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.

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

Imagine a doctor trying to infer the cause of a patient's symptoms from an x-ray image. Unsure about the evidence in the image, she asks a radiologist for a second opinion. If she tells the radiologist her suspicion, she may bias his report. If she does not, he may not detect a faint diagnostic pattern. If the evidence in the image is hard to detect or ambiguous, the radiologist's second opinion, and hence the final diagnosis, may be swayed by the doctors initial hypothesis. We argue that the brain faces a similar problem during perceptual decision-making: any decision-making area combines sequential signals from sensory brain areas, not directly from sensory input. If those signals themselves reflect inferences that combine both prior expectations and sensory evidence, we suggest that this can then lead to an observable confirmation bias (*1*).

Formalizing this idea in the context of approximate Bayesian inference requires extending classic evidence-integration models to include an explicit intermediate sensory representation (Figure 1b). We explicitly model the inferences of the intermediate sensory representation and find that task difficulty is modulated by two distinct types of information: the information between the stimulus and sensory representation (sensory information), and the information between sensory representation and category (category information) (Figure 1b). The balance between these distinct types of information can indeed explain puzzling discrepancies in the literature with regards to the temporal weighting of evidence across a wide range of studies. Even in tasks where all evidence is equally informative about the correct category, existing studies typically report one of three distinct motifs: some find that early evidence is weighted more strongly (a primacy effect) (*2, 3*) some that information is weighted equally over time (as would be optimal) (*4–6*), and some find late evidence being weighted most heavily (a recency effect) (*7*) (Figure 1a,c). There are myriad differences between these studies such as subject species, sensory modality, stimulus parameters, and computational frameworks (*2, 5, 7, 8*). However, none of these aspects can explain their different findings, whereas the differences arise naturally in a hierarchical approximate inference framework.

