On the computational basis of the confirmation bias

Richard D. Lange[1,2], Ankani Chattoraj[1], Ralf M. Haefner[1]
[1]Brain and Cognitive Sciences, University of Rochester
[2]Computer Science, University of Rochester
rlange@ur.rochester.edu, achattor@ur.rochester.edu, ralf.haefner@gmail.com

Abstract

The confirmation bias is one of the most ubiquitous of inferrential errors known in psychology [1]. Interestingly, it is also observed in low-level perceptual decision-making tasks [2]. Based on our previous work on a neural sampling based model of perceptual decision-making [3], we propose a new computational explanation for the confirmation bias that can be extended to higher cognitive processes. We suggest that the confirmation bias is the result of an online inference process in which internal beliefs are updated based on representations of the posterior belief (rather than likelihood, or based on the external evidence directly), which leads to a 'double-counting' of early evidence. We compare our model with a previous explanation [4] for the confirmation bias in low-level vision and present an empirical test that can distinguish between both models. Existing evidence from two recent empirical studies favors our model [5, 6].

1 Introduction

Integrating evidence from different sources is one of the most elementary computations carried out by the brain. The experimental paradigm of perceptual decision-making offers an ideal context for studying the neural and computational basis of this integration of evidence since the neural basis is well established, and due to the mathematical tractability of simple two alternative forced choice (2AFC) tasks. Here, we use this paradigm to investigate the confirmation bias – the empirical observation that when integrating a temporal sequence of pieces of evidence, subjects often weigh initial evidence more strongly than evidence presented later in a trial. We follow a top-down methodology along Marr's level of explanation that starts with a definition of the computational goal, before specifying representations and algorithms (process) for how the brain might achieve that goal [7].

We compare the standard model of perceptual decision-making [8, 4] with a model based on the neural sampling hypothesis, the idea that neural responses represent samples from the brain's posterior beliefs over the variables that they represent [9, 3]. Both models exhibit a confirmation bias; however, we will show that they differ in how the strength of this bias depends on the average magnitude (i.e. variance) of evidence being integrated, making them distinguishable empirically.

2 Results

2.1 Computation

In the class of perceptual decision-making tasks that we consider, the subject is presented with a sequence of n stimuli (e.g., images) that are predictive of the correct choice at the

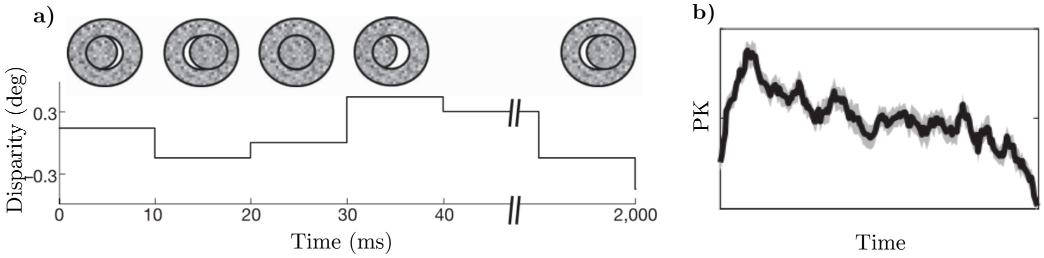

Figure 1: Example task. **a)** The subject has to judge whether the central disk is in front or behind the reference plane (binocular disparity of 0). 200 frames of differing disparity are shown over the course of a single trial, all constituting independent evidence. **b)** Empirical psychophysical kernel (PK). (Both panels adapted from [2].)

