for visualization purposes. **h)** *Change* in the exponential slope parameter between the two task contexts for each subject is consistently positive (individually significant in 9 of 12 subjects). Points are median slope values after bootstrap-resampling each subject's sub-threshold trials. A slope parameter $\beta > 0$ corresponds to recency and $\beta < 0$ to primacy (similar results for linear fits, Supplemental Figure S2).

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.

We recruited 15 human subjects, out of which 12 (9 naive and 3 authors) completed the experiment. For each subject, we compared 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 and each subject, we used logistic regression to infer 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,e). However, when the information is partitioned differently – in the HSLC condition – we find flat or increasing weights (Figure 3b,f). In fact, the *difference* in weights between conditions was remarkably consistent across subjects (Figure 3c,g). To quantify this change, we first used cross-validation to select a method for quantifying temporal slopes, and found that constraining weights to be a linear or exponential function of time worked equally well, and both outperformed plain or regularized logistic regression (Figure 3d; Methods). A within-subject comparison revealed that the change in slope between the two conditions was as predicted for all subjects (Figure 2h) ($p < 0.05$ 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 (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., 2012; 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 $e_1, \dots, e_f$, which can be expressed as the Log Posterior Odds (LPO),

$$\underbrace{\log \frac{\mathrm{p}(C=+1|e_1,\dots,e_f)}{\mathrm{p}(C=-1|e_1,\dots,e_f)}}_{\mathrm{LPO}_f} = \log \frac{\mathrm{p}(C=+1)}{\mathrm{p}(C=-1)} + \sum_{i=1}^{f} \underbrace{\log \frac{\mathrm{p}(e_i|C=+1)}{\mathrm{p}(e_i|C=-1)}}_{\mathrm{LLO}_i}, \quad (4)$$

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:

$$\mathrm{LPO}_f = \mathrm{LPO}_{f-1} + \mathrm{LLO}_f. \quad (5)$$

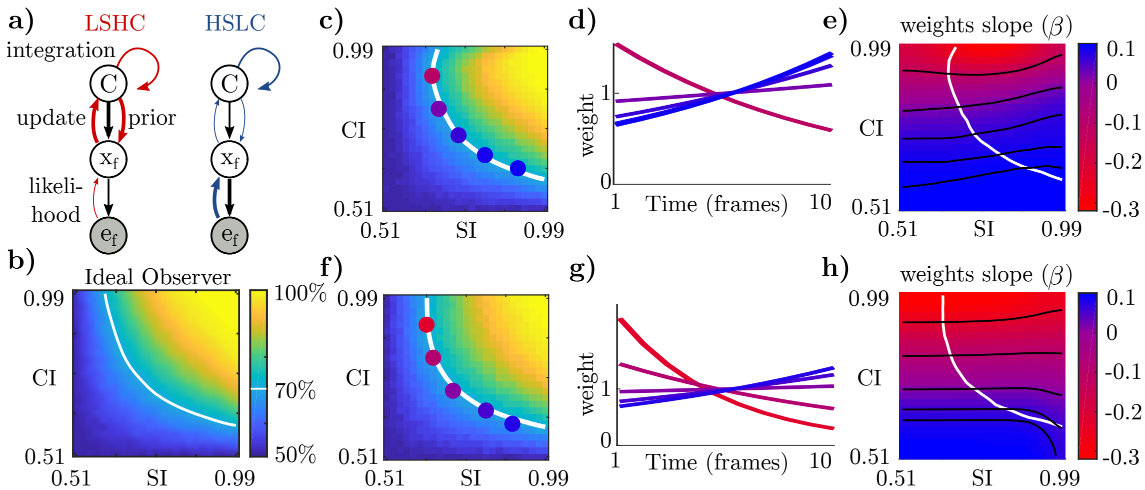

Figure 4: Approximate inference models explain results. **a)** The difference in stimulus statistics between HSLC and LSHC trade-offs implies that the relevant sensory representation is differentially influenced by the stimulus or by beliefs about the category C. A "confirmation bias" or feedback loop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

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 $p(e_f|C)$.

As we described earlier, we hypothesize that inference about x_f incorporates past information from e_1 through e_{f-1}, and this can be implemented online by feeding back information in LPO_{f-1} (equation (3)). Our models therefore assume a prior over x_f that depends on the current belief in C. 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). We emphasize again that in the case of *exact* inference, this bias that is fed back could be exactly "subtracted out" in the update to LPO_f; temporal biases arise from the combination of feedback of current beliefs *and* by the approximate nature of the representation of the posterior on x_f.

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 $\mathrm{p}_{f-1}(C)$ (equation (3), Methods). Dividing out the prior that was fed back (as in equation (2)) is naturally formulated as "importance sampling," which in our case weights each sample by the inverse of the prior (Shi and Griffiths, 2009; Murphy, 2012) (Methods). 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 likelihood from a single posterior sample. 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 the two (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 (2) 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 based on instantaneous evidence only ($\mathrm{p}(x_f|e_f)$) (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 – the "perceptual confirmation bias" mechanism introduced above. Importantly, this feedback loop is strongest when category information is high, corresponding to stronger feedback, and sensory information is low, since then x_f is both more dependent on the beliefs about C and 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-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 nearly 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:

$$\mathrm{LPO}_f \leftarrow (1-\gamma)\mathrm{LPO}_{f-1} + \hat{\mathrm{LLO}}_f \tag{6}$$

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 (6) when LPO_f is small, 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 (6) 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. We performed additional numerical experiments with the leak parameter, detailed in the Supplemental Text. Two findings are of note here. First, we found that in the regime where the confirmation bias is strongest (high category information), a moderate leak improves the model's performance, contrary to the behavior of leaky integration in models without feedback, where it impairs performance. Second, we found that if the optimal γ is used for all tasks (the value which maximizes performance), then temporal biases vanish. 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 1 hour in our case).

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., 2012) – 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., 2012; Raju and Pitkow, 2016). Inference proceeds by iteratively minimizing the Kullback-Leibler divergence between $\mathrm{q}(x)\mathrm{q}(C)\mathrm{q}(z)$ and $\mathrm{p}(x, C, z|e)$, where z is an auxiliary variable we introduce to make this a product of exponential families, as is common practice for mean field variational inference algorithms (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).

Confirmation bias, not bounded integration, explains primacy effects

The primary alternative explanation for primacy effects in fixed-duration integration tasks proposes that subjects integrate evidence to an internal *bound*, at which point they cease paying attention to the stimulus. In this scenario, early evidence almost always enters the decision-making process while evidence late in trial is often ignored. Averaged over many trials, this results in early evidence having a larger effect on the final decision than late evidence, and hence decreasing regression weights (and psychophysical kernels) just as we found in the LSHC condition (Kiani et al., 2008). While superficially similar, both models reflect very different underlying mechanism: in our approximate hierarchical inference models, a confirmation bias ensures that early evidence has a larger effect on the final decision than late evidence for every single trial. In the integration to bound (ITB) model, in a single trial, all evidence is weighed exactly the same before the bound is hit. and not at all afterwards. In order to test whether the integration to bound (ITB) mechanism could explain our results we developed a functional integration model that could be fit directly to subjects' behavior (Figure 5a). Our functional model is a simple extension to classic drift diffusion models, which can also be interpreted as integrating log odds (Gold and Shadlen, 2007). Until it hits a bound or the trial ends, the model integrates signals as follows:

$$\mathrm{LPO}_f = \begin{cases} +\mathrm{bound} & \text{if } \mathrm{LPO}_{f-1} \geq +\mathrm{bound} \\ -\mathrm{bound} & \text{if } \mathrm{LPO}_{f-1} \leq -\mathrm{bound} \\ (1-\gamma)\mathrm{LPO}_{f-1} + g(s_f) + \epsilon & \text{otherwise} \end{cases}, \tag{7}$$

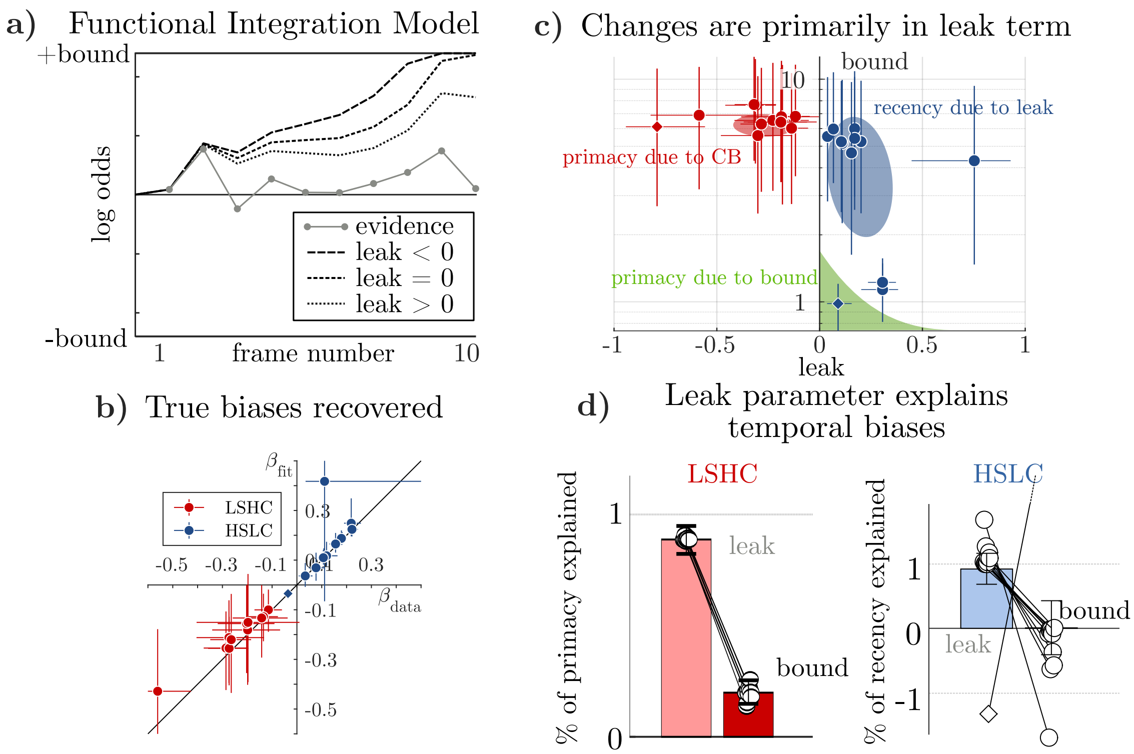

Figure 5: Results of fitting functional model show that a leak, rather than a bound, accounts for most of the observed biases. **a)** We fit a functional model of integration dynamics. As in classic drift-diffusion models, evidence is integrated to an internal bound, at which point subsequent frames are ignored. Compared to perfect integration (leak= 0), a *positive* leak (leak> 0) decays information away and results in recency effects, and a negative leak (leak< 0) amplifies already integrated information, resulting in primacy effects. Notice that $leak < 0$ may also result in more bound crossings – both the leak and the bound together will determine the shape of the temporal weights. **b)** Across both conditions, the temporal slopes (β) implied by the fit model closely match the slopes in the data. Recall that $\beta < 0$ corresponds to primacy, and $\beta > 0$ to recency. **c)** Inferred value of the bound and leak parameters in each condition, shown as median±68% confidence intervals. Ellipses depict the spread of subject means. The classic ITB explanation of primacy effects corresponds to a non-negative leak and a small bound – illustrated here as a shaded green area. Note that the three subjects near the ITB regime are points from the HSLC task – two still exhibit mild recency effects and one exhibits a mild primacy effect as predicted by ITB. **d)** We quantified the impact of the leak term and of the bound and noise terms by ablating them from the model then comparing the resulting temporal bias to the subject's actual bias (Methods). This lets us approximately quantify the fraction of each effect attributable to each parameter (but they do not necessarily sum to 1). In the LSHC condition, the (negative) leak parameter accounted for nearly all of the observed primacy effects. In the HSLC condition, the (positive) leak parameter accounted for *more* than 100% of the observed recency effects, since it was counteracted by the presence of a bound. Note that the single outlying subject (diamond symbols) corresponds to the outlying subject in panels (c) and (b) – see Supplemental Figure S13 for more information.