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

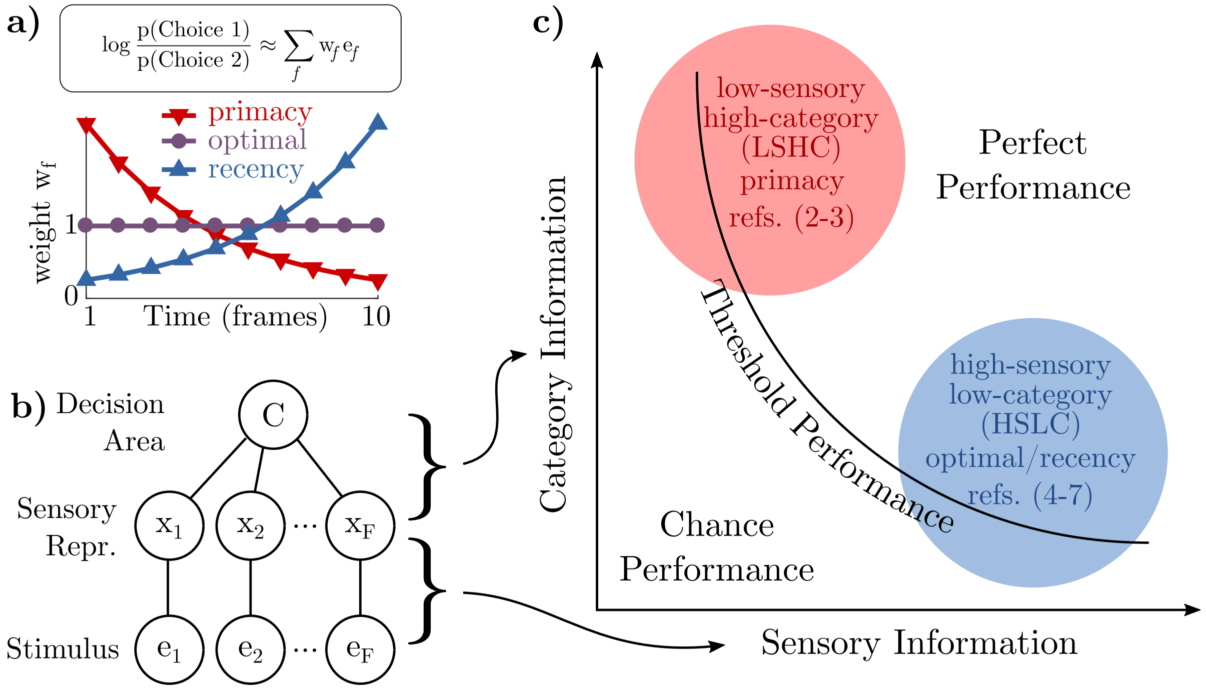

Figure 1: **a)** A subject's "temporal weighting strategy" is an estimate of how their choice is based on a weighted sum of each frame of evidence e_f. Three commonly observed motifs are decreasing weights (primacy), constant weights (optimal), or increasing weights (recency). **b)** Uncertainty in the stimulus about the category may be decomposed into uncertainty in each frame about a sensory variable ("sensory information") and uncertainty about the category given the sensory variable ("category information"). **c)** Category information and sensory information may be altered independently, creating a two-dimensional space of possible tasks, where subjects will be at threshold performance whenever the two sources of information are balanced. A qualitative placement of previous work into this space separates those that find primacy effects in the upper-left from those that find recency effects or optimal weights in the lower right (see Supplemental Text for detailed justification).

stimulus. On each trial in a typical task, the stimulus consists of multiple "frames" presented in rapid succession. (By "frames" we refer to discrete independent draws of stimulus values that are not necessarily visual). If the evidence in each frame, e_f, is independent and the categorical identity of the stimulus is a binary variable $C \in \{-1, +1\}$, then evidence in favor of $C = +1$ after F independent frames is $\mathrm{p}(C = +1|e_1, \ldots, e_F) \propto \mathrm{p}(C = +1) \prod_{f=1}^{F} \mathrm{p}(e_f|C = +1)$. The ideal observer reports the most likely category, for instance by reporting the sign of $\log \mathrm{p}(C = +1|e_1, \ldots, e_F) - \log \mathrm{p}(C = -1|e_1, \ldots, e_F)$.

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

To illustrate this difference, consider the classic dot motion task (*9*) and the Poisson clicks task (*5*), which occupy opposite locations in the space spanned by sensory and category information. In the classic low-coherence dot motion task, subjects view a cloud of moving dots, some percentage of which move "coherently" in one direction. Here, sensory information is low since evidence about the net motion direction at any time is weak. Category information, on the other hand, is high, since knowing the "true" motion on a single frame would be highly predictive of the correct choice (and of motion on subsequent frames). In the Poisson clicks task, subjects hear a random sequence of clicks in each ear and must report the side with the higher rate. Here, sensory information is high since each click is well above sensory thresholds, but category information is low since knowing the side on which a single click was presented provides only little information about the correct choice (and the side of the other clicks). Another way to think about category information is as "temporal coherence" of the stimulus: the more each frame of evidence is predictive of the correct choice, the more the frames must be predictive of each other, whether a frame consists of visual dots or of auditory clicks. Note that our distinction between sensory and category information is different from the well-studied distinction between internal and external noise; in general, both internal and external noise will reduce the amount of sensory and category information.

If we assume that the sensory representation, which itself is an inference about the actual stimulus, incorporates prior expectations (*10–12*), then, as we show below, approximate inference models predict that this will lead to a primacy effect when sensory information is low and category information is high, but not when sensory information high and category information is low. Indeed, a qualitative placement of prior studies in the space spanned by these two kinds of information demonstrates that studies that find early weighting are located in the upper left quadrant (low-sensory/high-category or LSHC) and studies with equal or late weighting in the lower right quadrant (high-sensory/low-category or HSLC) (Figure 1c). This suggests that the different trade-off between sensory information and category information may indeed underlie differences in temporal weighting seen in previous studies. Further, with this framework it is straightforward to predict how simple changes in stimulus statistics should change the temporal weighting (Table S1). To test this critical model prediction, we designed a visual discrimination task with two stimulus conditions that correspond to the two opposite sides of this task space, while keeping all other aspects of the design the same (Figure 2a). If our theory is correct, then we should be able to change individual subjects' temporal weighting strategy simply by changing the sensory–category information trade-off.

The stimulus in our task consisted of a sequence of ten visual frames (83ms each). Each frame consisted of band-pass-filtered white noise with excess orientation power either in the $-45°$ or the $+45°$ orientation (*13*) (Figure 2b,d). On each trial, there was a single true orientation category, but individual frames might differ. At the end of each trial, subjects reported whether the stimulus was oriented predominantly in the $-45°$ or the $+45°$ orientation. The stimulus was presented as an annulus around the fixation marker in order to minimize the effect of small fixational eye movements (Supplemental Methods).