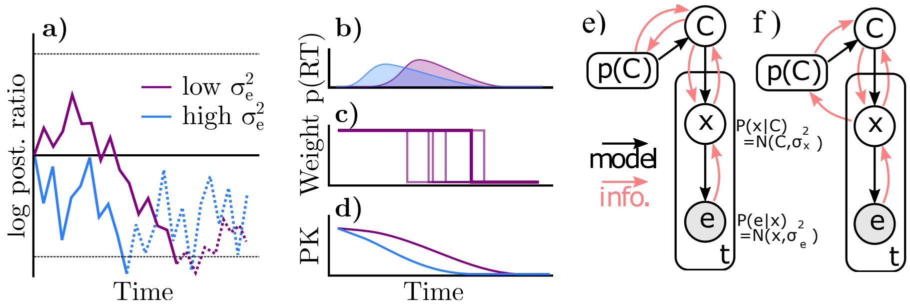

Figure 2: Bounded integration model. **a)** the log-posterior ratio is tracked over time. Once this quantity reaches the upper or lower bound (thin black lines), the subject commits to a decision. **b)** In trials with larger evidence variance (light blue), the bound is reached earlier on average (analogous to reaction time distributions). **c)** on any given trial, evidence weight is a step function. **d)** PK is the average of the step functions in **c**. This model predicts that higher σ_e^2 should result in a steeper slope of the PK. **e,f)** We compare two sampling-based update rules for $p(C)$. The graphical model that we sample from is shown with black arrows. The red arrows show information flow for the two update rules. The variables x and e are dependent on t, but only a single value for C is inferred.

end of the trial. The stimuli are uncorrelated and the correct choice is defined by the experimenter independent of the ordering. The evidence may consist of random moving dots [8], binocular disparity (Figure 1)[2], or oriented gratings [6].

In such tasks, the "weight" given to each piece of evidence as a function of when it was presented during the trials is commonly measured using the *psychophysical kernel* (PK). It can be computed either by reverse correlation between signal level and the subject's eventual choice [2], or logistic regression (as the weight of the stimulus presented at each point in time) [6]. A decreasing PK means that evidence at the beginning of the trial was given a bigger weight by the subject – a confirmation bias. An increasing PK would indicate a 'forgetful' subject (e.g., leaky integration of evidence [4]).

On each 'frame' in a trial, evidence e_t is drawn i.i.d. from some distribution. We can use Bayes' rule: $p(C|e_1, ..., e_n) \propto p(C) \prod_{i=1}^{n} p(e_i|C)$ to make explicit that $p(C|e_1, ...e_n)$ is independent of the order in which evidence is presented, implying a constant PK for the ideal observer. Our analysis will make predictions about how the *variance* of the evidence distribution (σ_e^2) affects the slope of the PK for the models we consider.

2.2 Bounded Integration Model

The bounded integration model [8] is illustrated and explained in Figure 2a-d. While an integration of all evidence over the entire duration of a trial would be the optimal strategy, it has been proposed that subjects only integrate until they have obtained sufficient evidence to convince themselves of the correct choice [4]. This means that evidence presented after that point is ignored, and since this point differs from trial to trial (sampled from the reaction time distribution), this will lead to a monotonically decreasing PK when averaged over many trials.

2.3 Sampling Model

Our sampling model assumes that the brain has learned a probabilistic internal model of the task, and that inference over unobserved variables is carried out by MCMC sampling [3]. Traditional ideal observer analyses only consider $p(C|e)$, while the classic perceptual decision-making framework models $p(C|r)$ and $p(r|e)$ separately. Instead, we assume that the brain performs inference over both C and x at the same time, and that the responses r represent the brain's posterior belief over x (Figure 2e,f). As a result, the sensory response reflect both feedforward and feedback information (shown by red arrows) [3].

We further assume that the brain maintains a graded representation of the current belief about C, $p_t(C) \equiv p(C|e_1, ...e_t)$ in a decision-making area like LIP [8]. Due to our assumption that the brain represents the posterior over its unobserved variables, during the inference process, samples drawn from (x, C) at time t in the trial will incorporate the accumulated belief as a prior.

We consider two "update rules" that the brain could plausibly use to estimate the posterior $p(C|e_1, ..., e_n)$ using sampled values of C and x. The first is an intuitive "counting" rule which estimates $\frac{p_t(C=+1)}{p_t(C=-1)}$ using the ratio of the total number of samples seen in each up until time t. As a result, sampling from the distribution that incorporates previous information and then *adding in* this new information effectively "double counts" information that is seen early in the trial, leading directly to a confirmation bias.