where s_f is the stimulus on frame f and ϵ is additive Gaussian noise. The function $g(s)$ translates from stimuli seen by the subjects into equivalent log odds, adjusting for the category and sensory information in the task (Methods). This model differs from our earlier hierarchical inference model in a few key ways. First, the signal that is integrated each frame, $g(s_f)$, is derived from the stimulus our subjects saw and contains no approximation nor inherent positive-feedback or confirmation-bias dynamics. Second, noise is added explicitly, whereas before all stochasticity came from the approximate computation of log odds, e.g. by sampling. Third, the model stops integrating when it hits an internal bound. Fourth, the model *functionally* replicates confirmation-bias dynamics by allowing the leak term, γ, to be negative; when γ is positive, information from earlier frames decays away, but when γ is negative, earlier information is amplified (Busemeyer and Townsend, 1993; Bogacz et al., 2006).

The functional model exhibits three distinct regimes of behavior. First, when the leak is positive and the bound is large, it produces recency biases. Second, when the bound is small, it produces primacy biases as in the ITB model (Kiani et al., 2008), as long as the leak is also small so that it does not prevent the bound from being crossed. Third, when the bound is large and the leak is *negative*, it also produces primacy biases but now due to confirmation-bias-like dynamics rather than due to bounded integration. In this regime, where $1 - \gamma > 1$, early evidence is "double-counted" and this model becomes functionally indistinguishable from our approximate hierarchical inference models (Supplemental Figure S10). Crucially, this means that this single model family can account for both primacy due to ITB and primacy due to a confirmation bias by different parameter values (recovery of ground-truth mechanisms shown in Supplemental Figures S11, S12) and we can use it to distinguish between the different proposals by fitting a single model to our data and examining its parameters.

We fit the functional model to sub-threshold trials from our subjects, separately for the LSHC and the HSLC conditions. We first asked whether the inferred model parameters reproduced the observed biases. Indeed, Figure 5b shows near-perfect agreement between the temporal biases implied by simulating choices from the fitted models and the biases inferred directly from subjects' choices. Figure 5c shows the posterior mean and 68% confidence interval for the leak parameter (γ) and bound parameter inferred for each subject. The model consistently infers a negative leak in the LSHC condition and positive leak in the HSLC condition for all subjects, suggesting that the confirmation-bias dynamics implied by the negative leak are crucial to explain subject's primacy biases in the LSHC condition, as well as the change in bias from LSHC to HSLC conditions. However, while the inferred bound for every single subject is so high as not to contribute at all *if the leak was zero*, it is possible that bounded integration still contributes to primacy effects, given that a stronger negative leak will hit a bound more often.

To determine the relative contribution of the leak and bound parameters to temporal biases, we simulated choices from the posterior over model parameters with either the leak parameter set to zero or after eliminating the bound (Methods). If ablating the bound leaves temporal biases unchanged, then we can conclude that biases were driven by the leak, and conversely, a temporal bias after ablating the leak must be due to the bound. We computed a population-level "ablation index" for each parameter, which is 0 if removing the parameter has no effect on β, and is 1 if removing it destroys all temporal biases. The ablation index can therefore be loosely interpreted as the fraction of the subjects' primacy or recency biases that are attributable to each parameter (but they do not necessarily sum to 1 because β is a nonlinear combination of parameters). In the LSHC condition, we found that our subjects' primacy effects are driven mostly by confirmation-bias-like integration dynamics rather than by bounded integration, though both mechanisms play some role (Figure 5d). The ablation index for the leak term was 0.89 (68% CI=$[0.87, 0.96]$), and for the bound term it was 0.19 (68% CI=$[0.15, 0.25]$) (Figure 5d). This indicates that although both mechanisms are present, primacy effects in our data are dominated by the self-reinforcing dynamics of a negative leak. In the HSLC condition, as expected, we found that recency effects are driven mostly by the leak parameter (Figure 5d). The ablation index for the leak term was 0.92 (68% CI=$[0.69, 1.17]$), and for the bound it was 0.01 (68% CI=$[-0.41, 0.43]$) (Figure 5d). The index above 1 for the leak and below 0 for the bound reflects the fact that recency effects can be balanced by the bound, so that in the absence of a leak, the bias reverts to a slight primacy effect due to an ITB mechanism, and in the absence of a mitigating bound, the recency effect appears stronger.

Interestingly, one subject exhibited a slight primacy effect in the HSLC condition, and our analyses suggest this was primarily due to bounded integration dynamics as proposed by Kiani et al (2008). This outlier subject is marked with a diamond symbol throughout Figure 5, and is further highlighted in Supplemental Figure S13. However, even this subject's primacy effect in the LSHC condition was driven by a confirmation bias (negative leak), and their change in slope between LSHC and HSLC conditions was in the same direction as the other subjects. Importantly, finding a primacy effect due to an internal bound confirms that our model fitting procedure is able to detect such effects when they are in fact present.

Two additional subjects appear to have low bounds in the HSLC condition (Figure 5c), but are dominated by the positive leak, resulting in an overall recency bias. For these subjects, the recency effect is further exaggerated when the bound is ablated, or flipped to primacy when the leak is ablated, resulting in ablation indices below 0 for the bound and above 1 for the leak (Figure 5d, steepest downward trending line in HSLC condition).

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. Fitting a phenomenological model to subjects' behavior confirmed that these changes in biases are functionally due to a change in integration dynamics rather than bounded integration. These results strongly suggest that the discrepancy in temporal biases reported by previous studies may be resolved by considering how their tasks trade off sensory and category information.

We used two distinct families of models to arrive at these conclusions. We first introduced a class of hierarchical inference models based on Importance Sampling (IS) or Variational Bayes (VB). Due to approximate inference dynamics – discussed in detail below – both of these models exhibit a confirmation bias in tasks with high category information, and they transition to recency effects in the high sensory information regime. Our hierarchical inference models distill the complexities of inference in large generative models down to just three scalar variables to isolate and study confirmation-bias dynamics, but the results generalize to higher-dimensional and deeper hierarchical models (Supplemental Figure S9). In our reduced models, we found that confirmation bias dynamics are *functionally* indistinguishable from noisy integration with a negative leak (Busemeyer and Townsend, 1993; Bogacz et al., 2006). This motivated the second class of functional or descriptive rather than mechanistic models, which allowed us to estimate the parameters of integration dynamics directly and compare this to an alternate explanation for primacy effects in the literature

(Kiani et al., 2008). Our conclusions thus proceed in two stages: first, the changes in our subjects' apparent weighting strategies are *functionally* explained by a change in the integration dynamics (primacy as $\gamma < 0$, recency as $\gamma > 0$). Second, these changes are themselves parsimoniously explained by hierarchical inference: *functional* changes in the leak parameter between tasks are a natural consequence of approximate hierarchical inference *with all model parameters, including the leak, constant across tasks*. While it is parsimonious to assume that the leak parameter is constant in the hierarchical inference models, we found that the *optimal* or normative leak parameter is high in the LSHC regime and low in the HSLC regime (Supplemental Figure S6) such that it balances the confirmation bias dynamics. Yet, we also considered the possibility that subjects infer the environment to be more volatile in the HSLC condition (Glaze et al. (2015); Figure S8), resulting in the opposite trend of stronger leak in the HSLC relative to LSHC condition. Our present data cannot speak to whether γ is truly fixed, or whether it is only constant by an accident of balancing bias-correction with a volatile environment. We leave this as a question for future work.

The "confirmation bias" emerges in our hierarchical inference models as the result of four key assumptions. Our first assumption is that inference in evidence integration tasks is in fact hierarchical, in particular that the different levels of the hierarchy require integrating evidence at different timescales, and that the brain approximates the posterior distribution over both the slow-changing category, C, and fast-changing 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., 2012) incorporating dynamic prior beliefs via feedback connections (Lee and Mumford, 2003; Yuille and Kersten, 2006; Beck et al., 2012; Nienborg and Roelfsema, 2015; Tajima et al., 2016; 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. Further, the use of discrete time in our experiment and models is only for mathematical convenience – we expect analogous dynamics to emerge in continuous-time problems that involve online inference at multiple timescales.

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 (Supplemental Figures S5-S7). The optimal amount of bias correction depends on the task statistics: in the LSHC regime where the confirmation bias is strongest, a stronger leak is needed to correct for it. While it is conceivable that the brain would optimize the amount of bias correction to the task (Brunton et al., 2013; Piet et al., 2018), our data suggest it 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 algorithmic approximations, while previously proposed "self-consistency" biases are not normative and require separate justification.

Our confirmation bias models predict attractor dynamics between different levels of the cortical hierarchy representing accumulated evidence and instantaneous sensory data. This contrasts with classic attractor models of decision-making which posit a recurrent feedback loop *within* a decision-making area (Wang, 2008; Wimmer et al., 2015). In our models, the strength of the coupling between decision-making and sensory areas depends on the category information in the stimulus. Given recent evidence that noise correlations contain a task-dependent feedback component (Bondy et al., 2018), we therefore suspect a reduction of task-dependent noise correlations in comparable tasks with lower category information. The confirmation bias mechanism may also account for the recent finding that stronger attractor dynamics are seen in a categorization task than in a comparable estimation task (Tajima et al., 2017).

Alternative models have been previously proposed to explain primacy and recency effects in evidence accumulation. We have already discussed the relation between our confirmation-bias models, bounded integration (Kiani et al., 2008), and a negative leak (Busemeyer and Townsend, 1993; Bogacz et al., 2006). 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, this model, as in the case of classic ITB models discussed earlier, depends only on the 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. While such other mechanisms can coexist with the confirmation bias dynamic proposed by our model, no previously proposed mechanism is sufficient to explain the pattern in our data for which the trade-off between sensory- and category-information is crucial. In general, *any* model based only on the total information per frame cannot explain the pattern in our data without additional parameters (such as separate leaks and bounds in each condition), which would beg additional justifications.

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 $-45°$ with $\kappa = 0.16$. 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

$$\mathbf{w}_f^{\text{exponential}} = \alpha \exp(\beta f) \tag{8}$$

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:

$$\mathbf{w}_f^{\text{linear}} = a + slope \times f \tag{9}$$

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, $\mathbf{w}/\text{mean}(\mathbf{w})$, 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 (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:

$$x_f^{(gen)} \sim \begin{cases} \mathcal{N}(+C, \sigma_x^2) \text{ with prob. equal to category info.} \\ \mathcal{N}(-C, \sigma_x^2) \text{ otherwise} \end{cases} \tag{10}$$

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

$$e_f^{(gen)} \sim \mathcal{N}(x_f, \sigma_e^2) \tag{11}$$

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

$$\text{sensory info.} = \Phi(\sqrt{2}/\sigma_e) \tag{12}$$

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 $\frac{1}{n_{\text{U}}}\hat{\text{LLO}}_f$. 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 (8)), so we use β to characterize the model's biases.