If the brain represents the orientation in each frame, then sensory information in our task is determined by how well each frame determines the orientation of that frame (i.e. the amount of "noise" in each frame), and category information is determined by the probability that any given frame's orientation matches the trial's category. For a ratio of $5 : 5$, a frame's orientation does not predict the correct choice and category information is zero. For a ratio of $10 : 0$, knowledge of the orientation of a single frame is sufficient to determine the correct choice and category information is high. For a more detailed discussion, see Supplementary Text.

Using this stimulus, we tested 12 human subjects (9 naive and 3 authors) comparing two conditions intended to probe the difference between the LSHC and HSLC regimes. Starting with both high sensory and high category information, we either ran a staircase lowering the sensory information while keeping category information high, or we ran a staircase lowering category information while keeping sensory information high (Figure 2a). These are the LSHC and HSLC conditions, respectively (Figure 2b,d). For each condition, we used logistic regression to infer, for each subject, the influence of each frame onto their choice. Subjects' overall performance was matched in the two conditions by defining threshold performance as 70% correct (supplementary materials).

In agreement with our hypothesis, we find predominantly flat or decreasing temporal weights

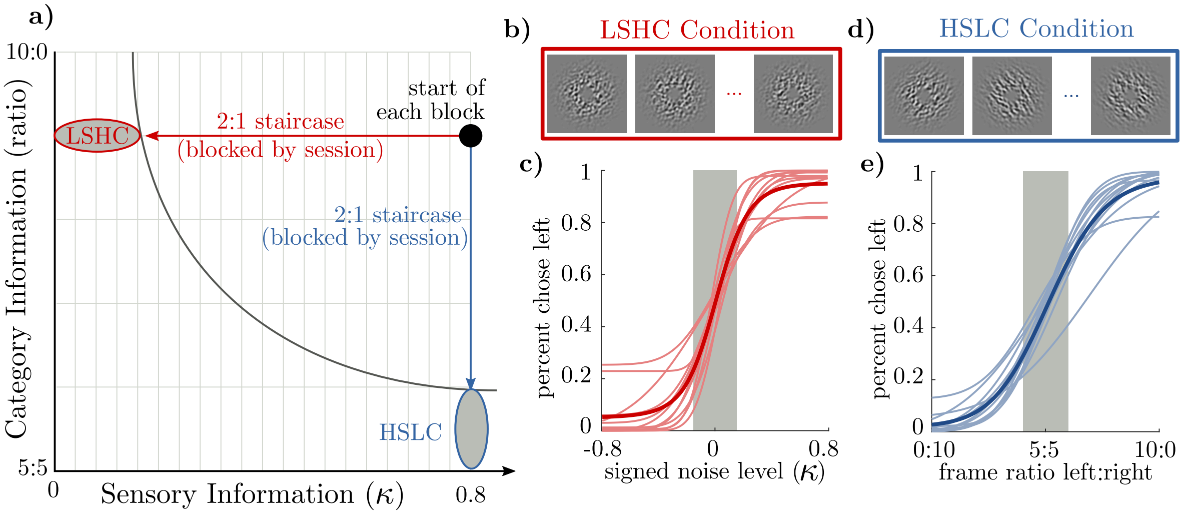

Figure 2: Summary of experiment design. **a)** In our task, category information is determined by the ratio of frame categories, and sensory information is determined by a noise parameter κ. At the start of each block, we reset the staircase to the same point, with category information at $9:1$ and κ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

