Brain mechanism of foraging: reward-dependent synaptic plasticity or neural integration of values?

Ulises Pereira-Obilinovic[1], **Han Hou**[2], **Karel Svoboda**[2], **and Xiao-Jing Wang**[1,*]

[1]Center for Neural Science, New York University, New York, NY 10003, USA; [2]Allen Institute for Neural Dynamics, Seattle, WA 98109, USA; [*]xjwang@nyu.edu

This manuscript was compiled on October 6, 2022

In foraging, behavior is guided by action values that are stored in memory and updated depending on the history of choices and rewards. What is the neural mechanism for action value maintenance and reward-dependent update? Here we explore two contrasting network models: neural integration and synaptic learning of action value. We show that both mechanisms reproduce existing experimental data, but they depend on distinct and falsifiable assumptions about the biological neural circuits. We demonstrate neural dynamical signatures and perturbations that could be carried out experimentally to differentiate the two contrasting scenarios.

Foraging | Value-based decision-making | Reinforcement learning | Neural integrator

During foraging, action values are stored in short-term memory over times of several behavioral trials, much longer than intrinsic neural time scales. The stored action values bias actions to maximize reward. In the cortex, single-neuron activity of the monkey posterior parietal cortex (1, 2), the mice medial prefrontal (mPFC) (3) and retrosplenial cortex (4), is correlated with estimated action values. Such value coding has traditionally been explained by reward-dependent synaptic plasticity (5, 6). However, recent findings of graded neural activity in proportion to newly updated action values during the inter-trial intervals (3, 4) have raised an alternative account by virtue of neural integration (7). In this work, we investigated these two contrasting scenarios (Fig. 1A, B).

According to the neural integrator model (Fig. 1A, top row), action values are stored in directions of slow change in population activity or neural integrator. The neural integrator integrates reward events by updating action values at each trial by an input proportional to the reward prediction error (RPE) in the direction of action value encoding (see Fig. 1A top row). Importantly, in this mechanism, changes in the synaptic efficacy are not needed for action value update, similar to a theoretical proposal of mPFC as a meta-reinforcement learning system (8). On the other hand, according to the synaptic model (Fig. 1A, bottom row), dopamine release triggered by reward enables Hebbian plasticity on the recurrent connections of the valuation network, in turn modifying in a reward-dependent manner the synaptic efficacy on the network recurrent connections (5, 6). Since synaptic inputs drive spiking activity, action values modulate neuronal activity.

Results

Foraging dynamics. We explore these two contrasting hypotheses for the neural mechanisms underlying action value maintenance and its reward-dependent update during foraging. We develop two experimentally testable circuit models. These models were guided by anatomical and physiological evidence of the neural circuits involved in foraging behavior. We aim for these models to be simple enough to be suitable for mechanistic understanding and mathematical analysis, but to include key details to the extent their predictions can be interpreted in actual brain circuits.

Here we focus in the dynamic foraging task (DFT) (2–4, 9, 10) (see SI).

In our models, the network architecture consists of a mesoscopic network of two inter-connected areas, modeled as recurrent networks (Fig. 1A, second column). Action values for the two contingencies are maintained and updated depending on reward, in a network we refer to as a valuation network. Action is selected depending on action values in what we refer to as the action selection network. This architecture is inspired by the separation between the limbic and motor streams of information during decision-making (11).

In our neural integrator model, action values for the two contingencies are maintained by the population activity in a two-dimensional manifold embedded in a high-dimensional neural space (the space spanned by the activity of all neurons in the network) (Fig. 1A, first row). After the action is selected and the reward is delivered, an RPE signal updates the corresponding action value represented in the action value manifold (Fig. 1B, first row). This manifold corresponds to overlapping populations of neurons. Unlike classic neural integrators where noise gets integrated, in our model, random perturbations hardly produce significant changes in the encoding of action values. This is because, in high dimensions, random perturbations are, on average orthogonal to the action value encoding directions (see SI). However, this comes with a price to pay. Namely, our model predicts that RPE signals should be aligned to the action value encoding axis corresponding to the chosen contingency. Therefore, such alignment should be contingency-dependent, with RPE signals projecting to different action value axes depending on the selected action (Fig. 1A, top row). This model predicts different RPE-correlated activity patterns in the valuation network for different selected contingencies.

In the synaptic model, feedback from the action selection network carrying choice information will transiently elevate the activity of the corresponding selective populations in the valuation network after the action is selected (Fig. 1B, bottom row). In contrast, little to no feedback activity is produced in the corresponding unchosen population. After reward delivery, the transient elevated activity in the corresponding chosen populations will drive Hebbian learning gated by a global

U.P.-O., K.S., and X.-J.W. designed research; U.P.-O. performed the research; H.H. performed the behavioral experiments; H.H. fitted the behavioral models with the help of U.P.-O.; U.P.-O. wrote the paper with the contributions of all authors.

The authors declare no competing interest.