Sampling model

The sampling model estimates $\text{p}(e_f|C)$ 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:

$$x_f^{(s)} \sim Q(x) \propto \text{p}(e_f|x_f) \sum_c \text{p}(x_f|C=c)\text{p}_{f-1}(C=c) \tag{13}$$

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

$$\hat{w}^{(s)} = \left(\sum_c \text{p}(x_f^{(s)}|C=c)\text{p}_{f-1}(C=c)\right)^{-1} \tag{14}$$

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 (6):

$$\hat{\text{LLO}}_f = \log \frac{\sum_{s=1}^{S} \text{p}(x_f^{(s)}|C=+1)w^{(s)}}{\sum_{s=1}^{S} \text{p}(x_f^{(s)}|C=-1)w^{(s)}} \tag{15}$$

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 $S = 5$, $n_{\text{U}} = 5$, 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:

$$z_f = \begin{cases} +1 & \text{with probability equal to category info} \\ -1 & \text{otherwise} \end{cases} \tag{16}$$

such that

$$x_f \sim \mathcal{N}(z_f C, \sigma_x^2). \tag{17}$$

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):

$$\begin{aligned} \log \mathrm{q}(x_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + \mathit{const} \\ \log \mathrm{q}(z_f) &\leftarrow \mathbf{E}_{\mathrm{q}(C)\mathrm{q}(x_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + \mathit{const} \\ \log \mathrm{q}(C) &\leftarrow \mathbf{E}_{\mathrm{q}(x_f)\mathrm{q}(z_f)}[\log \mathrm{p}(C, x_f, z_f | e_f)] + \mathit{const} \end{aligned} \tag{18}$$

After simplifying, the new $\mathrm{q}(x_f)$ term is a Gaussian with mean given by equation (19) and constant variance

$$\mu_{x_f} \leftarrow \frac{\sigma_e^2 \mu_C \mu_{z_f} + \sigma_x^2 e_f}{\sigma_e^2 + \sigma_x^2} \tag{19}$$

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:

$$\mathrm{LPO}_{z_f} \leftarrow \log \frac{\mathrm{p}(z_f = +1)}{\mathrm{p}(z_f = -1)} + 2 \frac{\mu_{x_f} \mu_C}{\sigma_e^2 + \sigma_x^2}. \tag{20}$$

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

$$\mathrm{LPO}_C \leftarrow \log \frac{\mathrm{p}(C = +1)}{\mathrm{p}(C = -1)} + 2 \frac{\mu_{x_f} \mu_{z_f}}{\sigma_x^2} \tag{21}$$

Note that the first term in equation (21) – the log prior – will be replaced with the log posterior estimate from the previous frame (see Supplemental Algorithm S2). Comparing equations (21) and (5), we see that in the variational model, the log likelihood odds estimate is given by

$$\hat{\text{LLO}}_f = 2\frac{\mu_{x_f}\mu_{z_f}}{\sigma_x^2} \tag{22}$$

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 (22) 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.

Integration to Bound (ITB) Model

We implemented an ITB model in our simplified 3-variable hierarchical task model, $C \to x_f \to e_f$. The dynamics of the integrator model were nearly identical to equation (6), using the exact log likelihood odds, but with added noise:

$$\text{LPO}_f = \text{LPO}_{f-1}(1-\gamma) + \text{LLO}_f + \epsilon \quad , \tag{23}$$

where ϵ is zero-mean Gaussian noise with variance σ_ϵ^2 (Wong and Wang, 2006; Usher and McClelland, 2001; Bogacz et al., 2006; Brunton et al., 2013; Drugowitsch et al., 2016). Whenever LPO_f crosses the bound at $\pm B$, it "sticks" to that bound for the rest of the trial regardless of further evidence. Not that in the unbounded case noise does not affect the shape of the temporal weights (only their magnitude), but noise interacts with the bound to determine the shape as well as overall performance.

Simulations in Figure S8a-c used $\sigma_x^2 = 0.1$, $\epsilon = 0.35$, $\gamma = 0$, and $B = 1.2$. This replicates the finding of Kiani et al (2008) that bounded integration results in primacy effects. Figure S8d-f were identical except for $\gamma = 0.1$. These parameters were chosen by hand to match the magnitude and shape of the IS model's temporal weights in the LSHC condition. For Figure S8g-i, we varied γ as a function of the category information, obeying the arbitrarily chosen relationship $\gamma = 1 - CI$. In all three simulations, the model parameters were first simulated across the full space of category and sensory information to find the threshold performance curve at 70% correct. Subsequent analyses were based on points chosen to lie on the threshold performance curve, resulting in slightly different stimulus statistics for each model. This resulted in values of $\gamma = 0.09$ in the LSHC condition and $\gamma = 0.35$ in the HSLC condition for the ground-truth ITB model simulations.

Ground-truth models

To benchmark inference and as a reference for interpreting results, we simulated choices from two ground-truth models (IS and ITB) on each of two conditions (LSHC and HSLC). Both ground-truth models used parameters already described above, summarized again in Table 1, which ensured constant performance at 70% as well as a primacy effect with shape $\beta \approx -0.1$ in the LSHC condition and a recency effect with shape $\beta \approx 0.1$ in the HSLC condition for both models.

	LSHC								HSLC							
Model	**SI**	**CI**	**S**	**γ**	**B**	**ε**	**T**	**λ**	**SI**	**CI**	**S**	**γ**	**B**	**ε**	**T**	**λ**
IS	0.65	0.91	5	0.1	∞	0	0.1	0	0.91	0.63	5	0.1	∞	0	0.1	0
ITB	0.65	0.91		0.09	1.2	0.35	0.1	0	0.91	0.65		0.35	1.2	0.35	0.1	0

Table 1: Parameters of ground-truth models. **SI** = sensory information. **CI** = category information. $\boldsymbol{\gamma}$ = leak. **S** = samples per batch (IS mode only). **B** = bound (ITB model only). $\boldsymbol{\epsilon}$ = integration noise. **T** = decision temperature. $\boldsymbol{\lambda}$ = lapse rate.

Inference of ITB model parameters

The model we fit to subjects is a simple extension of the above ITB model in which the leak (γ) is allowed to be negative. Per subject per condition, we used Metropolis Hastings (MH) to infer the joint posterior over seven parameters: the category prior (p_C), lapse rate (λ), decision temperature (T), integration noise (ϵ), bound (B), leak (γ), and evidence scale (s). The evidence scale parameter was introduced because although we can estimate the ground truth category information in each task (0.6 for HSLC and 0.9 for LSHC), the effective sensory information depends on unknown properties of each subject's visual system and will differ between the two tasks. Within each task, this mapping can be approximated by simply scaling the estimated signal per frame by the constant s. To predict a subject's choices, the model thus "observed" signals equal to $\mathbf{S}/s$, where $\mathbf{S}$ is the matrix of inferred signal strengths per frame defined earlier. (Using logistic regression, we explored plausible nonlinear monotonic mappings between $\mathbf{S}$ and e and found that none performed better than linear scaling). Given s, there is no need to *additionally* infer sensory information; in our models, changing the sensory information is equivalent to rescaling the observed signal for the purposes of computing log likelihood odds. Hence a single scaling parameter s captures both the effective sensory information – which depends on each subject's visual system – as well as the mapping from the effective log odds per frame to the space of model observations (e). However, we did not include additional observation noise. We fixed the sensory information (which determines the value of σ_e^2 during inference) in the model to 0.6 in the LSHC condition and 0.9 in the HSLC condition during fitting, such that any rescaling would be captured by s. The scale s was fixed to 1 when fitting the ground-truth models, as there was no unknown mapping in those cases.

Each trial, the model followed the noisy integration dynamics in (23), where $\text{LPO}_0 = \log \frac{p_C}{1-p_C}$ and LLO_f was computed exactly conditioned on evidence $\mathbf{S}/s$. After integration, the decision then incorporated a symmetric lapse rate and temperature:

$$\text{p}(\text{Choice} = +1|\text{LPO}_F, \lambda, T) = \lambda + (1-2\lambda)\sigma\left(\text{LPO}_F/T\right) \quad ,$$

where $\sigma(a)$ is the sigmoid function, $\sigma(a) \equiv (1+\exp(-a))^{-1}$. Note that if the bound is hit, then $\text{LPO}_F = \pm B$, but the temperature and lapse still apply. To compute the log likelihood for each set of parameters, we numerically marginalized over the noise, ϵ, by discretizing LPO into bins of width at most 0.01 between $-B$ and $+B$ (clipped at 3 times the largest LPO reached by the ideal observer) and computing the *probability mass* of LPO_f given LPO_{f-1}, LLO_f, and ϵ. This enabled exact rather than stochastic likelihood evaluations within MH.

The priors over each parameter were set as follows. $\text{p}(p_C)$ was set to $\text{Beta}(2,2)$. $\text{p}(\lambda)$ was set to $\text{Beta}(1,10)$. $\text{p}(\gamma)$ was uniform in $[-1,1]$. $\text{p}(s)$ was set to an exponential distribution with mean 20. $\text{p}(\epsilon)$ was set to an exponential distribution with mean 0.25. $\text{p}(T)$ was set to an exponential distribution with mean 4. $\text{p}(B)$ was set to a Gamma distribution with (shape,scale) parameters $(2,3)$ (mean 6). MH proposal distributions were chosen to minimize the autocorrelation time when sampling each parameter in isolation.

We ran 12 MCMC chains per subject per condition. The initial point for each chain was selected as the best point among 500 quasi-random samples from the prior. Chains were run for variable durations based on available shared computing resources. Each was initially run for 4 days; all chains were then extended for each model that had not yet converged according to the Gelman-Rubin statistic, $\hat{R}$ (Gelman and Rubin, 1992; Brooks and Gelman, 1998). We discarded burn-in samples separately per chain post-hoc, defining burn-in as the time until the first sample surpassed the median posterior probability for that chain (maximum 20%, median 0.46%, minimum 0.1% of the chain length for all chains). After discarding burn-in, all chains had a minimum of 81k, median 334k, and maximum 999k samples. Standard practice suggests that $\hat{R} < 1.1$ indicates good enough convergence. The slowest-mixing parameter was the signal scale (s), with $\hat{R} = 1.13$ in the worst case. All $\hat{R}$ values for the parameters relevant to the main analysis – γ, B, and β – indicated convergence ([min, median, max] values of $\hat{R}$ equal to $[1, 1.00335, 1.032]$ for γ, $[1.0005, 1.00555, 1.0425]$ for B, and $[1, 1.0014, 1.0178]$ for all β values in ablation analyses.

Estimating temporal slopes and ablation indices implied by model samples

To estimate the the shape of temporal weights implied by the model fits, we simulated choices from the model once for each posterior sample after thinning to 500 samples per chain for a total of 6k samples per subject and condition. We then fit the slope of the exponential weight function, β, to these simulated choices using logistic regression constrained to be an exponential function of time as described earlier (equation (8)). This is the β_{fit} plotted on the y-axis of Figure 5b. For the ablation analyses, we again fit β to choices simulated once per posterior sample of model parameters, but setting $\gamma = 0$ in one case or $(B = \infty, \epsilon = 0)$ in the other.