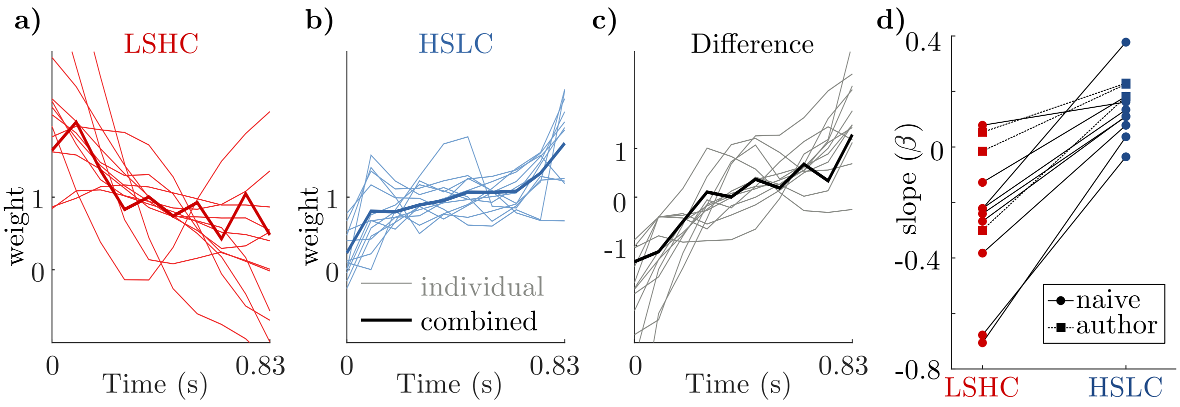

Figure 3: Regression of subjects' temporal weights. **a-b)** Temporal weights for individual subjects (thin lines) and the mean across all subjects (thick lines). Weights are always normalized to have a mean of 1. Individual subjects' curves were fit using a cross-validated smoothness term (Supplemental Methods). **c)** Difference of normalized weights (HSLC−LSHC). Despite variability across subjects in (a-b), each subject reliably changes in the direction of a recency effect. **d)** Change in slope between the two task contexts for each subject is consistently positive. We summarize subjects' temporal weighting strategy with an exponential fit; the slope parameter $\beta > 0$ corresponds to recency and $\beta < 0$ to primacy (Supplemental Methods).

when sensory information is low and category information is high (Figure 3a). When the information is partitioned differently – into high sensory and low categorical information – we find flat or increasing weights (Figure 3b). Despite variability between subjects in each condition, a within-subject comparison revealed that the change in slope between the two conditions was as predicted for every single subject (Figure 2c,d) ($p < 0.05$ for 9 of 12 subjects, bootstrap). This demonstrates that the trade-off between sensory and category information in a task robustly changes subjects' temporal weighting strategy as we predicted, and reconciling the discrepant results in the literature.

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 (*11, 14–16*) and a parametric (mean-field) approximation (*17,18*) to exact inference can explain the observed pattern of changing temporal weights as a function of stimulus statistics.

The crucial assumption in both models is that the brain computes a posterior belief over both C and x given the external evidence, i.e. $\mathrm{p}(x, C|e)$, not just over the categorical variable C. This assumption differs from some models of approximate inference in the brain that assume populations of sensory neurons strictly encode the *likelihood* of the stimulus (*19*), but is consistent with other models from both sampling and parametric families (*11, 12, 18*).

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

Importantly, the strength of this confirmation bias depends on the relative amount of sensory and category information in the stimulus (Figure 4a). It is weakest when the posterior over x is dominated by the likelihood, a case that occurs when the category information is much weaker than the sensory information. Conversely, the feedback loop is strongest when the category information is high compared to the sensory information, as assumed in (*11*) who found a primacy effect in their model.

To demonstrate and quantify the intuitions laid out above, we implemented approximate online inference (observing a single frame at a time) for a discrimination task using two previously proposed frameworks for how inference might be implemented in neural circuits: neural sampling (*11, 14–16*) and Mean Field Variational Inference (*17*) (Figure 4). The central operation in either case is the evaluation of the following ratio (Supplemental Methods)

$$\log \frac{\mathrm{p}(e_f|C=+1)}{\mathrm{p}(e_f|C=-1)} = \log \frac{\int_{x_f} \mathrm{p}(e_f|x_f)\mathrm{p}(x_f|C=+1)\mathrm{d}x_f}{\int_{x_f} \mathrm{p}(e_f|x_f)\mathrm{p}(x_f|C=-1)\mathrm{d}x_f} \tag{1}$$

which quantifies how much the brain's belief about C should be changed as the result of the current evidence e_f (*20*). The assumption that information about C is fed back to x is operationalized differently in each model, but the effects on the models' behavior are the same (Figure 4, Supplemental Figure S5).

