# Probing the sources of suboptimality in human Bayesian inference

Luigi Acerbi, Sethu Vijayakumar, Daniel M. Wolpert

**Summary.** In the simple case of a Gaussian distribution of stimuli, humans are often able to combine noisy sensory information with the statistics of the experiment (priors) in agreement with the 'optimal' solution of Bayesian Decision Theory (BDT). However, in the presence of a more complex experimental distribution (e.g. skewed or bimodal), such that the normatively optimal behavior is nonlinear, performance may appear suboptimal even after a long training. It is unclear whether such observed suboptimality arises from a mismatched internal representation of the complex prior, or from limitations in performing probabilistic inference on a correct internal representation of the complex distributions. We tested between these possibilities with a novel estimation task in which subjects were provided explicit probabilistic information in each trial, removing thereby the need for remembering a prior. The task consisted in estimating the location of a hidden target given a noisy cue and a visual representation of the prior probability density over locations, which changed from trial to trial. Priors belonged to different classes of distributions (Gaussian, unimodal, bimodal). Subjects' performance was in qualitative agreement with the predictions of BDT albeit generally suboptimal. However, the degree of suboptimality was mostly independent of the class of the prior and level of noise in the cue, suggesting that remembering the patterns of past events constitutes more of a challenge to decision making than manipulating the complex probabilistic information at hand. We performed an extensive model comparison across a large set of suboptimal Bayesian observer models. Our analysis rejects for our task many common models of variability, such as probability matching and a sample-averaging strategy. Instead we found that subjects' suboptimality was driven both by a miscalibrated internal representation of the parameters of the likelihood, and by decision noise that can be interpreted as a noisy representation of the posterior.

**Results.** Twenty-four subjects took part in a target estimation task (see Figure 1 a & b for description). Subjects performed a training session on Gaussian priors and were tested in a separate session with Gaussian, unimodal or bimodal priors (Figure 1 c). Individual performance in each trial was measured through the *optimality index*, the probability of hitting the target based on the subjects' responses divided by the probability of hitting the target for an optimal responder (Figure 2). The maximal optimality index is 1, for a Bayesian observer with correct internal model of the task and no sensorimotor noise. We found that the relative degree of suboptimality was similar across different priors and different cue noise.

**Model.** In order to find the best explanation for the data, we built a large set of Bayesian observer models that assumed various sources of suboptimality in the decision making process. For example, baseline model $M_0$ follows standard BDT; $M_{lapse}$ considers the possibility of random lapses; $M_{post}$ assumes a noisy representation of the posterior; $M_{sample}$, instead, assumes that the chosen target is obtained through averaging $\kappa$ samples from the posterior; $M_{sample1}$ models a posterior-matching decision strategy ($\kappa = 1$); conversely, $M_{sampleG}$ assumes that the samples are drawn from a Gaussian approximation of the posterior. These and several other observer models were compared by computing the DIC score [1] through sampling from the posterior distribution of the parameters given the individual datasets. In all models the parameters of the likelihood were allowed to differ from the true noise parameters of the task and we added motor noise to the subjects' responses. The most supported model was $M_{post}$, in which the subjects' choice of the optimal target is not deterministic, as per standard BDT, but emerges stochastically from a 'noisy posterior' which yields a target choice distribution approximated by a power function of the posterior. A 'postdiction' for each subjects' performance was obtained through simulating an observer model with parameters drawn from the posterior distribution of the parameters given the data (continuous line in Figure 2). Simulated performance of model $M_{post}$ correlated very well with subjects' individual performance in both training and test sessions ($R^2 = 0.92$; data not shown).

[1] Spiegelhalter, D. J. et al. (2002) Bayesian measures of model complexity and fit. *J R Stat Soc B*, 64.