[2]To whom correspondence should be addressed. E-mail: xjwang@nyu.edu

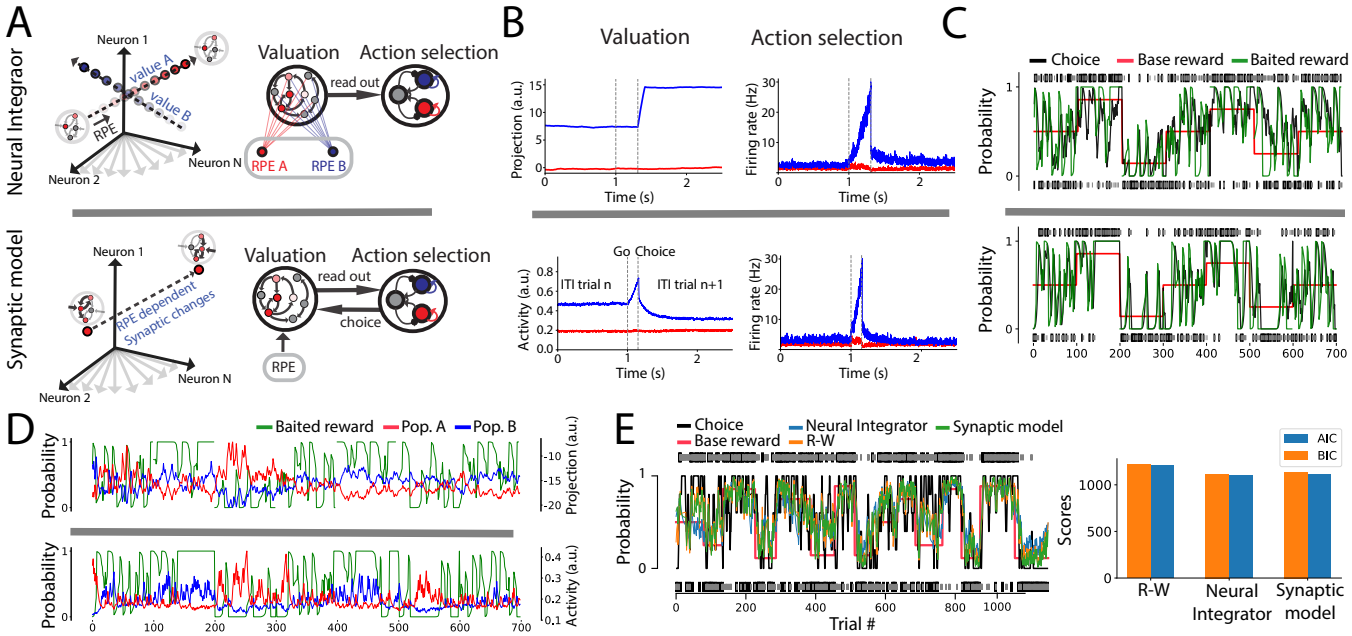

Fig. 1. Foraging behavior: Neural integrator vs. synaptic model. (A) Dynamical mechanism during foraging (first column) and network architecture (second column). (B) Single-trial network dynamics. (C) Foraging behavior during a DFT session. (D) Baited reward (green) and population activity averaged during the ITI for an example DFT session. (A-D) Top row: Neural integrator; Bottom row: Synaptic model. (E) Left: Foraging behavior of a mouse during an example session fitted using a Rescorla–Wagner model (R-W) (see SI), Neural integrator, and a synaptic model. Right: Model fitting AIC and BIC scores. Here the contingencies are licking left (A) and right (B).

RPE signal using a three-factor learning rule (5, 6) (see the bottom row in Fig. 1A). Thus, only the recurrent synapses of the corresponding chosen populations will modify their synaptic efficacy. Importantly, these selective feedback signals after action selection are not observed in the neural integrator (Fig. 1B).

During foraging, both models display value-dependent persistent activity during the inter-trial interval (ITI) (Fig. 1B) (3, 4). After the action selection network makes a choice, the activity is modified in both models, and action value representations are persistent during the next ITI (3, 4). Both models display matching behavior (2–4, 10) (Fig. 1C). Strikingly, as it has been observed in the cortex (3, 4), in both models, the valuation network presents population activity correlated with baited reward (Fig. 1D).

A non-selective multiplicative go signal to the action selection network is needed for stable foraging in a broad range of reward probabilities for both models (12). This signal is consistent with a non-selective input to the thalamus in a recurrent connected two-network action selection model involving the cortex and thalamus (13) (see SI). Our model predicts that the ITI activity in the action selection network also presents, a rather small, action value correlated persistent firing.

We fit both network models using mouse behavior during a dynamic foraging task (see SI). We found that both models have comparable performance with standard behavioral models for foraging (3, 4, 14).

Response to perturbations. A central difference between these models is their response to perturbations.

When a perturbation targets action value encoding populations in the valuation network during the ITI the models behave qualitatively differently. In the case of the synaptic model, because action values are stored in the synaptic efficacies, disruptions in the action value encoding due to selective perturbations always rebound to previous values (see Fig. 2A). By contrast, in the case of the neural integrator, perturbations of population activity during the ITI can modify the action value encoding (Fig. 2B). Importantly, however, the effect highly depends on the way of perturbation. While on-manifold perturbation dramatically modifies action value encoding, random perturbations have little effect because, in large networks, random perturbations independently distributed across neurons are orthogonal (on average) to the action value manifold (Fig. 2C). On the other hand, uniform (spatially correlated) perturbations where all neurons are all perturbed by the same constant amount, reminiscent of optogenetic manipulations, could also disrupt action value encoding as on-manifold perturbations if they are strong enough. In the neural integrator, perturbations can be further engineered by ranking neurons by the strength of their encoding to one of the two contingencies. We found that by perturbing the first 10% of the highest action value encoding neurons produce a 83% of change in action value encoding (see Fig. 2D).

In the synaptic model, the timescale of response to random fluctuations of the selective populations in the valuation network is correlated with action values, whereas in the neural integrator it is not. The reason is that in the synaptic model, an increase (decrease) in the strength of excitatory recurrent connections increases (decreases) network reverberations and slows down (speeds up) the network's dynamics (see SI). In contrast, for the neural integrator, action value updates do not necessarily require synaptic efficacy changes.

Discussion

In light of recent experimental findings, we considered two contrasting scenarios for action value estimation and updat-

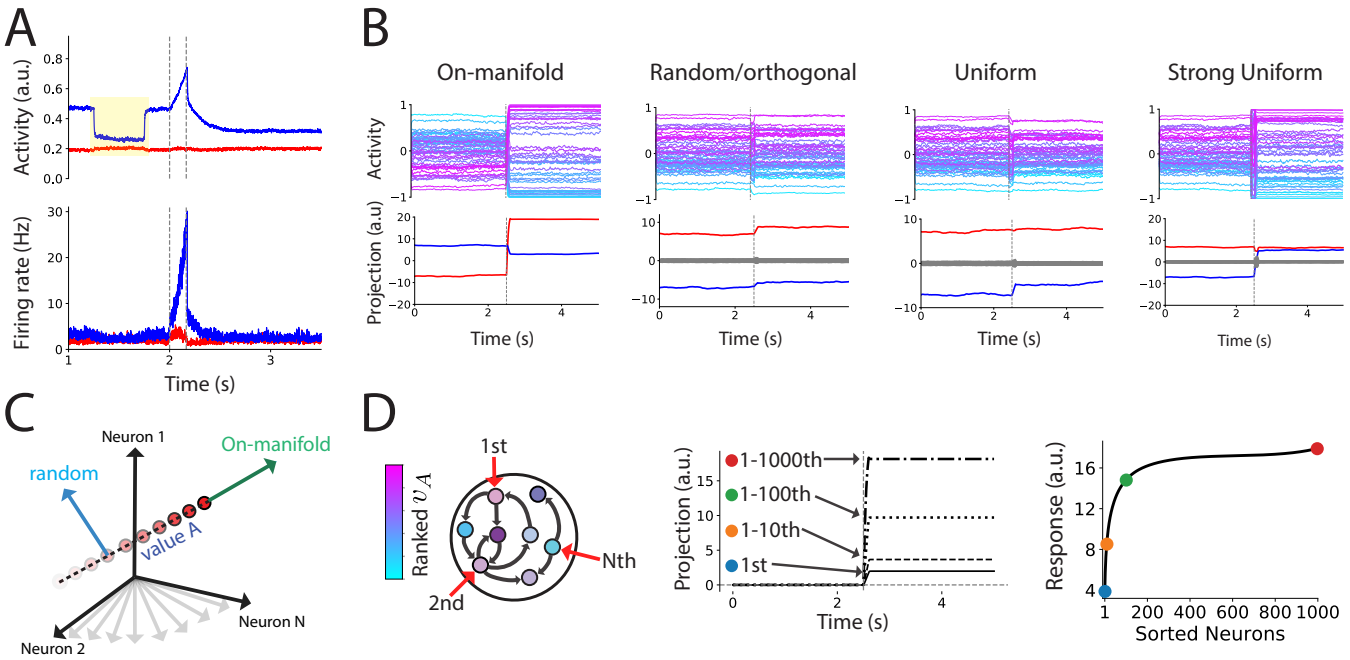

Fig. 2. Response to perturbations: Neural integrator vs. synaptic model. (A) Perturbation during the ITI (yellow shaded region) in the valuation network for the synaptic model. (B) Examples of on-manifold, random/orthogonal, and uniform perturbations (at the vertical dashed lines) in the valuation network for the neural integrator. (C) Schematic of on-manifold and random/orthogonal perturbations. (D) First column: schematics for the engineered perturbation experiment for the neural integrator. Neurons are ranked according to the entries of the action value encoding axis $\vec{v}_A$. Second column: change in action value encoding when the 1st, 1-10th, 1-100th, and all ranked neurons are stimulated with the same magnitude. Third column: number of stimulated neurons vs. the magnitude of the perturbation response (see SI).