The neural sampling hypothesis states that variable neural activity over time can be interpreted as a sequence of samples from the brain's posterior over x. The prior belief about C biases the distribution from which samples are generated. The canonical way to compute an expectation with respect to one distribution (the likelihood) using samples from another (the posterior) is 'importance sampling' which weights each sample so as to "subtract out" the prior as described above. While this approach is unbiased in the limit of infinitely many samples,

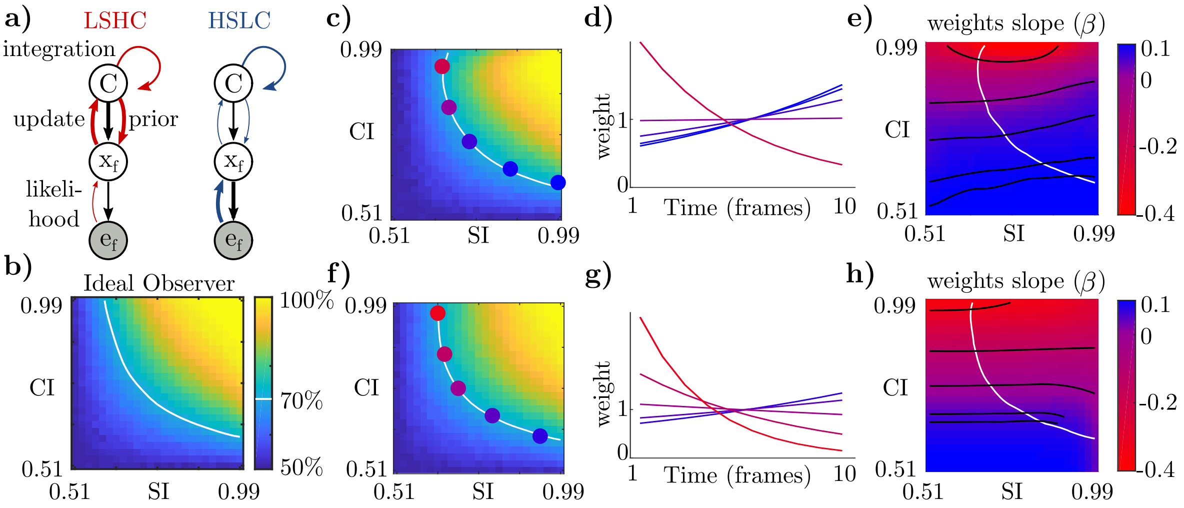

Figure 4: Approximate inference models explain results. **a)** The difference in stimulus statistics between HSLC and LSHC trade-offs implies that the relevant sensory representation is differentially influenced by the stimulus or by beliefs about the category C. A "confirmation bias" or feedback loop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

it incurs a bias for a finite number – the relevant regime for the brain. The bias is such that it under-corrects for the prior that has been fed back, resulting in a positive feedback loop (see Supplemental Methods). Figure 4b,c shows performance for the ideal observer and for the sampling model, respectively, across all combinations of sensory and category information. White lines show threshold performance (70% correct) as in Figure 1c. This model reproduces the primacy effect, and how the temporal weighting changes as the stimulus information changes seen in previous studies. Importantly, it predicted the same within-subject change seen in our data (*11*). However, double-counting the prior alone does not yet explain recency effects (Figure S5a-c,j-l). A simple and biologically-plausible explanation for recency effects is that the brain tries 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. This reduces the primacy effect in the upper left of the task space and leads to a recency effect in the lower right (Figure 4c-h, Supplemental Figure S5), as seen in the data. Interestingly, a linear bias correction term takes the same form as a "leak" term in the classic drift-diffusion framework (Supplemental Methods), and has been shown to be optimal for inference in non-stationary environments (*8, 20*).

The second major class of models for how probabilistic inference may be implemented in the brain – based on mean-field parametric representations (*17, 19*) – 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. (*18*). 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, $\mathrm{q}(x)\mathrm{q}(C)$ (*17, 18*). Inference proceeds by iteratively minimizing the Kullback-Leibler divergence between $\mathrm{q}(x)\mathrm{q}(C)$ and $\mathrm{p}(x, C|e)$ (Supplemental Methods). As in the sampling model, the running estimate the category C acts as a prior over x. Because this model is unable to explicitly represent posterior dependencies between sensory and decision variables, it is biased to commit early either to both x and C being positive or to both x and C being negative. This yields the same behavior as the sampling model: a transition from primacy to flat weights as category information decreases, with recency effects emerging only when approximate bias correction is added (Supplemental Figure 4f-h). Whereas the limited number of samples was the key deviation from optimality in the sampling model, here it is the assumption that the brain represents its beliefs separately about x and C in a factorized form and that its instantaneous belief about x is unimodal.