We used a hierarchical regression analysis to compute "ablation indices" per subject and per parameter. The motivation for this analysis is that subjects have different magnitudes of primacy and recency effects, but the *relative* impact of the leak or bound and noise parameters appeared fairly consistent throughout the population (Supplemental Figure S13), so a good summary index measures the *fraction* of the bias attributable to each parameter, which directly relates to the slope of a regression line through the origin. To quantify the net effect of each ablated parameter per subject, we regressed a linear model with zero intercept to β_{fit} versus β_{true}. If an ablated parameter has little impact on β, then the slope of the regression will be near 1, so we use 1 minus the linear model's slope as an index of the parameter's contribution. The regression model accounted for errors in both x and y but approximated them as Gaussian. Defining m to be the regression slope for the population and m_i to be the slope for subject i, the regression model was defined as

$$\sigma_m \sim \text{half-cauchy}(0, 5) \tag{24}$$

$$m_i \sim \mathcal{N}(m, \sigma_m) \tag{25}$$

$$\beta_{\text{true},i} \sim \mathcal{N}(x_i, \sigma_{x,i}) \tag{26}$$

$$\beta_{\text{fit},i} \sim \mathcal{N}(x_i m_i, \sigma_{y,i})\,. \tag{27}$$

This model was implemented in STAN and fit using NUTS (Carpenter et al., 2017). Equations (24) and (25) are standard practice in hierarchical regression – they capture the idea that there is variation in the parameter of interest (the slope m) across subjects which is normally distributed with unknown variance, but that this variance is encouraged to be small if supported by the data. The variable x_i is the "true" x location associated with each subject, which is inferred as a latent variable to account for measurement error in both x (26) and y (27) dimensions. Measurement errors in β_{true}, $\sigma_{x,i}$ were set to the standard deviation in β across bootstraps. Measurement errors in β_{fit}, $\sigma_{y,i}$ were set to the standard deviation of the posterior predictive distribution over β from simulated choices on each sample of model parameters as described above.

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					
ITB Model + fitting					
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, 2012.

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.

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.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

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.

Jerome R. Busemeyer and James T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459, 1993.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

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.

Andrew Gelman and Donald B Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 —- 511, 1992.

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.

Charles D Gilbert and Wu Li. Top-down influences on visual processing. 14(May):350–363, 2013.

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.

Georg B. Keller and Thomas D. Mrsic-Flogel. Predictive Processing: A Canonical Cortical Computation. *Neuron*, 100(2):424–435, 2018.

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.

Jan-Matthis Lueckmann, Jakob H. Macke, and Hendrikje Nienborg. Can serial dependencies in choices and neural activity explain choice probabilities? *The Journal of Neuroscience*, 38(14): 2225–17, 2018.

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.

Matthias Michel and Megan A.K. Peters. Confirmation bias without rhyme or reason. *Synthese*, 2020.

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.

Hendrikje Nienborg, Marlene R Cohen, Bruce G. Cumming, Marlene R. Cohen, and Bruce G. Cumming. Decision-related activity in sensory neurons: correlations among neurons and with behavior. *Annual review of neuroscience*, 35(1):463–483, jan 2012.

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.

Kong-fatt Wong and Xiao-jing Wang. A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *The Journal of Neuroscience*, 26(4):1314–1328, 2006.

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.

November 23, 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

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

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 (15,6) 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 (19)
    $\mathrm{LPO}_{z_f} \leftarrow \log\frac{\mathrm{p}(z_f=+1)}{\mathrm{p}(z_f=-1)} + 2\frac{\mu_{x_f}\mu_C}{\sigma_x^2+\sigma_e^2}$ ▷ equation (20)
    $\mu_{z_f} \leftarrow 2(1+\exp(-\mathrm{LPO}_{z_f})^{-1} - 1$ ▷ 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 (22)
    LPO $\leftarrow$ LPO$(1-\gamma/n_\mathrm{U}) + \eta\hat{\mathrm{LLO}}_f/n_\mathrm{U}$ ▷ Equations (6) and (21) amortized for n_U updates with update strength η
  end for
end for

Optimal bias correction

A leak term approximates optimal inference in a changing environment when total evidence is weak (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 across the space of category and sensory information (Figure S6a-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 S6c). 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, the higher γ must be to correct for it (Figure S6d). Analogous results were found for the variational model (Figure S7).

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 S6e). 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).

Detailed comparison with integration to bound (ITB)

The primary alternative explanation for primacy effects in fixed-duration integration tasks proposes that subjects integrate evidence to an internal *bound*, at which point they cease paying attention to the stimulus (Kiani et al., 2008). Because the bound is crossed at different times on different trials, the average weight subjects give to each frame is a decreasing function of the frame number, i.e. a primacy effect. We implemented an integration-to-bound (ITB) observer in our hierarchical inference framework and replicated the observation that bounded and noisy integration results in primacy effects (Figure S8a-b). Importantly, this mechanism depends only on the net log likelihood per frame regardless of how it is partitioned into category information and sensory information. Classic ITB therefore always predicts the same temporal weights as long as performance is held constant. ITB does, however, predict a change in temporal weighting as a function of task difficulty, because the bound is hit earlier in a trial when evidence is stronger (Figure S8c). However, this explanation is unlikely to explain the changes seen on our data given that our experiment used a continuous staircase procedure which sustained performance near 70% in both tasks.

We next investigated the behavior of a *leaky*, noisy, and bounded integrator. While the addition of a leak term shifts the effective weights in the direction of a recency effect, we again see no systematic changes across the space of tasks (Figure S8d-f). In order to produce different regimes of temporal biases at fixed performance levels, then, either the bound, the leak term, or both must change as a function of category information and sensory information. We next simulated a leaky ITB model in which the leak term, γ, varied with category information: small γ in the LSHC regime and large γ in the HSLC regime. This change is plausible because subjects may adopt a strategy that discounts past evidence more when the world appears more volatile (Glaze et al., 2015). This model is dominated by bounded integration in the LSHC condition and by leaky integration in the HSLC condition, qualitatively reproducing the trends in our data (Figure S8g-i).

There are thus two families of models in qualitative agreement with our subjects' data: hierarchical inference with a confirmation bias, or bounded integration with a leak that depends on the task. Both model families explain recency effects as the result of leaky integration but differ in their account of primacy effects. We reasoned that these models might be distinguished using data from our LSHC condition: whereas they agree on the sign and magnitude of the temporal bias as measured by an exponential fit β, they make divergent predictions for subjects' confidence, determined by the magnitude of the integrated log odds at the end of a trial. According to the confirmation-bias mechanism, subjects should count all evidence in a trial but *over*-count early evidence, inflating their confidence relative to an unbiased integrator. According to the ITB mechanism, however, the magnitude of the bound itself sets an upper limit on log odds, and thus an upper limit on confidence, truncating the range of confidences relative to an unbiased integrator. Because we did not ask subjects to report confidence in their choices, these predictions cannot be tested directly. However, this line of reasoning suggests that these mechanisms may nonetheless be distinguished by fitting models to subjects' data; confident choices are predictable choices.

We first tested whether the two primacy mechanisms – a confirmation bias or bounded integration – are quantitatively distinguishable in ground-truth data. We simulated choices from the ground-truth IS and ITB models already described (the models plotted in Figure 4c-e and Figure S8g-i, respectively). The models were matched both in performance and in their temporal biases, exhibiting a primacy effect ($\beta \approx -0.1$) in the LSHC condition and a recency effect ($\beta \approx +0.1$) in the HSLC condition. Due to the internal stochasticity of the IS model, it is infeasible to infer its parameters directly. However, we found that an ITB model with a large bound and negative leak ($\gamma < 0$) is *functionally* indistinguishable from the IS model (Figure S10). Recall that the leak term, γ, was introduced in equation (6) and explains recency effects when $\gamma > 0$. When $\gamma < 0$, this has the opposite effect of amplifying already accumulated evidence, leading to a primacy effect due to a mechanism that is *functionally* equivalent to a confirmation bias (Busemeyer and Townsend, 1993; Bogacz et al., 2006). The key question thus becomes: are the primacy effects in our data better explained by a negative leak term or by bounded integration? These mechanisms not mutually exclusive and in principle both may contribute. We therefore fit a single ITB model with $-1 < \gamma < 1$ to each condition. By fitting a single model that contains both mechanisms as special cases, we compare them on equal terms. In order to estimate the relative contribution of each mechanism, we used MCMC sampling to infer the full posterior over all parameters (Methods).

We verified that these two distinct parameter regimes – negative leak or bounded integration – are distinguishable in ground-truth data. Indeed, in the case of the IS model, the posterior concentrated on unbounded integration with $\gamma < 0$ in the LSHC condition and unbounded but leaky integration in the HSLC condition. In the case of ground truth data from the ITB model in Figure S8g-i, the posterior concentrated around the ground truth parameters (Figure S11).

Simulation of a larger hierarchical inference model

We simulated the hierarchical sampling-based inference model of Haefner et al. (2016). Unlike our reduced $C \to x \to e$ models in the main text with only scalar variables, the model of Haefner et al. (2016) decomposes as $C \to \mathbf{G} \to \mathbf{X} \to \mathbf{I}$ where $\mathbf{I}$ is an entire image, and $\mathbf{X}$ and $\mathbf{G}$ represent entire populations of V1 and V2 neurons respectively. We will refer to this as the HBF16 model in what follows. Trying to better understand inference dynamics and the source of primacy effects in the HBF16 inspired the present work. In particular, the original model was shown to produce primacy effects in a task which we would now categorize as having low-sensory and high-category information.

The original HBF16 model was run on a coarse orientation discrimination task between low-contrast vertical and horizontal gratings embedded in white noise with variance 1. As in our reduced models in the main text, we adapted the generative model to the statistics of the stimuli as we transitioned from LSHC to HSLC conditions. In the main text, we converted sensory information into the variance of two Gaussians centered at ± 1. In the HBF16 model, sensory information is instead determined by the *contrast* of a stimulus with fixed noise. We therefore made no change to the *generative* structure of $\mathbf{X} \to \mathbf{I}$ because higher contrast images immediately results in higher signal to noise in $\mathbf{X}$. We manipulated category information in the stimulus, as in the models in the main text, by randomly flipping the orientation of each of the 10 frames per trial with probability p_{match}. Lower category information in the stimulus requires a weaker coupling from $\mathbf{G}$ to C, parameterized by κ. For each V2-like grating element G_i with preferred orientation θ_i, the generative model couples C to $\mathbf{G}$ as follows:

$$\mathrm{p}(G_i = 1|C) \propto \begin{cases} \exp(\kappa \cos(\theta_i - \theta_{C=1})) & \text{if } C = 1 \\ \exp(\kappa \cos(\theta_i - \theta_{C=2})) & \text{if } C = 2 \end{cases} \tag{1}$$

where $\theta_{C=c}$ is the true grating orientation for category $c \in \{1, 2\}$. Note that each G_i is binary, indicating the presence or absence of a grating element (see Haefner et al. (2016) for additional details). Clearly, as κ goes to zero, C and $\mathbf{G}$ become independent, and as κ gets large, C uniquely determines which grating orientation is present, and, conversely, samples of $\mathbf{G}$ strongly determine C. Thus κ controls the strength of the positive feedback or confirmation bias in this model.