ing underlying foraging behavior. For the synapse model, we considered a valuation circuit where the recurrent connections undergo reward modulated plasticity, in contrast to most previous models where plasticity occurs only at feedforward input connections. Because of the trial-to-trial changes of recurrent connection strengths, neural activity during inter-trial intervals varies in proportion to action values, consistent with the new experimental observations (3, 4). For the neural integrator model, we assumed a line attractor encoding an action value. Crucially, when embedded in a high dimensional space of neural population activity, such a line attractor becomes robust against random noise. We showed that this model can account for graded changes of neural firing during inter-trial intervals, but requires that value updating is realized by a action-specific reward prediction error signal (mediated by dopamine cells) (15) projecting to value-coding neurons in a special way along a specific action value axis (e.g., for action A, not B), in a given trial. The latter implies that dopamine projection targets selective neural populations in a choice-dependent manner, an assumption not supported by presently known evidence but testable in the future. We demonstrate several other experimentally feasible ways to differentiate the two scenarios, especially by optogenetic perturbations.

Materials and Methods

For the methods, see the SI.

ACKNOWLEDGMENTS. We thank Nai-Wen Tien and Marton Rozsa for their help with the behavioral training in the pilot experiments. We thank Aldo Battista and Jorge Jaramillo for their constructive comments on the manuscript. This work was supported by The Swartz Foundation (to U.P.-O.), James Simons Foundation Grant 543057SPI (to X.-J.W.), National Institutes of Health grant R01MH062349 (to X.-J.W.), Janelia Visiting Scientists Program (to X.-J.W. and U.P.-O.), Howard Hughes Medical Institute (to K.S.), National Institute of Health U-19 grant 1U19NS123714-01 (to H.H. and K.S.), the Simons Collaboration on the Global Brain (to H.H., K.S., and X.-J.W.), and The Allen Institute for Neural Dynamics (to H.H. and K.S.).

1. ML Platt, PW Glimcher, Neural correlates of decision variables in parietal cortex. *Nature.* **400**, 233–238 (1999).
2. LP Sugrue, GS Corrado, WT Newsome, Matching behavior and the representation of value in the parietal cortex. *science* **304**, 1782–1787 (2004).
3. BA Bari, et al., Stable representations of decision variables for flexible behavior. *Neuron* **103**, 922–933 (2019).
4. R Hattori, B Danskin, Z Babic, N Mlynaryk, T Komiyama, Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872 (2019).
5. A Soltani, XJ Wang, A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J. Neurosci.* **26**, 3731–3744 (2006).
6. Y Loewenstein, HS Seung, Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proc. Natl. Acad. Sci.* **103**, 15224–15229 (2006).
7. HS Seung, How the brain keeps the eyes still. *Proc. Natl. Acad. Sci.* **93**, 13339–13344 (1996).
8. JX Wang, et al., Prefrontal cortex as a meta-reinforcement learning system. *Nat. neuroscience* **21**, 860–868 (2018).
9. RJ Herrnstein, Relative and absolute strength of response as a function of frequency of reinforcement. *J. experimental analysis behavior* **4**, 267 (1961).
10. B Lau, PW Glimcher, Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. experimental analysis behavior* **84**, 555–579 (2005).
11. J Cox, IB Witten, Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* **20**, 482–494 (2019).
12. HK Inagaki, et al., A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell* **185**, 1065–1081 (2022).
13. ZV Guo, et al., Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545**, 181–186 (2017).
14. GS Corrado, LP Sugrue, HS Seung, WT Newsome, Linear-nonlinear-poisson models of primate choice dynamics. *J. experimental analysis behavior* **84**, 581–617 (2005).
15. B Engelhard, et al., Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature* **570**, 509–513 (2019).

Supplementary Information for

Brain mechanism of foraging: reward-dependent synaptic plasticity or neural integration of values?

Ulises Pereira-Obilinovic, Han Hou, Karel Svoboda, Xiao-Jing Wang

Corresponding Author name.
E-mail: xjwang@nyu.edu

This PDF file includes:

Supplementary text
SI References

Supporting Information Text

Dynamic foraging task

We implement a dynamic foraging task (DFT) similar to the one in (1, 2) in both our network simulations and behavioral experiments. Briefly, in this task, after a go cue, the subject (in our case, a network model and the mice) choose freely among two alternatives that deliver reward probabilistically. Then, after an inter-trial interval (ITI) of the order of several seconds (3s in the case of the network and 5.79s median ITI in the case of the mice), the next go cue (auditory in the case of the mice) is presented. Then the subject chooses (by licking one of two ports in the case of the mice) either left or right for a reward. Depending on their choice, a reward is delivered with a probability that changes randomly in blocks of 100 trials (see Fig. 1C). For this task, the reward schedule uses baited rewards (e.g., see (1–4)), that is that once the reward is assigned, it remains in the corresponding contingency (network)/port (mice) for its consumption.

Network models

Network architecture. As described in the main text, we develop two network models for contrasting two different mechanisms for action value maintenance and update during foraging: (i) neural integration and (ii) synaptic learning. Both models have similar network architecture consisting of a mesoscopic network of two inter-connected areas. Each area is modeled as a recurrent network. Action values are maintained and updated depending on reward in a network we refer to as the valuation network (see Fig. 1A second column). For action selection, action values are read out from the valuation network and projected for biasing actions in a network we refer to as the action selection network (see Fig. 1A second column). The dynamics of the action selection network is the same for both models.

As described below, two critical differences exist between the neural integrator and the synaptic model architectures. First, there are feedback projections between the action selection and the valuation network in the synaptic model, while in the neural integrator, these projections do not exist. Second, in the neural integrator, reward prediction error (RPE) signals are selectively projected to the two action value axis, while in the synaptic model, RPE is a global signal projected to the whole valuation network.