Both models induce a confirmation bias by creating an "attractor" dynamic between different levels of the cortical hierarchy – the decision-making area and the relevant sensory areas. Our model therefore makes two testable neurophysiological predictions when subjects show a primacy effect: (i) the presence of so-called "differential correlations" (*11, 21*) in populations of task-relevant sensory neurons, and (ii) a reduction of those correlations, as well as any primacy effect, when cortical feedback is inactivated. Our model further predicts that attractor-like dynamics in sensory cortex will depend on the decision-making context, as was recently reported (*22*). This observation, as well as our two novel predictions, contrasts with classic attractor models which posit a recurrent feedback loop within a decision making area (*23*).

As in the classic sequential probability ratio test, both models maintain a running estimate of posterior odds over time (*20*). The confirmation bias mechanism is thus complementary to other aspects of evidence integration like "integration to bound" (*2*) or uncertainty over stimulus strength (*24*).

In the brain, decisions are not based directly on external evidence but on intermediate representations . If those intermediate representations themselves in part reflect prior beliefs, and if inference in the brain is approximate, then this is likely to result in a bias. The nature of this bias is directly related to the integration of internal "top-down" beliefs and external "bottom-up" evidence previously implicated in clinical dysfunctions of perception (*25*). Importantly, we have shown how the strength of this effect depends on the nature of the information in the task in a way that may generalize to cognitive contexts where the confirmation bias is typically studied.

References

1. R. Nickerson, *Review of general psychology* **2**, 175 (1998).
2. R. Kiani, T. D. Hanks, M. N. Shadlen, *The Journal of neuroscience* **28**, 3017 (2008).
3. H. Nienborg, B. G. Cumming, *Nature* **459**, 89 (2009).
4. V. Wyart, V. D. Gardelle, J. Scholl, C. Summerfield, *Neuron* **76**, 847 (2012).
5. B. W. Brunton, M. M. Botvinick, C. D. Brody, *Science* **340**, 95 (2013).
6. D. Raposo, M. T. Kaufman, A. K. Churchland, *Nature Neuroscience* **17**, 1784 (2014).
7. J. Drugowitsch, V. Wyart, A.-D. Devauchelle, E. Koechlin, *Neuron* **92**, 1398 (2016).
8. C. M. Glaze, J. W. Kable, J. I. Gold, *eLife* **4**, 1 (2015).
9. W. T. Newsome, E. B. Pare, *The Journal of Neuroscience* **8**, 2201 (1988).
10. T. S. Lee, D. Mumford, *Journal of the Optical Society of America A* **20**, 1434 (2003).
11. R. M. Haefner, P. Berkes, J. Fiser, *Neuron* **90**, 649 (2016).
12. C. I. Tajima, *et al.*, *Nature Scientific Reports* **5**, 1 (2016).
13. W. H. A. Beaudot, K. T. Mullen, *Vision Research* **46**, 26 (2006).
14. P. O. Hoyer, A. Hyvärinen, *Advances in neural information processing systems* **17**, 293 (2003).

15. J. J. Fiser, P. Berkes, G. Orbán, M. Lengyel, *Trends in cognitive sciences* **14**, 119 (2010).

16. G. Orbán, P. Berkes, J. Fiser, M. Lengyel, *Neuron* **92**, 530 (2016).

17. J. Beck, K. Heller, A. Pouget, *Advances in Neural Infromation Processing Systems* **25**, 3068 (2013).

18. R. V. Raju, X. Pitkow, *NIPS* **30** (2016).

19. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, *Nature neuroscience* **9**, 1432 (2006).

20. J. I. Gold, M. N. Shadlen, *Annual review of neuroscience* **30**, 535 (2007).

21. R. Moreno-Bote, *et al.*, *Nature Neuroscience* **17**, 1410 (2014).

22. S. Tajima, *et al.*, *eLife* **6**, 1 (2017).

23. X. J. Wang, *Neuron* **60**, 215 (2008).

24. S. Deneve, *Frontiers in Neuroscience* **6**, 1 (2012).

25. R. Jardri, S. Denéve, *Brain* **136**, 3227 (2013).

Acknowledgments

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).

Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1-S5

Tables S1-S2

References *1-15*