The strength of the coupling between C and $\mathbf{G}$ is naturally quantified with the ROC of the two cases of von Mises distributions in (1). As in the main paper, this quantifies category information (in the generative model rather than in the stimulus) on a scale between 0.5 and 1. Denoting this function as $\mathrm{p} = \mathrm{roc}(\kappa)$ and its inverse as $\kappa = \mathrm{roc}^{-1}(\mathrm{p})$, we set κ in our simulations to $\mathrm{roc}^{-1}(\mathrm{p}_{match})/\mathrm{roc}^{-1}(0.9)$. This way, κ scaled appropriately with the amount of information in the stimulus, and $\kappa = 1$ when category information is 0.9 to approximately the original parameter regimes of HBF16.

We additionally extended the model of HBF16 to include a leak parameter in the update to the log odds of C, and set the leak to 0.01 in the simulations (equivalent to $\gamma = 0.08$ in the main paper where we divided γ by the number of updates per frame). We simulated 200 trials from the HBF16 model across a range of contrast values from 0 to 10 and p_{match} values ranging from 0.51 to 0.99. We then smoothed the resulting performance grid and plotted the results in Figure S9a, and recapitulates the patterns seen in our reduced models. We selected two points in this space – corresponding to one LSHC and one HSLC condition – for 5000 additional trials. We then computed temporal weights using AR2-regularized logistic regression. Results are plotted in Figure S9b, showing a transition from primacy in the LSHC condition to recency in the HSLC condition. (Note that without any leak, the HBF 16 model only transitions to flat weights in the HSLC condition but requires higher sensory information for equivalent LSHC performance, exactly as in our reduced models; not plotted). This demonstrates that our insights from the reduced hierarchical inference models used in the main text can generalize to larger hierarchical inference settings with a large number of variables and nontrivial dynamics.

Additional model-fitting details

To determine whether subjects' strategies were better described by confirmation bias dynamics or bounded integration, we initially sought to use standard *model comparison* methods. Ideally, Bayesian model comparison is done by computing Bayes Factors, or the ratio of the marginal likelihoods of the data under two models being compared (Bernardo and Smith, 2000). The marginal likelihood may be estimated by procedures similar to cross-validation (Fong and Holmes, 2019), which requires repeatedly performing full Bayesian inference over model parameters conditioned on random splits or subsets of the full dataset. For this to be feasible, the "inner loop" of Bayesian inference must be efficient. The primary barrier to this approach is the fact that the likelihood in the IS model is only known implicitly through stochastic simulations. Simulation-based inference methods are an active area of research (van Opheusden et al., 2020; Greenberg et al., 2019; Lueckmann et al., 2018; Papamakarios and Murray, 2016; Sisson et al., 2018; Acerbi, 2020).

For all of our models, the likelihood of the subject's choice on trial t, written $\mathrm{p}(\mathrm{choice}_t|\mathbf{S}_t, \theta)$ for stimulus sequence $\mathbf{S}_t$ and model parameters θ, is the Bernoulli probability of the observed choice given the model's confidence on the final frame, *marginalizing over the internal stochasticity of the model.* That is, for a fixed stimulus $\mathbf{S}_t$ and parameters θ, the model may output a different final log odds, LPO_F, on multiple runs. The likelihood can be written

$$\mathrm{p}(\mathrm{choice}_t|\mathbf{S}_t, \theta) = \int_{-\infty}^{\infty} \underbrace{\mathrm{p}(\mathrm{choice}_t|\mathrm{LPO}_F, \theta)}_{(i)} \underbrace{\mathrm{p}(\mathrm{LPO}_F|\mathbf{S}_t, \theta)}_{(ii)} \mathrm{dLPO}_F \quad .$$

The first term, (i), is the lapse- and temperature-adjusted probability of making a choice given a final confidence or belief value of LPO_F. The second term, (ii), depends on the internal stochasticity of the model. In the case of ITB models, all internal stochasticity is due to the integration noise ϵ, and can be numerically marginalized by internally maintaining a *distribution* of possible log posterior odds each frame, and updating that distribution for each frame, computing a new distribution, $\mathrm{p}(\mathrm{LPO}_f|\mathrm{LPO}_{f-1}, \mathrm{LLO}_f, \theta)$, taking into account the total probability mass that has crossed the bounds $\pm B$. This is precisely how we estimate the likelihood for the Metropolis Hastings sampler used in the main text. We cannot, however, apply the same trick to the IS model. Whereas the ITB models' internal stochasticity is simply additive Gaussian noise with variance ϵ^2, internal stochasticity in the IS model comes from the location of generated samples in the SNIS algorithm. If drawing $S = 5$ samples per update, as in our main simulations, then marginalization would require integrating over $\mathbb{R}^5$. In general, the marginalization problem grows exponentially with S, which is a parameter we would in principle like to infer and may be large. As a final comment before discussing alternatives, we note that SNIS with S samples can be viewed as implicitly defining a 1-dimensional distribution over x after $S-1$ marginalization steps (Cremer et al., 2017); however, this distribution is not known in closed form (or whether it has a closed form), and we were unable to derive a sub-exponential-time expression for numerically approximating it.

An cheaper alternative approach to model comparison, compared to performing full Bayesian inference in an inner-loop, is to search for the maximum likelihood or maximum a posteriori estimate of the parameters (MLE or MAP), then approximately correct for biases by adjusting the model score by the number of parameters (as in AIC) or the number of parameters and amount of data (as in BIC). *Search* methods with a stochastic objective are in general more mature than *inference* methods with stochastic likelihood evaluations, suggesting this may be a promising approach. It requires two ingredients: a method to get unbiased (but possibly variable) estimates of the log likelihood, and a method to search for the maximum of a noisy objective. We implemented the Inverse Binomial Sampling (IBS) method of van Opheusden et al (2020) to get unbiased but noisy log likelihood estimates . Briefly, IBS estimates the likelihood of each trial by counting the number of repeated (stochastic) simulations it takes before the model makes the same choice as the subject. Let k_t be the number of simulations before the first match, then IBS estimates the log likelihood for that trial as

$$\hat{LL}_t = \psi(1) - \psi(k_t) \quad , \tag{2}$$

where ψ is the digamma function (van Opheusden et al., 2020). Crucially, $\hat{LL}_t$ is an *unbiased* estimator of the true LL_t. Other naive methods derived by considering how to estimate the likelihood directly (as opposed to the log likelihood) result in biases after taking the log. The full log-likelihood estimate is given by $\hat{LL} = \sum_{t=1}^{T} \hat{LL}_t$. Its variance grows with the number of trials, so we averaged together $\sqrt{T}$ repeats of the IBS estimator per evaluation. With an unbiased estimator of the log likelihood in hand, we used Bayesian Adaptive Direct Search (BADS) to search for the maximum likelihood parameters (Acerbi and Ma, 2017). We began with a quasi-random grid of 5k points sampled from the prior over each parameter and evaluated their estimated log likelihood. For each BADS run, we perturbed the set of evaluated log likelihoods by adding Gaussian noise proportional to the empirical standard deviation of $\hat{LL}$ (i.e. Thompson Sampling), then selected the maximum as the starting point. We re-ran this procedure for at least 20 and at most 1000 searches (stopping when enough runs agreed on the value of $\hat{LL}$ at the MLE). Using the best estimate of $\hat{LL}$ for each model and condition, re-estimated with $10\sqrt{T}$ repeats of IBS, we computed AIC:

$$\text{AIC} = -2\hat{LL} + 2P \quad ,$$

where P is the number of parameters in the model. Because $\hat{LL}$ is stochastic with known empirical variance, we plotted AIC for each model fit to ground-truth data with error bars in Figure S10.

Ultimately, our conclusion from this AIC-based comparison on ground-truth models was two-fold. First, although we were able to recover the ground-truth parameters in each case, this method gives no sense of the *uncertainty* over those parameters, which is crucial for answering the question posed in the main text of the *extent* to which either of two mechanism produces primacy effects. Second, we observed that although the standard ITB model is distinguishable from the IS model with the constraint of a positive leak ($0 < \gamma < 1$) enforced, allowing negative leak ($-1 < \gamma < 1$) it is no longer distinguishable (Figure S10). In other words, this means that a negative leak is *functionally* indistinguishable from the IS model in the LSHC condition. Further, the same ITB model family with a positive leak is *functionally* indistinguishable from the IS model in the HSLC condition. Taken together, this implies that the key question of whether primacy effects are due to bounded integration or due to self-reinforcing dynamics when integrating LPO can be answered even more directly and more fairly by comparing *parameter regimes* within the ITB model family with negative leak rather than comparing across model families of IS and ITB. For this reason, we pursued full inference over ITB model parameters in the main text rather than fitting a point estimate of the IS model directly to data.

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.

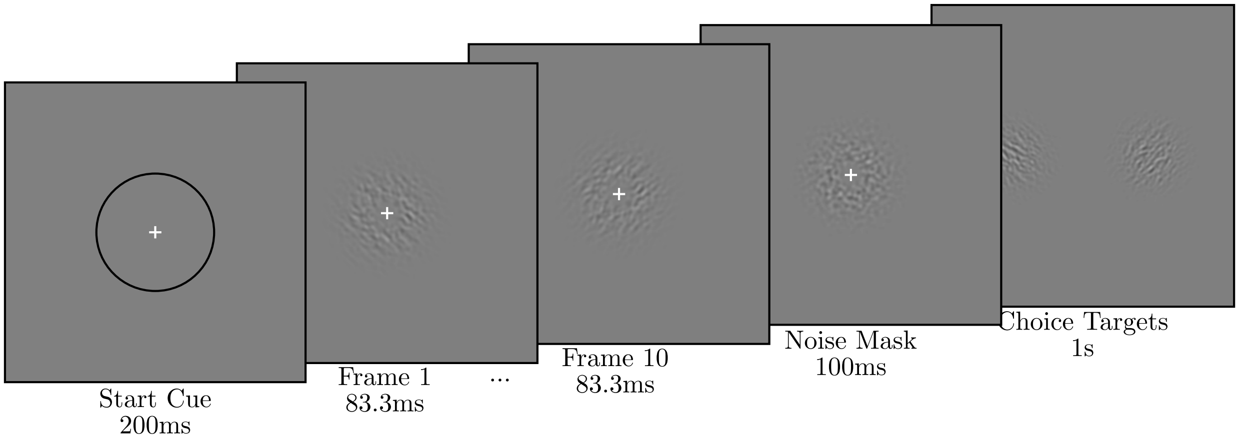

Figure S1: Stimulus timing for each trial in our visual discrimination task

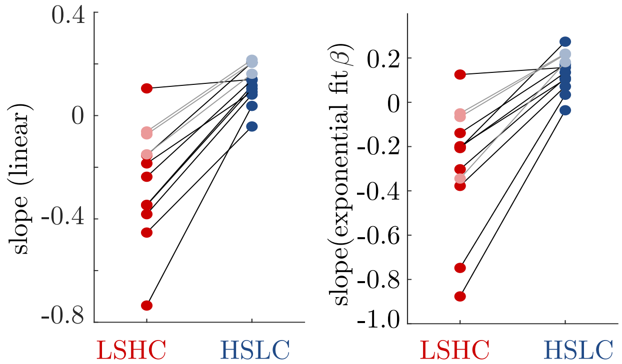

Figure S2: Same as Figure 3d in the main text, comparing slope of $\mathbf{w}$ by constraining $\mathbf{w}$ a linear (left) or an exponential (right) function of time. Using the linear fit, 10 of 12 subjects individually have a significant increase in slope ($p < 0.05$, bootstrap). Using the exponential fit, 9 of 12 subjects individually have a significant increase in slope ($p < 0.05$, bootstrap).