Action selection network. For the action selection network, we use a reduced *mean field model* that has been shown to be a good approximation of the dynamics of a full recurrent network of spiking neurons (5). This reduced model represents the dynamics of the fraction of activated N-methyl-D-aspartate (NMDA) receptors of two selective excitatory populations. These populations represent the two contingencies. The dynamics is described by the below equations

$$\frac{ds_A}{dt} = -\frac{s_A}{\tau} + \gamma(1 - s_A)F(I_A) \quad [1]$$

$$\frac{ds_B}{dt} = -\frac{s_B}{\tau} + \gamma(1 - s_B)F(I_B). \quad [2]$$

with $\gamma = 0.641$ and $\tau = 100$ms. The function F is the input-output transfer function given by

$$f(I) = \frac{A_f I - B_f}{1 - \exp(-D_f(A_f I - B_f))}, \quad [3]$$

where $A_f = 270$Hz/nA, $B_f = 108$Hz, and $D_f = 0.154$s (5, 6).

The synaptic currents are given by a recurrent (I_A^r and I_B^r) , a background synaptic noise (I_A^n and I_B^n),and a non-selective external go signal (I_A^{go} and I_B^{go}) components

$$I_A = I_A^r + I_A^n + I_A^{\text{go}} \quad [4]$$

$$I_B = I_B^r + I_B^n + I_B^{\text{go}}. \quad [5]$$

The recurrent component is given by

$$I_A^r = Ws_A - Js_B + I_0 + j^{\text{FF}} x_A(t) \quad [6]$$

$$I_B^r = Ws_B - Js_A + I_0 + j^{\text{FF}} x_B(t). \quad [7]$$

Here $W = 0.2609$ corresponds to the average recurrent synaptic weights for each selective population, $J = 0.0497$ corresponds to the effective shared inhibition weights, and $I_0 = 0.3381$ to the background current. The parameter j^{FF} corresponds to the feedforward connections from the valuation to the action selection network. The variables $x_A(t)$ and $x_B(t)$ represent the mean activity of the two selective populations $r_A(t)$ and $r_B(t)$ in the case of the synaptic model and the two projections $m_A(t)$ and $m_B(t)$ in the case of the neural integrator respectively (see following sections for the detailed implementation of the valuation network). For the synaptic model $j^{\text{FF}} = 0.2$ for Fig. 1B and $j^{\text{FF}} = 0.04$ for Fig. 1C,D. For the neural integrator $j^{\text{FF}} = 0.0015$ for Fig. 1B and $j^{\text{FF}} = 0.0008$ for Fig. 1C,D. The equations for the background synaptic noise due to AMPA synapses are given by

$$\tau_{\text{AMPA}} \frac{dI_A^n}{dt} = -I_A^n + \sigma_{AMPA}\sqrt{\tau_{AMPA}}\eta_A(t) \quad [8]$$

$$\tau_{\text{AMPA}} \frac{dI_B^n}{dt} = -I_B^n + \sigma_{AMPA}\sqrt{\tau_{AMPA}}\eta_B(t) \quad [9]$$

Here $\eta_A(t)$ and $\eta_B(t)$ are Gaussian noise variables. The time scale of AMPA synapses is $\tau_{\text{AMPA}} = 2$ms. Lastly, the intensity of the noise for the synaptic model is $\sigma_{AMPA} = 0.03$ for Fig. 1B-D and for the neural integrator $\sigma_{AMPA} = 0.03$ for Fig. 1B and $\sigma_{AMPA} = 0.05$ for Fig. 1C,D.

In our model, the go input currents I_A^{go} and I_B^{go} correspond to a transient non-selective inputs that lasts until one of the populations reaches a set firing rate ($F(I_A)$ or $F(I_B)$) threshold of 30Hz. Consistent with recordings in the Anterior Lateral Motor cortex (ALM) during motor initiation (7) and as proposed in theoretical studies (8, 9), we hypothesize transient non-selective go signals gate actions to the motor thalamus. Since the recurrent thalamocortical loop involving ALM and ventromedial (VM) thalamic nucleus is causally involved in motor initiation (10), the net effect of an input to the thalamus corresponds to transient effective synaptic weights (8, 9) given by

$$I^{\text{go}}(t)_A = w^{\text{go}}(t)s_A \quad [10]$$

$$I^{\text{go}}(t)_B = w^{\text{go}}(t)s_B, \quad [11]$$

where

$$w^{\text{go}}(t) = \begin{cases} 0 & \text{t} < t_{\text{go}} \\ 0.1 & t_{\text{go}} \leq \text{t} \leq t^* \\ 0 & t^* < t \end{cases}. \quad [12]$$

We define t^* as the time when one of the populations firing rates $F(I_A)$ or $F(I_B)$ reaches the threshold of 30Hz.

As an alternative model, we also implemented an additive go cue

$$I^{\text{go}}(t) = \begin{cases} 0 & \text{t} < t_{\text{go}} \\ I^{\text{go}} & t_{\text{go}} \leq \text{t} \leq t^* \\ 0 & t^* < t \end{cases}, \quad [13]$$

and simulate the network for several values of I^{go}. The numerical integration time-step used was $dt = 0.1$ms.

Synaptic model.

Network dynamics. For the synaptic model, the network dynamics are given by the linearized dynamics of a network of two excitatory selective populations to the two contingencies in the DFT with shared inhibition

$$\tau \frac{dr_A}{dt} = -r_A + w_A(k)r_A - \frac{\rho}{2}(r_A + r_B) + I_A + I_{fb,A} \quad [14]$$

$$\tau \frac{dr_B}{dt} = -r_B + w_B(k)r_B - \frac{\rho}{2}(r_A + r_B) + I_B + I_{fb,A}. \quad [15]$$

Here the variables r_A and r_B correspond to the mean firing rates of the two selective populations. The parameters $w_A(k)$ and $w_B(k)$ correspond to the mean synaptic efficacy of the recurrent connections in the trial k. The synaptic efficacies are plastic and change on a trial-to-trial basis according to a learning rule specified in the next section. The parameter $\rho = 0.05$ corresponds to the effective, shared inhibition. The background input current is given by the parameters $I_A = I_B = 0.1$. The timescale of the dynamics is taken as $\tau = 10$ms. The numerical integration time-step used was $dt = 10$ms. The currents $I_{fb,A} = w_{fb}s_A$ and $I_{fb,B} = w_{fb}s_B$ correspond to the feedback current from the action selection network. We use $w_{fb} = 0.2$. Since during the ITI the feedback current is small, for the below analysis, we will assume $I_{fb,A} = I_{fb,B} = 0$.

Although extremely simple, this model captures essential features of an E-I network and has the advantage that several quantities can be calculated analytically.

