

## Supplementary information S2 (box): Variational free energy

Here, we derive the free-energy and show how its various formulations relate to each other. We start with the quantity we want to bound; namely, the surprise or log-evidence associated with sensory states  $\tilde{s}(t)$  that have been caused by some unknown quantities  $\vartheta \dots \{\tilde{x}, \theta\}$ , which include the hidden states and parameters in box (S1)

$$-\ln p(\tilde{s}(t)) = -\ln \int p(\tilde{s}(t), \vartheta) d\vartheta \quad \text{S2.1}$$

To create a free-energy bound on surprise  $F(\tilde{s}(t), q(\vartheta))$ , we simply add a non-negative cross-entropy between an arbitrary (recognition) density on the causes  $q(\vartheta)$  and their posterior density  $p(\vartheta | \tilde{s})$  (dropping the dependency on  $m$  for clarity).

$$\begin{aligned} F &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta | \tilde{s})} d\vartheta - \ln p(\tilde{s}) \\ &= D(q(\vartheta) || p(\vartheta | \tilde{s})) - \ln p(\tilde{s}) \end{aligned} \quad \text{S2.2}$$

The cross-entropy term is non-negative by Gibb's inequality. In short, free-energy is *cross-entropy plus surprise*. Because surprise depends only on sensory states, we can bring it inside the integral and use  $p(\vartheta, \tilde{s}) = p(\vartheta | \tilde{s})p(\tilde{s})$  to show free-energy is *expected energy minus entropy*

$$\begin{aligned} F &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta | \tilde{s})p(\tilde{s})} d\vartheta \\ &= \int q(\vartheta) \ln q(\vartheta) d\vartheta - \int q(\vartheta) \ln p(\vartheta, \tilde{s}) d\vartheta \\ &= -\langle \ln p(\vartheta, \tilde{s}) \rangle_q - \langle -\ln q(\vartheta) \rangle_q \end{aligned} \quad \text{S2.3}$$

where  $-\ln p(\vartheta, \tilde{s})$  is Gibb's energy. A final rearrangement, using  $p(\vartheta, \tilde{s}) = p(\tilde{s} | \vartheta)p(\vartheta)$ , shows free-energy is also *complexity minus accuracy*, where complexity is the cross-entropy between the recognition  $q(\vartheta)$  and prior density  $p(\vartheta)$

$$\begin{aligned} F &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\tilde{s} | \vartheta)p(\vartheta)} d\vartheta \\ &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta)} d\vartheta - \int q(\vartheta) \ln p(\tilde{s} | \vartheta) d\vartheta \\ &= D(q(\vartheta) || p(\vartheta)) - \langle \ln p(\tilde{s} | \vartheta) \rangle_q \end{aligned} \tag{S2.4}$$