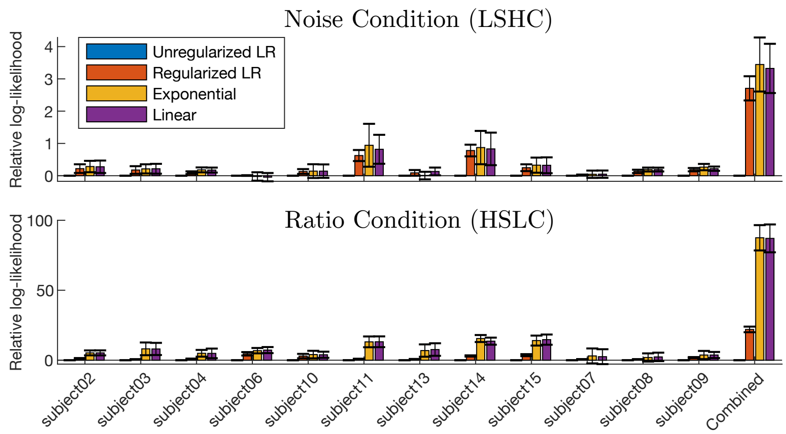

Figure S3: Cross-validation selects linear or exponential shapes for temporal weights, compared to both unregularized and AR2-regularized logistic regression. Panels show 20-fold cross-validation performance of four methods to fit evidence-weighting profiles, separated by task type and by subject. All values are relative to the log-likelihood, per fold, of the unregularized model. Error bars show standard error of the mean difference in performance 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 separately for each subject. "Exponential" is is the 3-parameter model where weights are an exponential function of time (equation (8) plus a bias term). Similarly, the "Linear" model constrains the weights to be a linear function of time as in equation (9), plus a bias term.

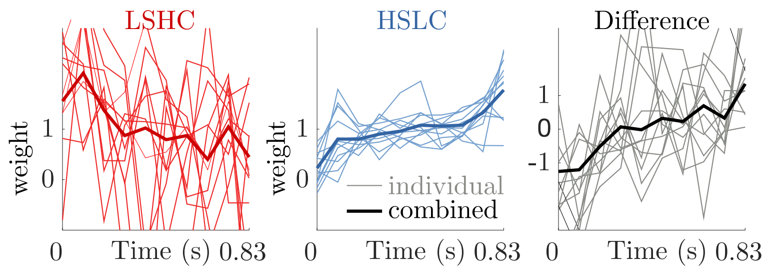

Figure S4: Same as Figure 3a-c in the main text, but with no regularization applied to logistic regression for individual subjects. Both here and in the main text, the "combined" weights are computed using the un-regularized individual weights.

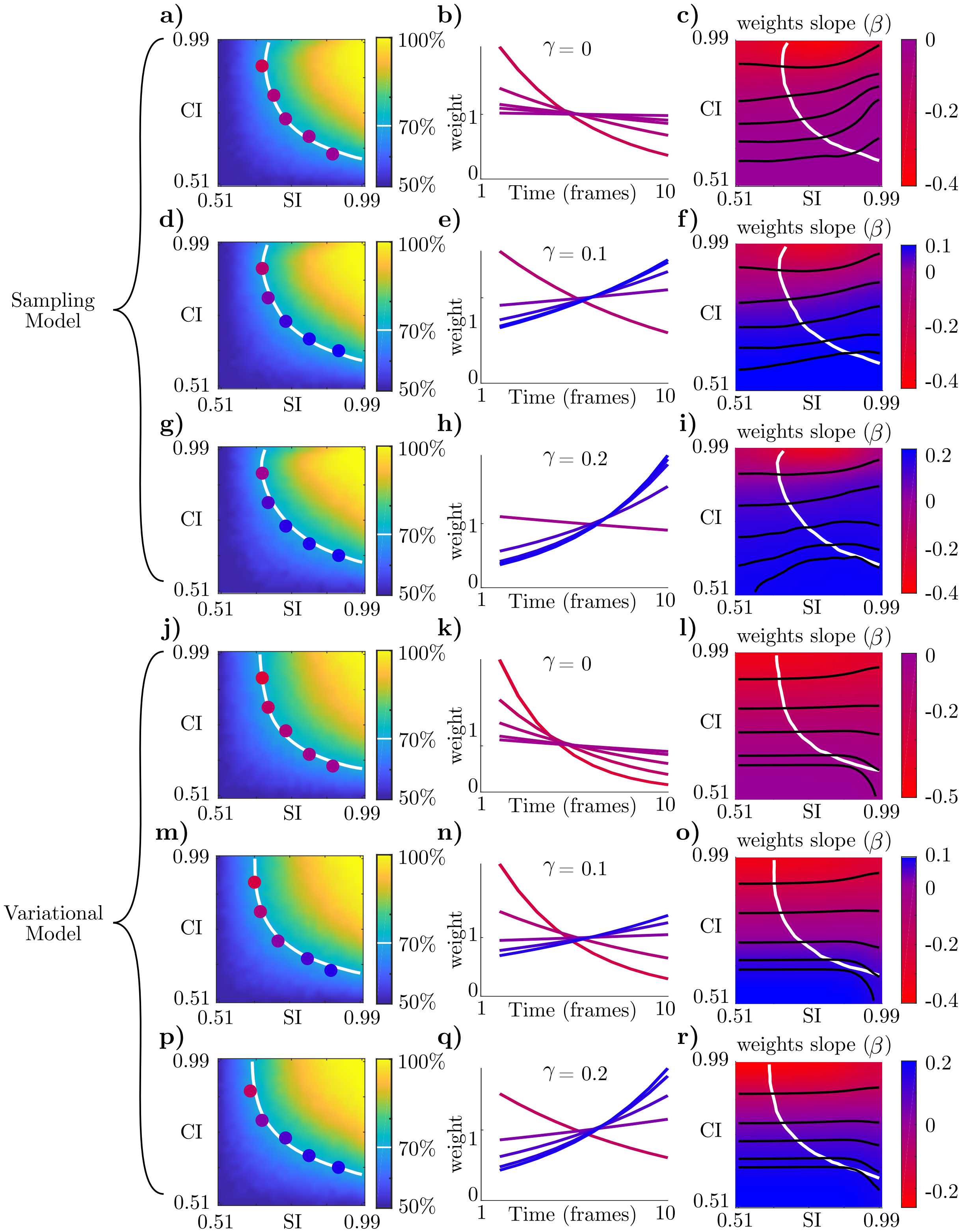

Figure S5: In both models, larger γ increases the prevalence of recency effects across the entire task space. Panels are as in Figure 4 in the main text. **a-c** sampling model with $\gamma = 0$. **d-f** sampling model with $\gamma = 0.1$. **g-i** sampling model with $\gamma = 0.2$. **j-l** variational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

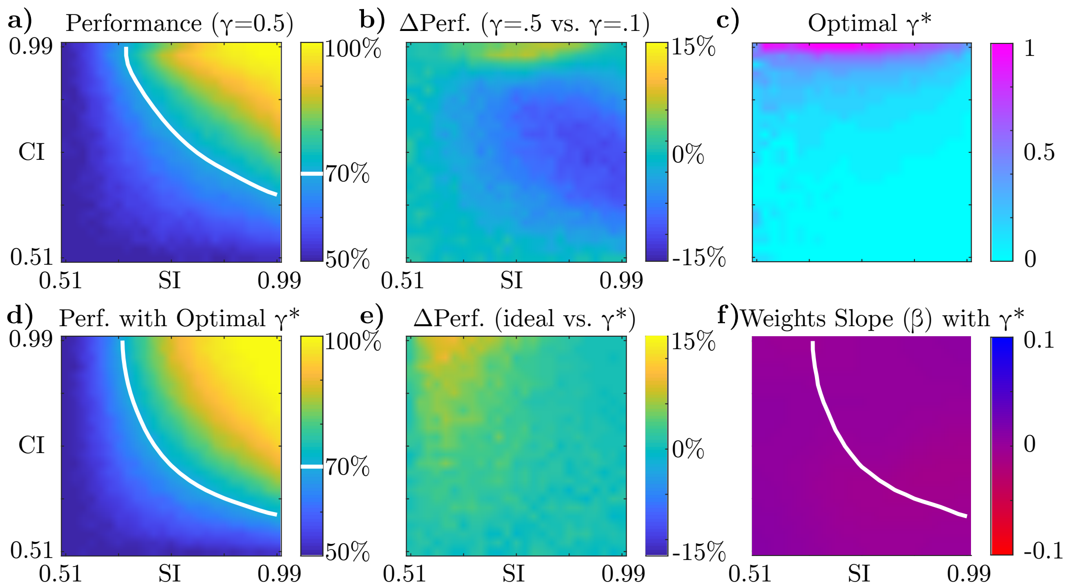

Figure S6: Optimizing performance with respect to γ (see also Figure S7). **a)** Sampling model performance across task space with $S = 5$ and $\gamma = 0.5$ (compare with Figure 4c in which $\gamma = 0.1$). **b)** Difference in performance for $\gamma = 0.5$ versus $\gamma = 0.1$. Higher γ improves 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).