This model has a single fixed-point attractor that corresponds to the stationary state of the dynamics. For calculating its fixed-point, let us first define the following trial-dependent matrix

$$J(k) = \begin{pmatrix} w_A(k) - 1 - \frac{\rho}{2} & -\frac{\rho}{2} \\ -\frac{\rho}{2} & w_B(k) - 1 - \frac{\rho}{2} \end{pmatrix}. \quad [16]$$

We have that Eqs. (14,15) can be written in a vectorial form

$$\dot{\vec{r}} = J(k)\vec{r} + \vec{I}. \quad [17]$$

The fixed-point of Eq. (17) is given by

$$\vec{r}^*(k) = -J^{-1}(k)\vec{I}. \tag{18}$$

The inverse is given by

$$J^{-1}(k) = \frac{1}{w_A(k)w_B(k) - (1+\frac{\rho}{2})(w_A(k)+w_B(k)+1+\rho} \begin{pmatrix} w_B(k) - 1 - \frac{\rho}{2} & \frac{\rho}{2} \\ \frac{\rho}{2} & w_A(k) - 1 - \frac{\rho}{2} \end{pmatrix}. \tag{19}$$

Then trial-dependent fixed-point is given by

$$r_A^*(k) = \frac{(1+\frac{\rho}{2} - w_B(k))I_A - \frac{\rho}{2}I_B}{w_A(k)w_B(k) - (1+\frac{\rho}{2})(w_A(k)+w_B(k)) + 1 + \rho} \tag{20}$$

$$r_B^*(k) = \frac{(1+\frac{\rho}{2} - w_A(k))I_B - \frac{\rho}{2}I_A}{w_A(k)w_B(k) - (1+\frac{\rho}{2})(w_A(k)+w_B(k)) + 1 + \rho}. \tag{21}$$

If $I_A = I_B = I_0$, then the fixed-points in Eqs. (20, 21) become

$$r_A^*(k) = \frac{I_0\left[1 - w_B(k)\right]}{w_A(k)w_B(k) - (1+\frac{\rho}{2})(w_A(k)+w_B(k)) + 1 + \rho} \tag{22}$$

$$r_B^*(k) = \frac{I_0\left[1 - w_A(k)\right]}{w_A(k)w_B(k) - (1+\frac{\rho}{2})(w_A(k)+w_B(k)) + 1 + \rho}. \tag{23}$$

The above equations show that the stationary firing rate depends on the recurrent synaptic efficacies $w_A(k)$ and $w_B(k)$. In our model, these quantities will change in a trial-by-trial by a reward-dependent learning rule that will be specified in the next section.

For the calculation of the autocorrelation function, we first need to calculate the eigenvalues and eigenvectors of the matrix in Eq. (16). The eigenvalues are given by

$$\lambda_+ = \frac{1}{2}\left[w_A(k) + w_B(k) - 2 - \rho + \sqrt{\left(w_A(k) - w_B(k)\right)^2 + \rho^2}\right] \tag{24}$$

$$\lambda_- = \frac{1}{2}\left[w_A(k) + w_B(k) - 2 - \rho - \sqrt{\left(w_A(k) - w_B(k)\right)^2 + \rho^2}\right]. \tag{25}$$

and the eigenvectors are

$$\vec{v}_+ = \begin{bmatrix} \frac{w_B(k) - w_A(k) - \sqrt{(w_A(k)-w_B(k))^2+\rho^2}}{\rho^2 + \left(\sqrt{(w_A(k)-w_B(k))^2+\rho^2} + w_A(k) - w_B(k)\right)^2} \\ \frac{\rho^2}{\rho^2 + \left(\sqrt{(w_A(k)-w_B(k))^2+\rho^2} + w_A(k) - w_B(k)\right)^2} \end{bmatrix} \tag{26}$$

and

$$\vec{v}_- = \begin{bmatrix} \frac{w_B(k) - w_A(k) + \sqrt{(w_A(k)-w_B(k))^2+\rho^2}}{\rho^2 + \left(\sqrt{(w_A(k)-w_B(k))^2+\rho^2} + w_B(k) - w_A(k)\right)^2} \\ \frac{\rho^2}{\rho^2 + \left(\sqrt{(w_A(k)-w_B(k))^2+\rho^2} + w_B(k) - w_A(k)\right)^2} \end{bmatrix}. \tag{27}$$

Given a initial condition $\vec{r}_0$, the solution to the network dynamics in Eq. (14) is given by

$$\vec{r}(t) = [\vec{v}_+, \vec{v}_-] \begin{pmatrix} e^{\lambda_+ t} & 0 \\ 0 & e^{\lambda_- t} \end{pmatrix} \begin{bmatrix} \vec{v}_+ \\ \vec{v}_- \end{bmatrix} \vec{r}_0. \tag{28}$$

Using the above expression and the definition of autocorrelation and autocovariance functions (11), and defining the following coefficients

$$R_+^A = \frac{\left(w_A(k) - w_B(k) + \sqrt{(w_A(k) - w_B(k))^2 + \rho^2}\right)^2}{\rho^2 + \left(\sqrt{(w_A(k) - w_B(k))^2 + \rho^2} + w_A(k) - w_B(k)\right)^2} \quad [29]$$

$$R_-^A = \frac{\left(w_B(k) - w_A(k) + \sqrt{(w_A(k) - w_B(k))^2 + \rho^2}\right)^2}{\rho^2 + \left(\sqrt{(w_A(k) - w_B(k))^2 + \rho^2} + w_B(k) - w_A(k)\right)^2} \quad [30]$$

$$R_+^B = 1 - C_+^A \quad [31]$$

$$R_-^B = 1 - C_-^A, \quad [32]$$

we find that the autocovariance functions are given by

$$C_A(t') = \left(\frac{R_+^A}{|\lambda_+|}\right) e^{\lambda_+ t'} + \left(\frac{R_-^A}{|\lambda_-|}\right) e^{\lambda_- t'} \quad [33]$$

$$C_B(t') = \left(\frac{R_+^B}{|\lambda_+|}\right) e^{\lambda_+ t'} + \left(\frac{R_-^B}{|\lambda_-|}\right) e^{\lambda_- t'}. \quad [34]$$

Lastly, the autocorrelation functions are given by

$$ACF_A(t') = \frac{C_A(t')}{C_A(0)} \quad [35]$$

$$ACF_B(t') = \frac{C_B(t')}{C_B(0)}. \quad [36]$$

Notice that the autocorrelation functions for both populations $ACF_A(t')$ and $ACF_B(t')$ implicitly depend on the synaptic efficacies $w_A(k)$ and $w_B(k)$.

Reward-dependent learning rule. In the synaptic model, the synaptic efficacies of the recurrent connections of the two selective populations in the valuation network $w_A(k)$ and $w_B(k)$ are learned using a reward-dependent plasticity rule (12–14) in a trial-by-trial (where k is the trial index) basis. In our three-factor learning rule the synaptic changes are given by

$$\Delta W_x = W_x(k+1) - W_x(k) = (R(k) - \bar{R}_x(k)) f(r_x)^2 \qquad x = A, B. \quad [37]$$

Here the reward is a binary variable $R \in \{0, 1\}$. The variable $\bar{R}_x$ is the predicted reward specific to the chosen contingency (15) which is estimated by the following running trial average (15, 16)

$$\bar{R}_A(k+1) = R(k) + \frac{(R(k) - \bar{R}_A(k))}{\tau_R} \qquad \text{if A is chosen} \quad [38]$$

$$\bar{R}_B(k+1) = R(k) + \frac{(R(k) - \bar{R}_B(k))}{\tau_R} \qquad \text{if B is chosen} \quad [39]$$

and

$$\bar{R}_B(k+1) = \bar{R}_B(k) \qquad \text{if A is chosen} \quad [40]$$

$$\bar{R}_A(k+1) = \bar{R}_A(k) \qquad \text{if B is chosen.} \quad [41]$$

We choose $\tau_R = 5$.