The second update rule considers how $p(e_t|C)$ can be directly estimated by marginalizing over x_t. Assuming that both x and C are sampled from their respective posteriors, and taking the limit where $p(C)$ is updated after each sample of x (i.e. that the brain cannot use an internal "buffer" to "remember" and "swap" batches of samples; $S = 1$ below), one finds[1]

$$
\begin{aligned}
p_t(C) &\propto p_{t-1}(C) p(e_t|C) \propto p_{t-1}(C) \int p(e_t|x) p(x|C) dx \\
&\propto p_{t-1}(C) \frac{1}{S} \sum_{x^{(s)} \sim p(e_t|x) p(x|C_{t-1})}^{s=1..S} \frac{p(x^{(s)}|C)}{p(x^{(s)}|C_{t-1})} \\
&\propto p_{t-1}(C) p(x^{(s)}|C)^{\frac{1}{n_s}} \quad \text{for } S = 1
\end{aligned}
$$

where the $x^{(s)}$ are samples from the current posterior over x and n_s is the number of samples taken per stimulus frame (determined by the sampling rate). The exponent results from the fact that e_t is constant for all n_s samples from x[2].

2.4 Differential predictions

While all three models exhibit a confirmation bias as evidenced by their decreasing PK, how this bias depends on the magnitude of the evidence presented (here parameterized by σ_e^2) differs between the BI model and the sampling-based models. In the BI model, larger variance in the evidence leads to shorter bound-crossing times which in turn appears as a steeper PK when averaged over many trials [4] (Figure 2b-d). In the sampling-based model, on the other hand, stronger individual evidence e_t will decrease the influence of the trial-specific prior, $p(x, C|e_1, ..., e_{t-1})$ on the representation of x and C. This will lead to a shallower PK.

[1] note $p(x^{(s)}|C_{t-1})$ cancels in the last step acting merely as a normalization factor if the sum is evaluated using only $S = 1$ sample for $x^{(s)}$.

[2] likely, the sensory sampling rate for the brain is fixed and the brain learns the exponent from experience with the task.

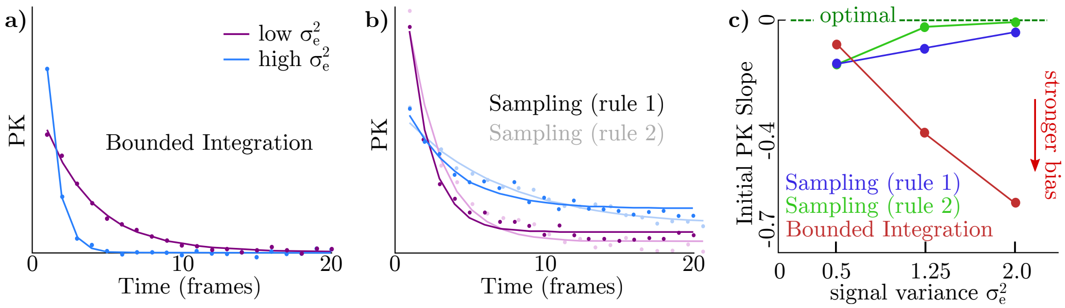

Figure 3: Results from model simulations. Results were fitted with exponential curves. **a)** In the BI model, higher signal variance results in earlier decisions and, therefore, a PK that decreases more quickly. **b)** In the sampling models, in contrast, the PK is steeper for the low signal variance condition. **c)** Summarizing model predictions: the two models make opposite predictions for how the slope of the PK changes with signal variance.

Figure 3a shows the simulation results of an IB model. Figure 3b shows simulation results from a toy-model instantiation of the model in Figure 2e,f where we assumed $p(x|C) = \mathcal{N}(C, 0.2)$, $C = \pm 1$, and $p(e|x) = \mathcal{N}(x, 0.9)$. As expected, BI and sampling-based models make opposing predictions and are therefore distinguishable using empirical data. Figure 3c shows the dependencies of the slope of the PK for the BI and sampling-based models as a function of the signal variance.

The fact that the decreasing PK in [2] was found using weak evidence on each frame (the frames were perceptually indistinguishable) while [6, 5] found a constant PK using stimuli that elicited strong percepts less likely to be influence by prior expectations (auditory clicks and high contrast Gabors, respectively), agrees with our model's predictions.

References

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

[2] 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.

[3] Ralf M. Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *arXiv*, (1409.0257v1), 8 2014.

[4] 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 : the official journal of the Society for Neuroscience*, 28(12):3017–3029, 2008.

[5] 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.

[6] 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.

[7] Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229, 2015.

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

[9] József 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, 3 2010.