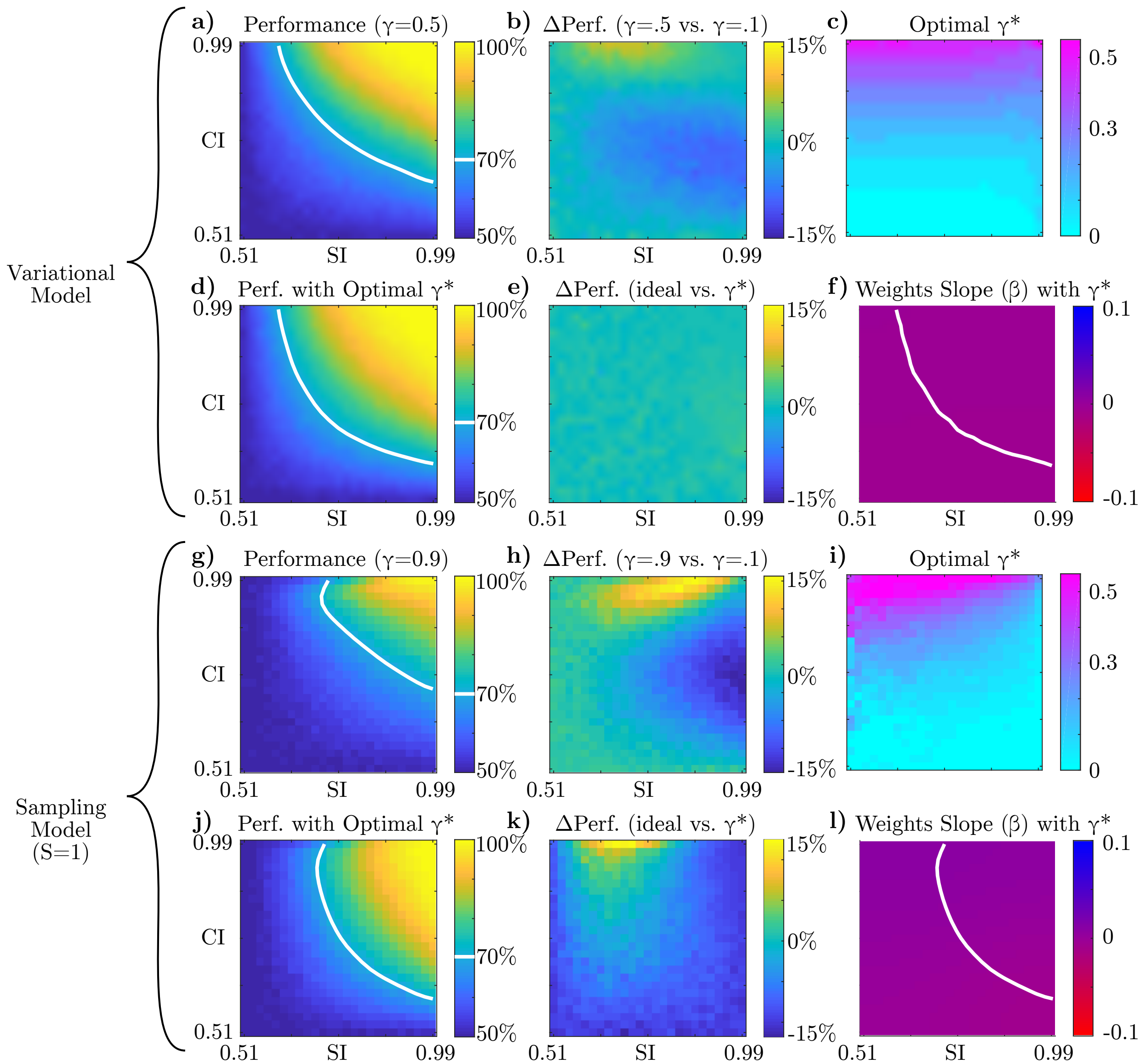

Figure S7: Simulation results for optimal leak (γ) for two further model variations, panels as in Figure S6. **a-f** Variational model results. As in the sampling model, we see that the optimal value of $\gamma*$ increases with category information, or with the strength of the confirmation bias. **h-l** Sampling model results with $S = 1$ (in the main text and Figure S6 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.

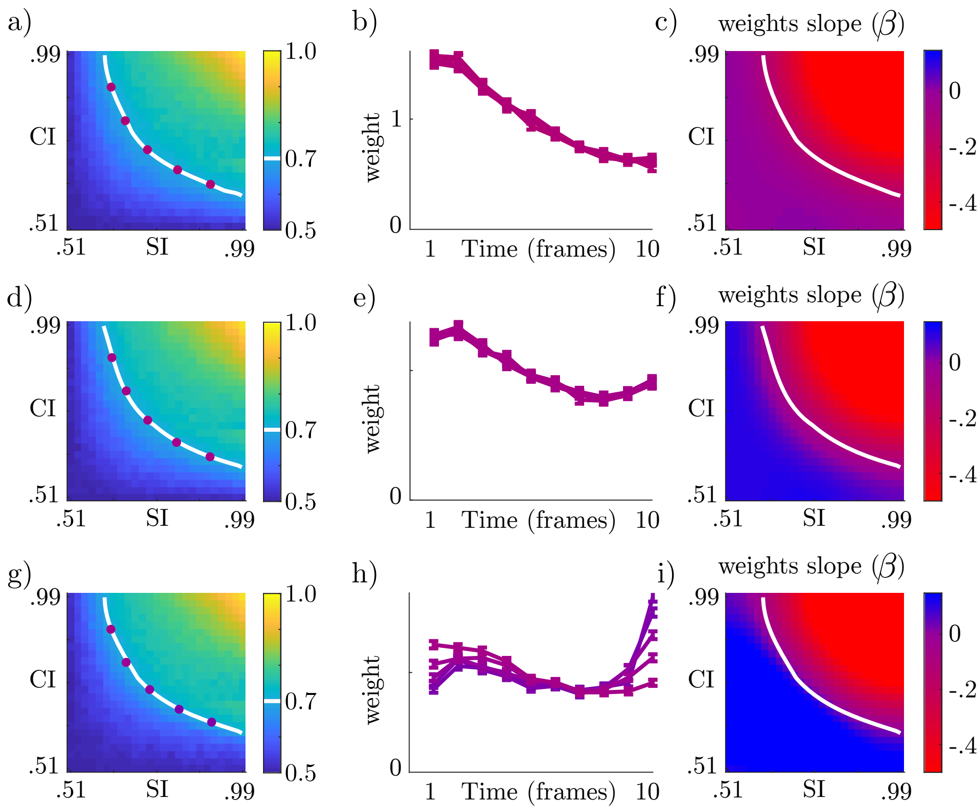

Figure S8: Simulation of bounded integration (ITB) model. **a)** Performance of an ITB model is not differentially modulated by sensory and category information. **b)** ITB consistently produces primacy effects, as in (Kiani et al., 2008). **c)** The primacy effect becomes more extreme in regions where evidence is stronger. **d-f)** As in (a-c), but with an additional leak term, resulting in less extreme primacy effects and a transition to recency for *difficult* tasks, but no transition from primacy to recency along the iso-performance contour. (Also note the departure from monotonic exponential-like weight profiles). **g-i)** We now vary the leak term, γ, as a function of category information. This reproduces the qualitative transition from primacy in LSHC to recency in HSLC. As measured by an exponential fit (β), slopes are matched to those in the confirmation bias models (Figure 4d,g).

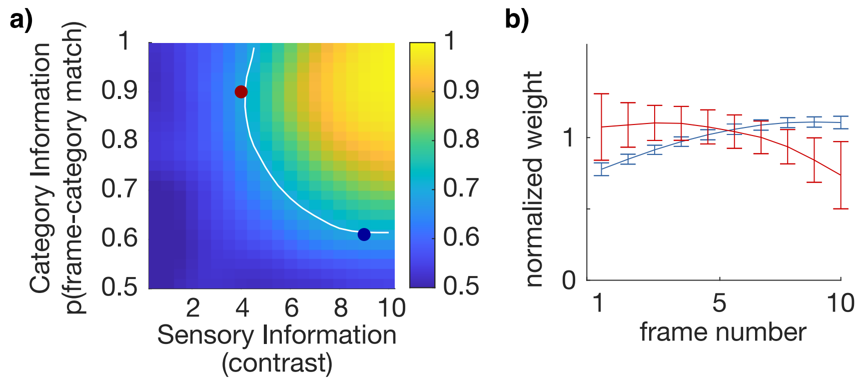

Figure S9: Simulation results on the larger model of Haefner et al. (2016). **a)** Performance as a function of sensory information (grating contrast) and category information (probability that each frame matches the trial category). White line is iso-performance contour at 70%, and dots correspond to LSHC and HSLC parameter regimes plotted in (b). Simulation details in the Supplemental Text. **b)** Temporal weights from LSHC and HSLC simulations corresponding to colored points in (a), normalized in each condition so the weights have mean 1. As in the reduced models in the main text, we see a transition from primacy to recency.

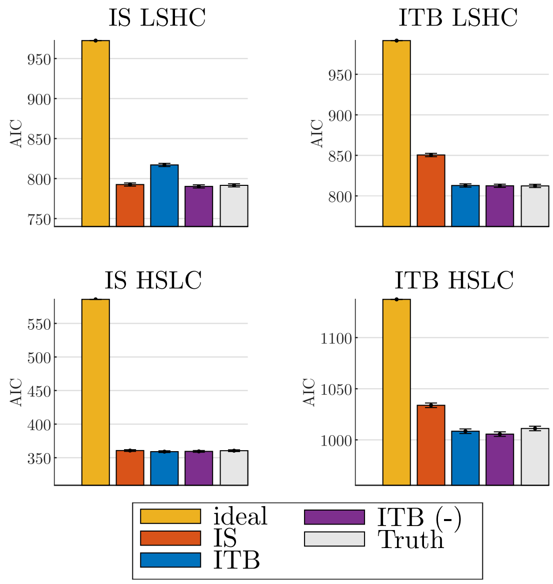

Figure S10: Results of direct model comparison between IS model and ITB model(s) fit to ground-truth data. We employed methods to search the log likelihood landscape of each model despite the stochastic likelihood evaluations of the IS model (van Opheusden et al., 2020; Acerbi and Ma, 2017). Lower AIC indicates better fit. An ideal integrator (gold) and ground-truth (gray) values serve as upper- and lower-bounds, respectively, on plausible AIC values. In all cases, the best fitting model recovered parameters that are as good as the ground truth. The standard ITB model (with positive leak enforced) is distinguishable from the IS model in the LSHC simulation (top row). However, an extended ITB model that allows for negative leak ("ITB (-)", purple), fits all data in all conditions as well as the ground-truth. For this reason, we state in the main text that a negative leak is *functionally* indistinguishable from the true IS model. We pursued *parameter comparison* within this extended ITB (-) model class, rather than *model comparison* between IS and ITB, in the main text.

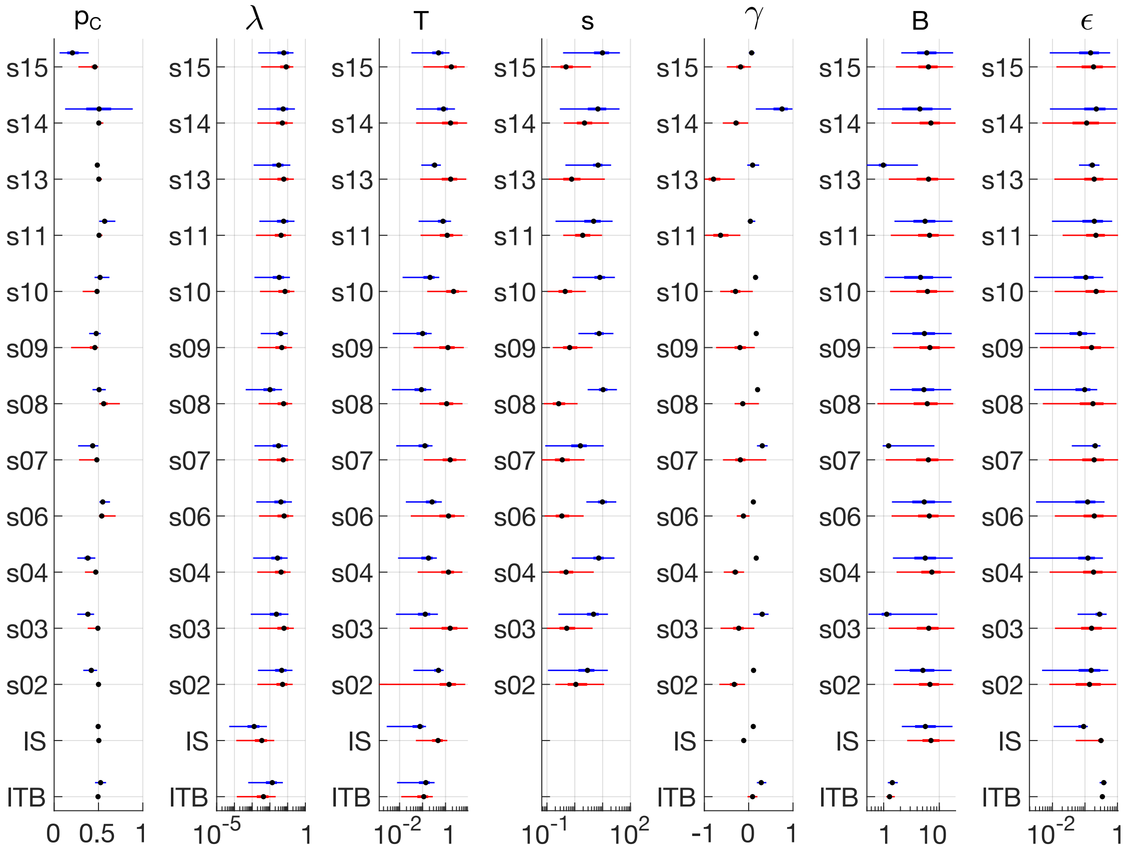

Figure S11: Box and whisker plots of inferred parameter values for each of 12 subjects as well as the ground truth models (IS and ITB). Each parameter and subject has two fits, one for the LSHC condition (lower/red) and one for the HSLC condition (upper/blue). Thin lines are 95% posterior interval, thick lines are 50% interval, and points are posterior median. Parameter names are as in the main paper, restated here: p_C = prior over categories, λ = symmetric lapse rate, T = decision temperature, s = signal scale (fixed to 1 for ground truth models), γ = leak, B = bound, ϵ = noise.