The reward prediction error (RPE) at trial k is given by the difference $\text{RPE}(k) = R(k) - \bar{R}_B(k)$. Then, Eq. (37) can be written as

$$\Delta W_x = \text{RPE}(k) f(r_x)^2 \qquad x = A, B. \quad [42]$$

We assume that both the average reward $\bar{R}_B(k)$ and RPE is computed elsewhere in the brain, most probably in the Ventral Tegmental Area (VTA) (17).

The Hebbian term of the three-factor learning rule (14), which is the product of a non-linear function of pre and post-synaptic activity of the selective populations, is given by the term $f(r_x)^2 = f(r_x) \cdot f(r_x)$ for $x = A, B$. Notice that the Hebbian term is squared because the pre and post-synaptic activity is the activity of the selective populations A or B. This term will modify the recurrent connections of the selective populations $W_A(k)$ and $W_B(k)$. In our model, the function f is a highly non-linear step function

$$f(s;k) = \begin{cases} 0 & \text{if } s < \mu_x(k) + \Delta \\ \sqrt{A_H} & \text{if } \mu_x(k) + \Delta \leq s \end{cases} \qquad x = A, B. \tag{43}$$

Here $\mu_A(k)$ and $\mu_B(k)$ correspond to the running temporal firing rate average of population A and B in the valuation network, respectively, at trial k. The above function induces plasticity in synapses where the firing rates are Δ larger than the average firing rate of the previous trial. This class of non-linear plasticity rule that potentiates firing rate outliers has been recently inferred from *in vivo* data (18, 19).

In our model, the running average $\mu_A(k)$ and $\mu_B(k)$ we assume a slow exponential kernel of the order of seconds. Since, in our task, the ITI period is much longer than the response period, we can replace $\mu_A(k)$ and $\mu_B(k)$ by the mean firing rates during the ITI. After the go signal, when the choice is selected by the action selection network, the chosen selective population in the action selection network will elevate its firing rate (see Fig. 1B, bottom row right). The feedback projections from the action selection to the valuation network will produce a transient elevated activity in the corresponding population. In contrast, little to no feedback activity is produced in the corresponding unchosen population (see Fig. 1B, bottom row left). This selective feedback activity in the valuation network and our learning rule in Eq. (43) produce changes in the recurrent synaptic efficacies of only the chosen populations in the valuation network. Therefore, the final learning rule we use in our numerical simulations is

$$w_A(k+1) = \rho_w w_A(k) + A_H \mathrm{RPE}(k) \qquad \text{if A is chosen} \tag{44}$$

$$w_B(k+1) = \rho_w w_B(k) + A_H \mathrm{RPE}(k) \qquad \text{if B is chosen} \tag{45}$$

and

$$w_A(k+1) = \rho w_A(k) \qquad \text{if B is chosen} \tag{46}$$

$$w_B(k+1) = \rho w_B(k) \qquad \text{if A is chosen.} \tag{47}$$

The parameter $\rho_w \in [0, 1]$ represents to a trial-by-trial forgetting parameter. If $\rho_w = 1$, there is no forgetting across trials, while if $\rho_w = 0$, the synaptic efficacies modifications are forgotten in one trial. We choose no forgetting $\rho_w = 1$ for simulations in Fig. 1A-C, and we fit this parameter using behavioral data in Fig. 1D (see below section on fitting network models in behavioral data). The parameter A_H in Eqs. (43-45) corresponds to the maximum amplitude of the Hebbian term in our three-factor learning rule. For the network simulations in Fig. 1A-C $A_H = 0.1$ while we fit this parameter using behavioral data in Fig. 1D (see last section).

Neural integrator.

Low-dimensional neural integrator. For completeness, we describe a network model corresponding to two one-dimensional neural integrators for storing action value in the valuation network. In this model, we assume that action values are encoded for two distinct recurrent networks encoding the action value of each corresponding contingency. This dynamics is given by

$$\frac{dr_A}{dt} = I_A + \sqrt{2\sigma}\xi_A(t) \tag{48}$$

$$\frac{dr_B}{dt} = I_B + \sqrt{2\sigma}\xi_B(t). \tag{49}$$

The variables $\xi_A(t)$ and $\xi_B(t)$ correspond to Gaussian noise with mean zero and variance equal to 1. This model integrates perturbations which makes it extremely labile to noise.

High-dimensional neural integrator. For constructing a neural integrator in high dimensional neural space, our starting point is a recurrent linear network with N number of units. The following linear equations give the dynamics

$$\frac{dr_i}{dt} = -r_i + \sum_{j=1}^{N} J_{ij} r_j + I_i. \tag{50}$$

In this network model, the connectivity is random. Each entry is drawn from a normal distribution with zero mean and variance equal to g^2/N. The external input to the network is given by the vector $\vec{I}$. For all the simulations in the paper, we used $g = 0.5$.

By defining the below matrix

$$\mathbf{W} = \mathbf{J} - \mathbf{I}, \tag{51}$$

where $\mathbf{I}$ is the identity matrix. Then Eq. (50) can be written in vectorial form as

$$\frac{d\vec{r}}{dt} = \mathbf{W}\vec{r} + \vec{I}. \quad [52]$$

Using the singular value decomposition (SVD) the matrix W can be written as the sum of rank-1 matrices scaled by the corresponding singular vectors

$$\mathbf{W} = \sum_{\mu=1}^{N} \sigma_\mu \vec{u}_\mu \vec{v}_\mu^T. \quad [53]$$

Here the singular vectors are ordered from larger (σ_1) to the smaller (σ_N). We constructed a neural integrator for encoding two action values using similar ideas as in (20). For constructing a neural integrator with two integration modes, we subtract to the matrix $\mathbf{W}$ the rank-1 matrices in Eq. (53) corresponding to the two smaller singular values, i.e.,

$$\mathbf{L} = \mathbf{W} - \sigma_N \vec{u}_N \vec{v}_N^T - \sigma_{N-1} \vec{u}_{N-1} \vec{v}_{N-1}^T - \epsilon \left[\vec{v}_N \vec{v}_N^T + \vec{v}_{N-1} \vec{v}_{N-1}^T \right]. \quad [54]$$