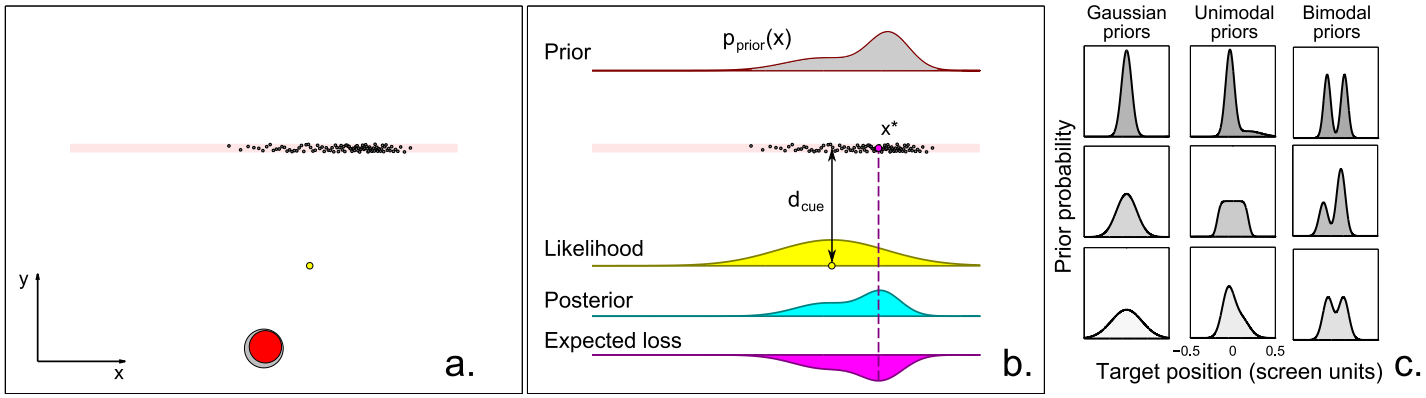


Figure 1: **Experimental procedure. a: Screen setup.** The screen showed a home position (grey circle), the cursor (red circle), a line of potential targets (dots) and a visual cue (yellow dot). The task consisted in locating the true target among the one-dimensional array of potential targets, given the position of the noisy cue. A 'hit' ensued if the cursor encircled the target. **b: Model of the task.** The potential targets constituted a discrete representation of the trial-dependent prior distribution $p_{prior}(x)$. The hidden target $x$ was chosen uniformly at random from the potential targets, and the horizontal position of the cue (yellow dot) was drawn from a Gaussian distribution centered on the true target and whose standard deviation was proportional to the distance $d_{cue}$ from the target line. A Bayesian ideal observer combines the prior distribution with the likelihood to obtain a posterior distribution. The posterior is convolved with the loss function (whether the target will be encircled by the cursor) and the observer picks the 'optimal' target location $x^*$ (purple dot) that corresponds to the minimum of the expected loss (dashed line). **c: Prior distributions.** Some examples of the prior distributions of targets. In each trial the prior was drawn from a class of distributions (Gaussian, unimodal or bimodal), which depended on the session and experimental group. Each class contained eight distinct priors, which were then randomly shifted or flipped along the mean when displayed on screen, so to preserve task symmetry.

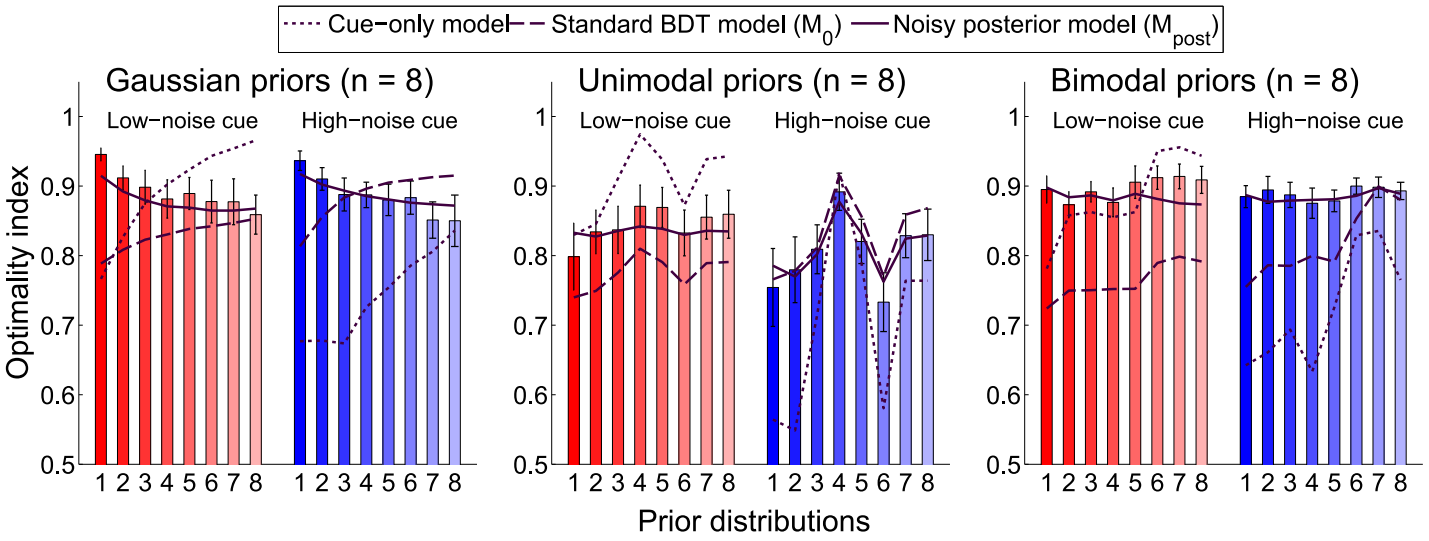


Figure 2: **Optimality index and model 'postdictions'.** Each bar represents the group-averaged optimality index for a specific test session, for each prior (indexed from 1 to 8) and cue type, either low-noise cues (red bars) or high-noise cues (blue bars). Error bars are s.e.m. across subjects. Priors are arranged in the order of differential entropy (i.e. increasing variance for Gaussian priors). The continuous line represents the 'postdiction' of the best suboptimal Bayesian observer model, model $M_{post}$ (noisy posterior). For comparison, the dashed line is the 'postdiction' of the standard BDT observer model $M_0$ and the dotted line represents instead a suboptimal observer that takes into account only the cue.