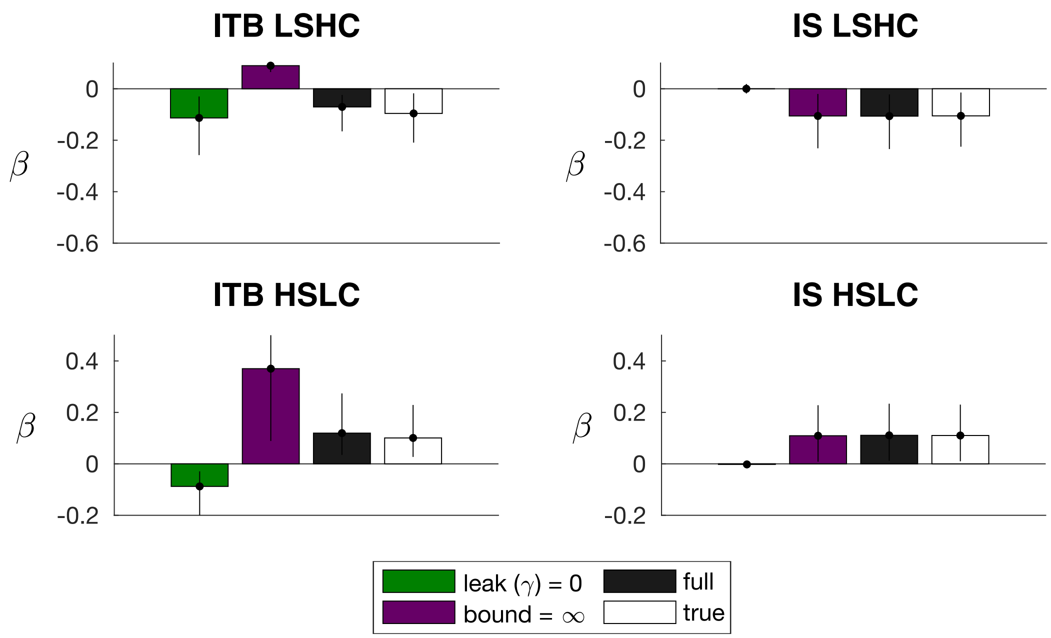

Figure S12: Recovery of true temporal weight slopes (β) and ablations on ground-truth models. White bars ("true") are bootstrapped (β) values on the ground-truth choices. Black bars ("full") are (β) values implied by simulating choices from the full inferred model. Green and purple bars are (β) values either after ablating the leak or after ablating the bound and noise, respectively, as described in Methods of the main text. In the **ITB LSHC** panel, note that ablating the leak has little effect, but ablating the bound reverse the effect to recency; this is consistent with the ground-truth mechanism: primacy due to bounded integration rather than a negative leak. In contrast, the **IS LSHC** primacy effect is completely destroyed by ablating the (negative) leak but unaffected by ablating the bound. Taken together, these **ITB LSHC** and **IS LSHC** simulations suggest we can identify which mechanism is responsible for primacy effects. In both **HSLC** panels (bottom row), ablating the (positive) leak term has the strongest effect, destroying recency in the **IS** case and reversing the effect to primacy in the **ITB** case, since the resulting model is purely a bounded integrator. Ablating the bound in the **ITB HSLC** case leaves the leak unmitigated, resulting in an even stronger primacy effect.

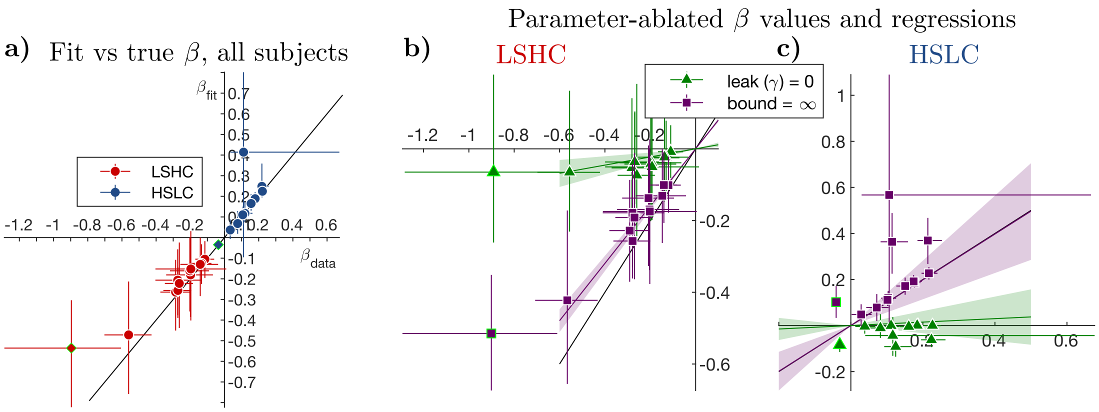

Figure S13: Additional information on model fits and ablation regressions. **a)** Identical to Figure 5b in the main text, but zoomed out to show outlying subjects as well. The diamond symbol in (a) and lime green borders in (b-c) indicate the one identified outlier. **b)** As in (a), this shows the model's temporal slope (β) on the y-axis versus the subject's actual temporal slope on the x-axis, but with either the leak parameter ablated (green triangles) or the bound and noise parameters ablated (purple squares). Each subject appears as 2 points that share an x-coordinate (slightly jittered for for visualization), plotted as mean±68% confidence intervals. The fact that β is near 0 when the leak term is ablated implies that the leak term is the primary driver of primacy effects in the LSHC condition. Population-level regression slope ("m" from equation (25)) mean and 65% error bars are shown as lines with shading. **c)** Same as (b) but for the HSLC condition. All subjects except the one outlying subject (lime green border) had a recency effect which disappears or is reversed to primacy when the leak is ablated (green points).

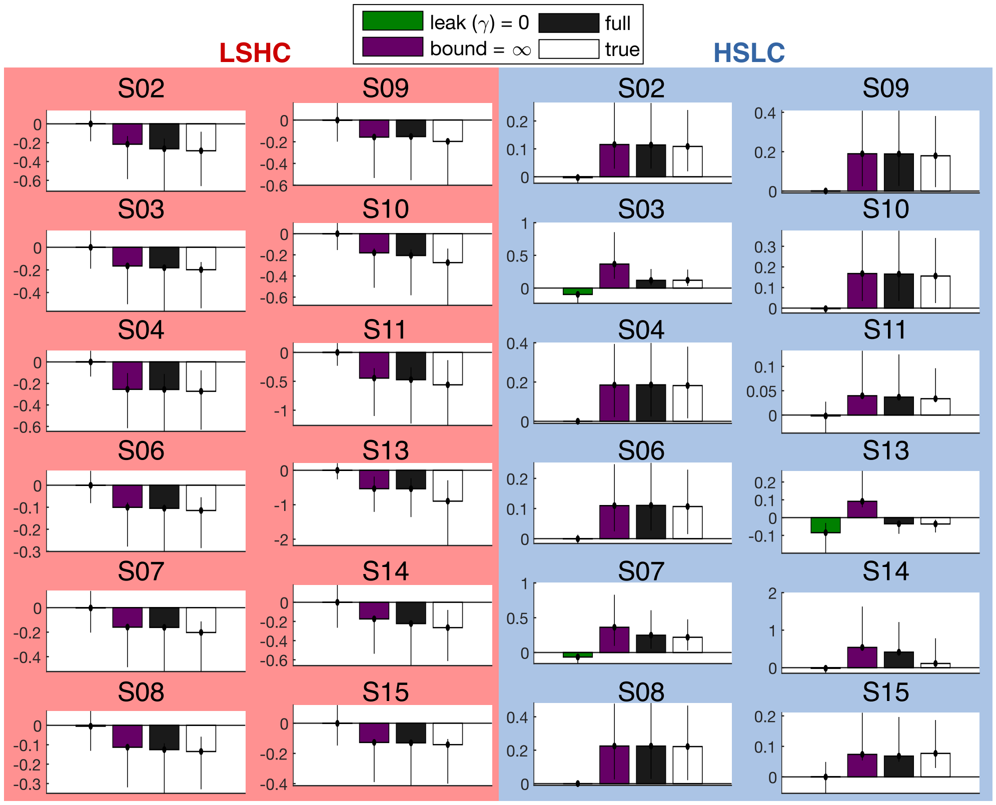

Figure S14: Temporal weight slopes (β) and ablations, broken out by individual subject. White bars ("true") are bootstrapped β_{data} values on the subject's choices. Black bars ("full") are β_{fit} values implied by simulating choices from the full inferred model. Green and purple bars are β_{fit} values either after ablating the leak or after ablating the bound and noise, respectively, as described in Methods of the main text.

References

Luigi Acerbi. Variational Bayesian Monte Carlo with Noisy Likelihoods. *arXiv*, 2020.

Luigi Acerbi and Wei Ji Ma. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems*, 30:1837–1847, 2017.

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

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, West Sussex, England, 2 edition, 2000.

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.

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.

Jerome R. Busemeyer and James T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3): 432–459, 1993.

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

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.

Edwin Fong and Chris Holmes. On the marginal likelihood and cross-validation. *arXiv*, (1): 1–16, 2019.

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

David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic Posterior Transformation for Likelihood-Free Inference. 2019.

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

Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H. Macke. Likelihood-free inference with emulator networks. pages 1–21, 2018.

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.

George Papamakarios and Iain Murray. Fast e-free inference of simulation models with Bayesian conditional density estimation. *Advances in Neural Information Processing Systems*, (Nips):1036–1044, 2016.

Scott A. Sisson, Yanan Fan, and Mark A. Beaumont. Overview of Approximate Bayesian Computation. In Scott A. Sisson, Yanan Fan, and Mark A. Beaumont, editors, *Handbook of Approximate Bayesian Computation*, chapter 1. CRC Press, 2018.

Bas van Opheusden, Luigi Acerbi, and Wei Ji Ma. Unbiased and Efficient Log-Likelihood Estimation with Inverse Binomial Sampling. pages 1–89, 2020.

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.