Notice that here introduce a rank-2 normal perturbation that will lead to drifts of time scale $\tau \sim 1/\epsilon$ in the directions $\vec{v}_{N-1}$ and $\vec{v}_N$. We then use this new matrix $\mathbf{L}$ as the connectivity matrix of our neural integrator network, which now has two integration modes. The dynamics of the neural integrator is given by

$$\frac{d\vec{r}}{dt} = \mathbf{L}\vec{r} + \vec{I}. \quad [55]$$

By projecting the network activity $\vec{r}(t)$ onto the right singular vectors we decompose the activity in N network modes

$$m_k(t) = \vec{v_k}^T \vec{r}(t) = \sum_{j=1}^{N} v_{k,j} r_j(t) \qquad k = 1, \ldots, N. \quad [56]$$

The dynamics in the action value axis Eq. (55)

$$\frac{dm_N}{dt} = \langle \vec{I}, \vec{v}_N \rangle - \epsilon m_N \quad [57]$$

$$\frac{dm_{N-1}}{dt} = \langle \vec{I}, \vec{v}_{N-1} \rangle - \epsilon m_{N-1}. \quad [58]$$

For $\epsilon = 0$ reads

$$\frac{dm_N}{dt} = \langle \vec{I}, \vec{v}_N \rangle \quad [59]$$

$$\frac{dm_{N-1}}{dt} = \langle \vec{I}, \vec{v}_{N-1} \rangle, \quad [60]$$

which correspond to the equations of two independent one-dimensional neural integrators (see Eq. (48,49)). Where $\langle \cdot, \cdot \rangle$ corresponds to the inner product of two vectors.

In our model we assume $\mathbf{W}$ to be independently and identically distributed (i.i.d.) from a Gaussian distribution, i.e., $W \overset{i.i.d.}{\sim} N(0, g^2/N)$. For a random input $\vec{I}$, since the singular vectors of $\mathbf{W}$ are uniformly distributed on a sphere of radius 1, the projections $\langle \vec{I}, \vec{v}_{N-1} \rangle$ and $\langle \vec{I}, \vec{v}_N \rangle$ vanish for large enough networks (i.e., for large N). Therefore, in our this model random perturbations would hardly perturb the encoding of the action value.

The input current to the network is given by

$$\vec{I} = \vec{I}_{\mathrm{RPE}}(t) + \vec{I}_{\mathrm{ext}}(t). \quad [61]$$

Here $\vec{I}_{\mathrm{RPE}}(t)$ is an input proportional to the RPE that updates the action value after reward delivery by briefly pushing or pulling in the direction of the action value axis corresponding to the chosen contingency (i.e., in the directions $\vec{v}_{N-1}$ and $\vec{v}_N$ for the contingency A and B respectively). We called the subspace spanned by $\vec{v}_{N-1}$ and $\vec{v}_N$ the action value manifold. The $\vec{I}_{\mathrm{RPE}}(t)$ is given by

$$\vec{I}_{\mathrm{RPE}}(t) = \begin{cases} 0 & t \leq T_{\mathrm{choice}} \\ A_{\mathrm{RPE}} \mathrm{RPE}(k) \vec{v}_x & T_{\mathrm{choice}} < t < T_{\mathrm{choice}} + 100\mathrm{ms} \\ 0 & T_{\mathrm{choice}} + 100\mathrm{ms} \leq t \end{cases}, \quad [62]$$

where $\vec{v}_x = \vec{v}_{N-1}$ if A is chosen and $\vec{v}_x = \vec{v}_N$ if B is chosen. The parameter A_{RPE} is the strength by which the RPE signal modifies the population activity. For Fig. 1B $A_{\mathrm{RPE}} = 0.1$ and for Fig. 1C,D $A_{\mathrm{RPE}} = 0.15$. The RPE signal stimulates the network after the reward is delivered, which happens instantly after the choice is made at time T_{choice} in our model. T_{choice} is $T_{go} = 3$s plus the reaction time of the action selection network at a given trial and therefore varies from trial to trial.

The $\vec{I}_{\mathrm{ext}}(t)$ is an external perturbation to the network. For Fig. 1 and Fig. 2A we set $\vec{I}_{\mathrm{ext}}(t) = \vec{0}$.

The numerical integration time-step used was $dt = 10$ms.

Three classes of external external perturbations. For Fig. 2C, D, we investigate the effect of perturbations by stimulating the network for a brief period of 100ms with a constant external input $\vec{I}_{\text{ext}}$. We use three classes of external stimulation patterns with the same norm $\|\vec{I}_{\text{ext}}\| = 5$, which are summarized below.

1. The on-manifold perturbation is given by
$$\vec{I}_{\text{ext}} = \alpha \vec{v}_N + \beta \vec{v}_{N-1}. \quad [63]$$
For Fig. 2C we use $\alpha = 5$ and $\beta = 0$.

2. The random Gaussian perturbation is given by
$$\vec{I}_{\text{ext}}(t) = \frac{A_{\text{Random}}}{\sqrt{N}} \vec{x}, \quad [64]$$
with $x_i \overset{i.i.d.}{\sim} N(0,1)$. For Fig. 2C we use $A_{\text{Random}} = 5$. Notice that, the average norm of this stimulus is $\langle \|\vec{I}_{\text{ext}}\| \rangle = A_{\text{Random}} = 5$.

3. The uniform perturbation is given by
$$\vec{I}_{\text{ext}} = \frac{-A_{\text{Uniform}}}{\sqrt{N}} \vec{1}, \quad [65]$$
where $\vec{1}$ is the vector with all its entries equal to one. For Fig. 2C we use $A_{\text{Uniform}} = 5$.

Engineered perturbations. For constructing the engineered perturbations in Fig. 2D we first ranked the entries of the action value encoding vector used for the external stimulation $\vec{v}_N$ as $v_{l_1}, v_{l_2}, \cdots, v_{l_N}$, where v_{l_1} has the largest positive value and v_{l_N} the largest negative value. Higher ranked neurons (e.g., neuron l_1) will strongly encode action value for the contingency encoded by $\vec{v}_N$. Then we constructed several stimuli of the form
$$\vec{I}_{\text{ext}}(K) = \frac{A_{\text{Engineered}}}{\sqrt{\sum_{i=1}^{K} v_{l_i}^2}} \left(0, \cdots, v_{l_1}, 0, \cdots, v_{l_2}, 0, \cdots, v_{l_K}, 0, \cdots\right)^T. \quad [66]$$
Here the stimulation vector $\vec{I}_{\text{ext}}(K)$ has only non-zero entries for the corresponding first K ranked entries. For the network with $N = 1000$ neurons in Fig. 2D, we use $K = 1, 10, 100, 1000$ and $A_{\text{Engineered}} = 4$.

Fitting network models in behavioral data.

