S1 Text: A statistical inference model

Uncertainty-dependent Extinction of Fear Memory in an Amygdala-mPFC Neural Circuit Model

Yuzhe Li¹, Ken Nakae², Shin Ishii² and Honda Naoki³,*
*: Corresponding author

1. Graduate School of Biostudies, Kyoto University, Kyoto, Japan
2. Graduate School of Informatics, Kyoto University, Kyoto, Japan
3. Imaging Platform of Spatio-temporal Information, Graduate School of Medicine, Kyoto University, Kyoto, Japan
**Fear memory as a statistical inference**

We aimed to reveal the type of statistical processing that animals perform during fear conditioning with partial reinforcement and subsequent extinction. The fact that extinction learning largely depends on the uncertainty of the US during fear conditioning (Fig 2I-K) suggests that the uncertainty must be encoded in the brain. Thus, animals may process the statistical properties of sequential US observations.

As statistical processing performed by the animals, we adopted a statistical model based on sequential updating of Bayesian logistic regression (see Statistical inference model below). Because continuous and discontinuous presentation of the US during full and partial reinforcement paradigms leads to fear memories that differ in their resistance to extinction, the effect of US continuity was incorporated into the model; the probability that the US would occur was inferred based on the previous observation of the US (equation (S1)). Then, our simulation demonstrated that, like the basic model, the sequentially predicted US probability reproduced (Fig 2) the characteristics of the extinction of fear memory acquired through both the full and partial reinforcement schedules (S3C Fig and S3D Fig) in a manner consistent with biological fear neurons.

This statistical model also generated several properties that depended on the US probability, as obtained in the neural circuit model. These properties included the time constant of extinction (S4B Fig), the amount of residual fear memory that remained after extinction (S4C Fig), and the degree of surprise to the no-US (S4D Fig), which was quantified as the amount of information (S3E Fig and S3F Fig). Moreover, we compared the measure of surprise in the neural circuit model (the learning signals) with the measure of surprise in the statistical model (amount of information) when the same US pattern was applied, and we found a high degree of correlation, providing additional evidence that the learning signal in the neural circuit model represents the degree of surprise in terms of statistics (S4E Fig and S4F Fig). Taken together, these commonalities between the two models suggest that the neural circuit model that consisted of fear, persistent and extinction neurons effectively processed the statistical property of the occurrence of the US through sequential updating of Bayesian logistic regression.
**Statistical inference model**

To quantify the degree of surprise from the perspective of inferring whether the US would occur, we developed a statistical model based on logistic regression with sequential Bayesian updating [1]. We assumed that the animals inferred the probability that the US would occur based on the previous observation of the US as

\[
\pi_t = \sigma(w_{i0} + w_{i1} US_{t-1}),
\]

(S1)

and sequentially updated \( w_{i+1} = [w_{i+1,0}, w_{i+1,1}] \) to predict the probability based on a new instance of the US, where \( US_t \) denotes a binary variable, i.e., \( US_t = 0 \) and \( = 1 \) represent the presence and absence of the US, respectively; \( w_{i0} \) and \( w_{i1} \) are internal variables in animals, representing the effects of inferring the probability that the US would occur independent of and dependent on the previous US, respectively; \( \sigma(\cdot) \) is the logistic function \( \sigma(x) = 1/(1+ e^{-x}) \); \( \pi_t \) represents the probability of the US occurrence that the animals infer based on the previous US, \( US_{t-1} \). We also assumed that the degree of surprise \( S \) that animals feel in response to the US or no-US is quantified by the amount of information [2]:

\[
S(\text{US}) = -\log \pi_t,
\]

\[
S(\text{no-US}) = -\log(1 - \pi_t).
\]

(S2)

We considered \( w_{i+1} \) to be updated by sequential Bayesian updating as

\[
P(w_i | US_{t-1}) \propto P(US_t | w_i, US_{t-1}) P(w_i | US_{t-1})
\]

\[
= P(US_t | w_i, US_{t-1}) \int P(w_i | \mathbf{w}_{t-1}) P(w_{t-1} | US_{t-1}) d\mathbf{w}_{t-1},
\]

(S3)

where \( P(w_i | US_{t-1}) \) and \( P(w_i | US_{t-1}) \) represent the prior and posterior distributions, respectively. The animals were assumed to think that instances of the US were simply generated by a Bernoulli process as

\[
P(US_t | w_i, US_{t-1}) = \pi_t^{US_t}(1-\pi_t)^{1-US_t}.
\]

(S4)

In addition, the animals were assumed to believe that \( \mathbf{w} \) remained almost constant with a small degree of noise as \( \mathbf{w} = \mathbf{w}_{i-1} + \mathbf{e}_i \), where \( \mathbf{e}_i \) represents independent and identically distributed Gaussian noise with a mean of zero and a low variance, \( \sigma \). Then,

\[
P(w_i | \mathbf{w}_{i-1}) = N(w_i | \mathbf{w}_{i-1}, \sigma \mathbf{I}),
\]

(S5)

where \( \mathbf{I} \) represents the unit matrix.

The Bayesian belief update was implemented as an extended Kalman filter, in which the prior and posterior distributions are approximated by Gaussian distributions. \( P(w_i | US_{t-1}) \) was represented by

\[
P(w_i | US_{t-1}) = N(w_i | \mu_{t-1}, \Sigma_{t-1}),
\]

(S6)

where \( N \) represents the function of the Gaussian distribution parameterized by \( \Sigma \) and \( \mu \), which are the variance-covariance matrix and mean vector of \( w_i \), respectively. Then, equation (S3) can be transformed into

\[
P(w_i | US_{t-1}) \propto P(US_t | w_i, US_{t-1}) N(w_i | \mu_{t-1}, \sigma + \Sigma_{t-1}).
\]

(S7)

Note that the prior and posterior distributions in equation (S7) do not have conjugate relationship. Thus, the posterior distribution, \( P(w_i | US_{t-1}) \), was approximated by a Gaussian distribution using
Laplace approximation as
\[ P(w_t | U_{S_i}) = N(w_t | \mu_t, \Sigma_t) , \]
where \( \mu_t \) and \( \Sigma_t \) are updated as
\[ \mu_t = \arg \max_{w} E(w_t) , \]
\[ \Sigma_t^{-1} = -\nabla E(w_t) = (s + \Sigma_{t-1})^{-1} + \sigma(\mu_{t-1}^T \phi_{t-1}) (1 - \sigma(\mu_{t-1}^T \phi_{t-1})) \phi_{t-1} \phi_{t-1}^T , \]
where \( \phi \) indicates \((1, US_t)^T\)

\[ E(w_t) = \frac{1}{2} (w_t - \mu_{t-1})^T (s + \Sigma_{t-1}) (w_t - \mu_{t-1}) \]
\[ + US_t \log \sigma(w_t^T \phi_{t-1}) + (1 - US_t) \log(1 - \sigma(w_t^T \phi_{t-1})) + \text{const} \]

To calculate \( \mu_t \) requires numerical optimization because \( E(w_t) \) is a non-linear function of \( w_t \).

References
1. Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Dietterich TG, Becker S, Ghahramani Z, editors. Advances in neural information processing systems 14, Vol. 2. Cambridge, MA: MIT Press; 2002. pp. 841-848.
2. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27: 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x