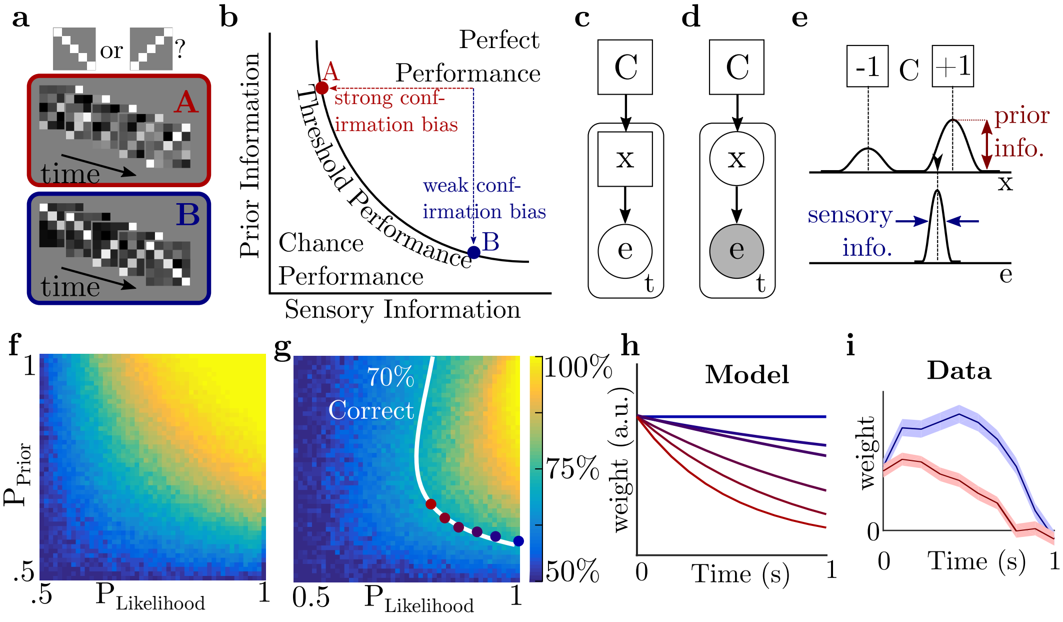
Perceptual confirmation biases from approximate online inference

Richard D. Lange*, Ankani Chattoraj*, Matthew Hochberg, Jacob Yates, Ralf M. Haefner

Brain and Cognitive Sciences, University of Rochester

{rlange, achattor, jyates7, rhaefne2}@ur.rochester.edu, mhochbe2@u.rochester.edu

Summary The confirmation bias (CB) is ubiquitous in psychological studies, but its computational and neural basis is unclear. An analogous effect is seen in some psychophysics studies: using reverse correlation, it has been shown that subjects overweight information presented early in a trial to make a choice[1,2] (a CB). Other studies, however, have found that subjects equally weight all information in a trial[3,4]. We introduce an intuitive probabilistic framework that distinguishes between these studies as having different sources of uncertainty in the mapping between the stimulus at any moment and the correct choice at the end of the trial. Those studies in which a CB is seen use stimuli with strong temporal correlations and weak sensory information, while those with no CB have the reverse. While exact inference in either case entails no CB, approximate inference methods may. We simulate decision-making in a sampling-based model[5,6,7], in which we make explicit that the brain relies on intermediate representations between sensory and decision-making areas. Our model qualitatively shows the same trends in evidence weighting between the two types of task. Finally, we report preliminary data from a visual discrimination task that enables us to directly explore these different sources of uncertainty and their effect on measured CBs. These results offer a concise and unified perspective on existing evidence integration studies, and provide a valuable framework for future task design.



Additional Detail Consider the two orientation discrimination tasks in Fig 1a. In task A, low-contrast gratings embedded in dynamic pixel noise each give only weak information about the correct choice. In task B, each frame is high-contrast, but the class itself alternates from frame to frame and subjects must report the majority. In both tasks, subjects must integrate uncertain sensory information over time, but the source of the uncertainty is very different. In Fig 1b we show these two tasks as occupying different regions of a "prior information" versus "sensory information" space. We define "prior information" as the probability that an ideal observer could predict the category of a single frame given the full trial's correct choice. We similarly define "sensory information" as the ideal observer's ability to identify each frame's category given a single image. When both prior and sensory information are high (e.g. a single high contrast image), tasks are extremely easy. When both are low (e.g. a sequence of noisy and flickering frames), tasks are extremely difficult. In between, there is a line of threshold performance where the different sources of uncertainty are traded off to calibrate task difficulty. Existing studies have fallen in different places along this line. Like in task A described above, Nienborg et al. (2009) and Kiani et al. (2008) both use stimuli with low sensory information. In contrast, Brunton et al. (2013) and Wyart et al. (2012) both use stimuli with high sensory but low prior information, as in task B. Interestingly, while the former studies found

*these authors contributed equally

that subjects weight early information more strongly (consistent with a CB), the latter found constant evidence weighting over time, and so far this discrepancy has been unexplained.

