Linking Tonic Dopamine and Biased Value Predictions in a Biologically Inspired Reinforcement Learning Model

Sandra Romero Pinto *
Division of Medical Sciences
Harvard University
Cambridge, MA 02139
sromeropinto@g.harvard.edu

Naoshige Uchida
Department of Molecular and Cellular Biology
Harvard University
Cambridge, MA 02139
uchida@mcb.harvard.edu

Abstract

Some psychiatric disorders are characterized by excessively optimistic or pessimistic predictions of future events, as well as changes in dopamine levels. However, how changes in dopamine lead to biased value predictions is unknown. Here, we examine this connection by examining the role of baseline dopamine levels in value learning. Value learning is thought to depend, in part, on synaptic plasticity driven by dopamine reward prediction errors acting upon D1 and D2 receptors in spiny projection neurons of the striatum. At reported striatal dopamine levels, D1 receptors are mostly unoccupied by dopamine, while D2 receptors are mostly occupied, making them sensitive to increases and decreases of dopamine, respectively. Accordingly, studies have reported that potentiation in SPNs expressing D1 or D2 receptors is triggered by phasic increases or decreases of dopamine. Moreover, given the receptors' sigmoidal dose-occupancy relationship, shifts in the dopamine baseline should change their sensitivity to dopamine transients (i.e., take the baseline to a "steep" or "shallow" region of the dose-occupancy curve). Here, we show that a reinforcement learning model incorporating these plasticity rules develops positive or negative biases in predictions of probabilistic rewards when baseline dopamine is increased or decreased, respectively. We validate the model using experimental data from a previous study. This study showed that lesions of the habenula resulted in positive biases both in reward-seeking behavior (anticipatory licking) and dopamine neurons' responses to cues predictive of probabilistic rewards. In our model, an increase in baseline firing of dopamine neurons, as observed in the data, is sufficient to lead to these optimistic biases. Taken together, our biologically inspired RL model highlights a causal impact of baseline dopamine on biasing value predictions, which may underlie abnormalities in psychiatric patients, including altered risk preferences.

Keywords: reinforcement learning, dopamine, affective biases, neuropsychiatric disorders

Acknowledgements

We thank B. Sabatini, S. Gershman, D. Polley, E. Phelps and M. Andermann for discussions and feedback; I. Green and members of the Uchida lab for discussions. This work was supported by US National Institutes of Health grants (U19NS113201, R01NS116753), Harvard Brain Science Institute Bipolar Seed grant, and Simons Collaboration on Global Brain.

*https://scholar.harvard.edu/sandraromeropinto

1 Outline

A hallmark of various psychiatric disorders is abnormal future predictions. For example, patients with major depressive disorder (MDD) make unrealistically negative predictions about future events. However, how these biased predictions arise remains unknown. In this study, we use reinforcement learning (RL) frameworks and propose a new mechanism that explain biased value learning. We begin by introducing two RL models that have been used to model biased value learning – risk-sensitive RL and distributional RL (DRL). We then introduce a new mechanism for biased value learning inspired by recent experimental findings on synaptic plasticity and anatomical features of the key RL circuit in the brain. We will then show that our new model can explain the neural recording and behavioral data in animals with lesions of the habenula, the brain region implicated in MDD.

2 Emergence of Biases in Value Predictions in Reinforcement Learning

2.1 Previous RL proposals to account for biases in value

In RL, an agent learns to predict future rewards by learning the expected cumulative future rewards held by different states (i.e., a value function $V(s)$) (*1*). Here, for simplicity, we will develop our model in a single state environment, dropping the V dependence on the state. Learning of value V is driven by reward prediction errors (RPEs), defined as the discrepancy between the actual and predicted reward (δ) (Eq 1). The value can be learned by updating the V estimate so as to minimize RPEs (Eq 2). If rewards are stochastic, V converges to the expected value of the reward distribution.

$$\delta = r - V \tag{1}$$

$$V \leftarrow V + \delta \cdot \alpha \tag{2}$$

Individuals MDDs exhibit enhanced sensitivity to losses relative to gains and attenuated sensitivity to rewarding stimuli. Furthermore, learning from negative outcomes is either less affected or augmented (*2*). Thus, individuals with MDD usually excel at loss minimization but underperform at gain maximization in decision-making tasks (*3*). This unbalanced sensitivity results in biases in value predictions (*4*) and excessive risk-aversive behaviors(*5*) (Fig 1). Furthermore, previous studies (*6*) have observed biases in value learning in humans with MDD. Biased value learning observed in these situations have been modeled using so-called risk-sensitive RL (*7*). Risk sensitive RL arises from asymmetric updates for positive versus negative RPEs (Eq 3): An agent learning probabilistic rewards with biased learning rate parameters for positive versus negative RPEs (i.e. $\alpha^+ \neq \alpha^-$) converges on value estimates higher or lower than the expected value, and therefore develop optimistic or pessimistic expectations, respectively. In other words, the agent becomes risk-seeking or risk-averse (Fig 1). While risk-sensitive RL captures biased value learning, the mechanism that regulate learning rates (α^+, α^-) remain unknown.