Reduced network models. For the fitting of the neural integrator and synaptic network models on behavioral data, as in (12), we first reduced the action selection network for selecting the choice $c \in \{A, B\}$ to Bernoulli random variable with probability $p(c)$. The function $p(c)$ corresponds to a sigmoidal function of the difference of the input currents that project to the action selection network from the valuation network, i.e., the Softmax function given by
$$p(c) = \begin{cases} \frac{1}{1+\exp(-\beta(I_A - I_B)+\alpha)} & \text{if } c = A \\ 1 - \frac{1}{1+\exp(-\beta(I_A - I_B)+\alpha)} & \text{if } c = B. \end{cases} \quad [67]$$
Here α and β are two parameters fitted from behavior.

From Eqs. (44-47) for the synaptic model, the synaptic weights change acording to the below equations
$$w_{ch}(k+1) = \rho_w w_{ch}(k) + A_H (R(k) - [r^*_{ch}(k)]_+) \qquad ch \in \{A, B\} \quad [68]$$
and for the unchosen site
$$w_{uch}(k+1) = \rho_w w_{uch}(k) \qquad uch \in \{A, B\}. \quad [69]$$
The term $r^*_{ch}(k)$ is the valuation network's stationary firing rate of the chosen population given by Eqs. (22, 23). The function $[\bullet]_+$ is the rectifier linear function. Lastly, the term $[r^*_{ch}(k)]_+$ correspond to the predicted reward (21) decoded directly from the valuation network. Using this term is a modification of the learning rule in Eqs. (37-41) in which the trial averaged reward is used instead of the predicted reward.

The input currents to the action selection softmax in Eq. (67) are given by
$$I_A = j_A r^*_A(k) \quad [70]$$
$$I_B = j_B r^*_B(k) \quad [71]$$
Here $j_A = j_B = 1$. We approximate the firing rates after the go cue by the stationary firing rates $r^*_A(k)$ and $r^*_B(k)$ at trial k given by Eqs. (22, 23).

For the neural integrator, by solving the dynamics during the ITI in Eqs.(57, 58) we found that at the time of go cue $T(k)$, the projection to the action value axis for the chosen contingency is given by

$$m_{ch}(k+1) = \left[m_{ch}(k) + A_{\mathrm{RPE}}(R(k) - \bar{R}_{ch}(k))\right] e^{-\epsilon T(k)} \qquad ch \in \{A, B\}, \tag{72}$$

and for the unchosen one

$$m_{uch}(k+1) = m_{uch}(k) e^{-\epsilon T(k)} \qquad ch \in \{A, B\}. \tag{73}$$

Notice that ϵ corresponds to the drift parameter in Eq. (54). The input currents to the action selection softmax in Eq. (67) are given by

$$I_A = j_A m_A(k) \tag{74}$$

$$I_B = j_B m_B(k). \tag{75}$$

Here $j_A = j_B = 1$.

Fitting procedure. Then the likelihood is given by

$$p(\{\text{choices}, \text{rewards}\} | \{\alpha, \beta, A, \epsilon\}) = \prod_{k=1}^{N_T} p(c_k | m_R(k), m_L(k), \{\alpha, \beta, A, \epsilon\}) \tag{76}$$

where c_k is the animal's actual choice on trial k and N_T is the total trial number. We fitted the parameters $\{\alpha, \beta, A, \epsilon\}$ by maximizing this likelihood using differential evolution (python package **scipy.optimize**). For the behavioral session shown in Fig. 1E, $N_T = 1,067$.

For Fig. 1E, the fitted parameters are specified below.

- For the neural integrator are $\beta = 100$ (Eq. (67)), $\alpha = -0.44$ (Eq. (67)), $A_{\mathrm{RPE}} = 0.01$ (Eqs. (72, 73)), and $\epsilon = 0.042/s$ (Eqs. (72, 73)).
- For the synaptic model are $\beta = 100$ (Eq. (67)), $\alpha = -0.48$ (Eq. (67)), $A_H = 1$ (Eqs. (68, 69)), $\rho_w = 0.19$ (Eqs. (68, 69)), $I_0 = 0.09$ (Eqs. (22, 23), and $\sigma = 0.11$ (Eqs. (22, 23).
- For the Rescorla–Wagner model (22), the fitted learning rate was 0.15. For producing the choice we use the Softmax in Eq. (67) with the fitted parameters $\beta = 4.16$ (Eq. (67)) and $\alpha = -0.89$ (Eq. (67)).

Code

The code for simulating the network models is available at the GitHub repository https://github.com/ulisespereira/foraging-integrator-vs-synaptic. The code for fitting the reduced network models on mice behavior is available upon request.

References

1. BA Bari, et al., Stable representations of decision variables for flexible behavior. *Neuron* **103**, 922–933 (2019).
2. R Hattori, B Danskin, Z Babic, N Mlynaryk, T Komiyama, Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872 (2019).
3. LP Sugrue, GS Corrado, WT Newsome, Matching behavior and the representation of value in the parietal cortex. *science* **304**, 1782–1787 (2004).
4. GS Corrado, LP Sugrue, HS Seung, WT Newsome, Linear-nonlinear-poisson models of primate choice dynamics. *J. experimental analysis behavior* **84**, 581–617 (2005).
5. KF Wong, XJ Wang, A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).
6. LF Abbott, FS Chance, Drivers and modulators from push-pull and balanced synaptic input. *Prog. brain research* **149**, 147–155 (2005).
7. A Finkelstein, et al., Attractor dynamics gate cortical information flow during decision-making. *Nat. Neurosci.* **24**, 843–850 (2021).
8. S Recanatesi, U Pereira-Obilinovic, M Murakami, Z Mainen, L Mazzucato, Metastable attractors explain the variable timing of stable behavioral action sequences. *Neuron* **110**, 139–153 (2022).
9. L Logiaco, L Abbott, S Escola, Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell reports* **35**, 109090 (2021).
10. ZV Guo, et al., Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545**, 181–186 (2017).
11. RN Bracewell, RN Bracewell, *The Fourier transform and its applications.* (McGraw-Hill New York) Vol. 31999, (1986).
12. A Soltani, XJ Wang, A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J. Neurosci.* **26**, 3731–3744 (2006).

13. Y Loewenstein, HS Seung, Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proc. Natl. Acad. Sci.* **103**, 15224–15229 (2006).
14. N Frémaux, W Gerstner, Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. neural circuits* **9**, 85 (2016).
15. N Frémaux, H Sprekeler, W Gerstner, Functional requirements for reward-modulated spike-timing-dependent plasticity. *J. Neurosci.* **30**, 13326–13337 (2010).
16. TA Engel, W Chaisangmongkon, DJ Freedman, XJ Wang, Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat. communications* **6**, 1–12 (2015).
17. W Schultz, P Dayan, PR Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
18. S Lim, et al., Inferring learning rules from distributions of firing rates in cortical neurons. *Nat. neuroscience* **18**, 1804–1810 (2015).
19. U Pereira, N Brunel, Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron* **99**, 227–238 (2018).
20. S Druckmann, DB Chklovskii, Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol.* **22**, 2095–2103 (2012).
21. RS Sutton, AG Barto, *Reinforcement learning: An introduction.* (MIT press), (2018).
22. RA Rescorla, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Curr. research theory*, 64–99 (1972).