We use a simple task model to explore this space of prior-vs-sensory information (Fig 1c). Binary variables are shown in boxes and continuous variables in circles. Each trial has a single true category C chosen at random. The category underlying each frame, x_t, is chosen to either agree or disagree with C with some probability that is equivalent to the "prior information." The actual stimuli $e_1, \ldots, e_T$ are selected from a normal distribution centered on the corresponding x_t where the variance of this distribution controls "sensory information." The performance of an ideal observer in this space is shown in Fig 1f.

We assume that the brain has learned this generative model and makes a choice each trial by *approximating* the posterior over C given the sequence of $e_1, \ldots, e_T$ (Fig 1d). We use a continuous x_t to represent an instantaneous sensory percept (Fig 1d-e), with a bimodal normal distribution for $P(x_t|C)$. (Note that this is equivalent to the discrete model when the variance of each mode is zero). Exact inference in this model also entails no CB. We make three core assumptions that together lead to it displaying a CB. First, we assume that the brain uses sampling to do approximate inference[5,6,7]. Second, we assume a decision-making area that updates a running estimate of $P(C|e_1, ..., e_t)$ using each incoming sample of x_t. Third, we assume feedback from the decision area[8] to x_t such that samples of x_t are drawn from the full posterior, accounting for the current belief about $P(C)$. Together, these assumptions result in a feedback loop between x_t updating $P(C)$ and $P(C)$ influencing subsequent samples of x_{t+1}, and this feedback loop causes early evidence to have a larger impact on the ultimate choice[6]. The performance of this model is illustrated in Fig 1g. Because the influence of C on x is strongest in tasks with high "prior information," our model exhibits a stronger CB as the task statistics move from high-sensory to high-prior information (Fig 1h; weight profiles corresponding to points in g), despite being at 70% performance in all cases.

Finally, we report preliminary data from the visual tasks illustrated in Fig 1a. In one condition, we lower the contrast to reach threshold performance (task A, Fig 1a), and in another we hold the images at high contrast and lower their "prior information" (task B, Fig 1a). We use regularized logistic regression to compute subjects' evidence-weighting profile over time and plot the resulting (unnormalized) weights in Fig 1i. Across subjects, the high-likelihood experiment (blue) yields relatively flat weights for the first two thirds of the trial, compared to the steadily decreasing weights seen in the high-prior experiment (red). These early results are consistent with our model's predictions (and, to our knowledge, are the first results to demonstrate different weighting profiles within subjects in a single paradigm).

Prior work by Kiani et al. 2008 modeled this decreasing weight effect using a classic ideal observer model. In their model, subjects reach an internal confidence "bound" and ignore subsequent information, which appears as a slow ramp down in weight when averaged across trials. This model and ours are not mutually exclusive; indeed, the decrease in weights we observe at the end of high sensory information trials can likely be explained by such a bound. However, the difference between conditions for the first two thirds of trials cannot easily be explained in the same way.

The "feedback loop" in our model discussed above is reminiscent of attractor models of decision making[9]. Rather than being attractor dynamics within a decision making area, however, the feedback loop in our model exists between the sensory and decision making areas, enabling us to make further testable predictions for top-down influences on sensory neurons during decision making[6,8,10]. Furthermore, our model builds a bridge to the idea of perception as Bayesian inference[11]; our model exhibits a confirmation bias as direct consequence of approximate inference in a generative model of the task. Finally, we believe that the distinction between prior-information and sensory-information is a useful and intuitive framework that will inform the design of future experiments.

[1] Nienborg & Cumming *Nature* 2009. [2] Kiani et al. *The Journal of Neuroscience* 2008. [3] Brunton et al. *Science* 2013. [4] Wyart et al. *Neuron* 2012. [5] Fiser et al. *TICS* 2010. [6] Haefner et al. *Neuron* 2016. [7] Orbán et al. *Neuron* 2016. [8] Nienborg et al. *Annual Review of Neuroscience* 2012. [9] Wong & Wang. *The Journal of Neuroscience Neuroscience* 2006. [10] Lange & Haefner. *Biorxiv* 2016. [11] Knill & Richards. *Perception as Bayesian Inference* 1996.