

# Supplementary Information

## Bayesian Surprise

We compute the Bayesian surprise for each trial under a Beta-Binomial model of Bayesian learning of the stimulus probabilities (Ostwald et al., 2012). The model assumes that the brain implements a sequential learning procedure starting from an uninformative prior, updates this prior according to the subsequent observations, and computes Bayesian surprise as the Kullback-Leibler divergence between prior and posterior (Ostwald et al., 2012). Following Ostwald et al. (2012), we use a variant of this model that assumes an exponential forgetting of stimuli which are observed in the distant past.

Formally, we assume that the probability of observing a low ( $x = 0$ ) or high ( $x = 1$ ) intensity stimulus at a given trial is described by a Bernoulli distribution with parameter  $\mu \in [0, 1]$ :

$$\Pr(x|\mu) = \mu^x (1 - \mu)^{1-x}, \quad x \in \{0, 1\}.$$

The true value of  $\mu$  is unknown by the subject, and the initial uncertainty is modelled by means of an informative prior of type Beta, whose density function is uniform on the unit interval:

$$f(\mu) = 1, \quad \forall \mu \in [0, 1].$$

On each trial, the prior is sequentially updated according to the observed data likelihood to form a posterior distribution over  $\mu$ . After  $N$  trials, let  $\mathbf{x}^N$  denote the sequence of observed stimuli

$$\mathbf{x}^N = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

where  $x_i \in \{0, 1\}$  for each  $i = 1, \dots, N$ . Under standard Bayesian learning, the posterior over  $\mu$  is computed as follows. The probability of observing a  $\mathbf{x}^N$  when the true parameter is  $\mu$  is:

$$\Pr(\mathbf{x}^N|\mu) = \mu^{n_N} (1 - \mu)^{m_N}.$$

Bayes rule implies that the posterior over  $\mu$  is a Beta distribution with density

$$f(\mu|\mathbf{x}^N) = \frac{\Pr(\mathbf{x}^N|\mu) f(\mu)}{\int_0^1 \Pr(\mathbf{x}^N|\nu) f(\nu) d\nu} = \frac{\mu^{n_N} (1 - \mu)^{m_N}}{B(1 + n_N, 1 + m_N)}$$

where  $n_N = |\{i : x_i = 1\}|$  is the number of high-intensity stimuli,  $m_N = N - n_N$  is the number of low-intensity stimuli and  $B$  is the Beta function (Ostwald et al., 2012).

In order to account for a forgetting dynamics, instead of using the accumulative stimulus counts  $n_N$  and  $m_N$  in previous formula, the model employed here weights past observations according to an exponential function. Define the weighted stimulus counts  $n_N^\tau$  and  $m_N^\tau$  by

$$n_N^\tau := \sum_{i=1}^N \exp\left(-\frac{1}{\tau}(N-i)\right) x_i$$

and

$$m_N^\tau := \sum_{i=1}^N \exp(-\tau(N-i))(1-x_i).$$

$\tau \geq 0$  is a parameter governing the forgetting dynamics: for  $\tau = 0$  we have  $n_N^\tau, m_N^\tau = n_N, m_N$ , whereas increasing the value of  $\tau$  implies that past observations are weighted less and less. This results in a posterior density given by

$$f^\tau(\mu|\mathbf{x}^N) = \frac{\mu^{n_N^\tau} (1-\mu)^{m_N^\tau}}{B(1+n_N^\tau, 1+m_N^\tau)}.$$

Finally, the model quantifies the degree of learning as the Bayesian surprise, or Kullback–Leibler divergence, between the prior and posterior distribution over  $\mu$  after a given trial. Let  $\mathbf{x}$  denote the full sample of observed stimuli. The Bayesian surprise after the N-th trial is given by:

$$\begin{aligned} \text{surprise}_N^\tau(\mathbf{x}) &\equiv KL(f^\tau(\cdot|\mathbf{x}^N) \| f^\tau(\mu|\mathbf{x}^{N-1})) \\ &\equiv \int_0^1 f^\tau(\mu|\mathbf{x}^{N-1}) \log \frac{f^\tau(\cdot|\mathbf{x}^{N-1})}{f^\tau(\mu|\mathbf{x}^N)} d\mu. \end{aligned}$$

Due to the use of conjugate priors, the Kullback–Leibler divergence can be evaluated analytically, which significantly simplifies the computation.<sup>1</sup>

## References

- [1] Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, 62(1), 177–188. <https://doi.org/10.1016/j.neuroimage.2012.04.050>

---

<sup>1</sup>In our particular case, it can be shown that

$$\begin{aligned} \text{surprise}_N^\tau(\mathbf{x}) = \log & \frac{B(1+n_N^\tau, 1+m_N^\tau)}{B(1+n_{N-1}^\tau, 1+m_{N-1}^\tau)} + (n_N^\tau - n_{N-1}^\tau + m_N^\tau - m_{N-1}^\tau) \psi(2+n_{N-1}^\tau + m_{N-1}^\tau) \\ & - (n_N^\tau - n_{N-1}^\tau) \psi(1+n_{N-1}^\tau) - (m_N^\tau - m_{N-1}^\tau) \psi(1+m_{N-1}^\tau) \end{aligned}$$

where  $B$  is the Beta function introduced above and  $\psi$  is the Digamma function.