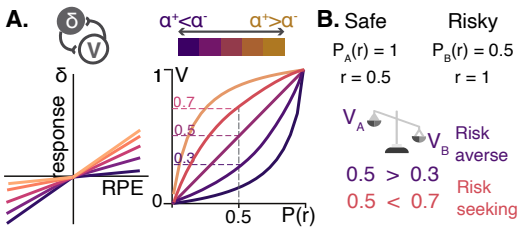

Figure 1: **A** Risk sensitive RL model, with asymmetric updates implemented here via asymmetric scaling of RPE evoked responses (α^+,α^-). **B** In choosing between a safe and risky option of equal expected value (E[V]=0.45), agents will behave as risk seeking or averse if it developed positively or negatively biased estimates of V.

The concept of asymmetric updates has been deployed in the novel RL framework called distributional RL (DRL) (*8*). DRL allows an agent to learn the entire distribution of rewards, as opposed to the expected value. In DRL, an agent is equipped with a set of value predictors (V_i), updated asymmetrically (Eq 4) for positive and negative RPEs with a bias characterized by their asymmetric scaling factor (τ_i, Eq 5). Using this learning rule (Eq 4), each V_i converges on the τ_i^{th} expectile of the distribution (Fig 2). Expectiles are the solutions to asymmetric least squares minimization and generalize the mean of a distribution as quantiles generalize the median (*8*). Since a set of expectiles defines a distribution, the diversity of τ_i across the population enables learning of a distributional value code.

$$\begin{aligned} V &\leftarrow V + \delta \cdot \alpha^+, \text{ if } \delta > 0 \\ V &\leftarrow V + \delta \cdot \alpha^-, \text{ if } \delta < 0 \end{aligned} \tag{3}$$

A previous study has found a potential neural substrate of asymmetric learning rate parameters. Dabney et al., (*9*) showed that dopamine neurons vary in terms of how their firing rate responses scale with positive and negative RPEs. DRL can then be implemented exploiting such diversity, with α_i^+ and α_i^- corresponding to the slopes of each dopamine neuron's response function for positive and negative RPEs.

$$\begin{aligned} V_i &\leftarrow V_i + \delta_i \cdot \alpha^+, \text{ if } \delta_i > 0 \\ V_i &\leftarrow V_i + \delta_i \cdot \alpha^-, \text{ if } \delta_i < 0 \end{aligned} \tag{4}$$

$$\tau_i = \frac{\alpha_i^+}{(\alpha_i^+ \alpha_i^-)}, \text{ with } \{\tau_i, \alpha_i^+, \alpha_i^-\} \in [0; 1] \tag{5}$$

In both risk sensitive RL and DRL, biases in value prediction can arise because learning rates for positive versus negative RPEs become unbalanced at a global level. At the mechanistic level, this can arise via the scaling of dopamine RPE responses, as shown in (*9*). In addition, biased value learning can, in principle, arise from difference in the efficacy of value updating from positive and negative RPEs, i.e. the efficacy of synaptic plasticity. It remains unknown whether these mechanisms play a role in causing biases in value prediction or what is the biological basis of these mechanisms.

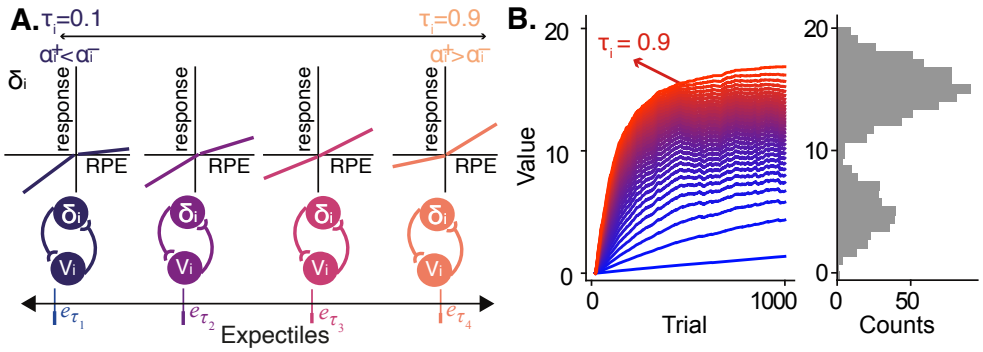

Figure 2: **A** Schematic of DRL. The dopamine population (δ_i) presents a distribution of asymmetric scaling factors (τ_i) determined by the single-cell specific scaling of responses to positive and negative RPEs (α_i^+,α_i^-).The scaled RPEs are used to update the respective value estimates (V_i) that will converge to a given expectile of the distribution (e_{τ_i}). **B** Learning of a distribution of expectiles for a bimodal distribution.

3 Accounting for biological plasticity rules in reinforcement learning models

The basic assumptions of simple RL models do not match the structure of neural circuits implicated in RL in the brain. In standard RL models, each value predictor is typically updated by both positive and negative RPEs. If the value is computed based on a linear sum of inputs ("feature vectors"), the update rules described above (Eq 2,3,4) are equivalent to performing a gradient descent that minimizes RPEs (δ_i) (*1*). However, this simple architecture does not necessarily match with the anatomy of the brain neural circuit: dopamine-recipient cells (spiny projection neurons, or SPNs) in the striatum are categorized into direct- and indirect-pathway SPNs, which express D1- and D2-type dopamine receptors, respectively. Anatomically, D1- and D2-SPNs exert opposing effects on downstream neurons (*10*). Importantly, recent studies (*11*, *12*) have shown that D1- and D2-SPNs use different plasticity rules: D1-SPNs are potentiated by a phasic increase in dopamine, whereas D2-SPNs are potentiated by a transient decrease in dopamine.

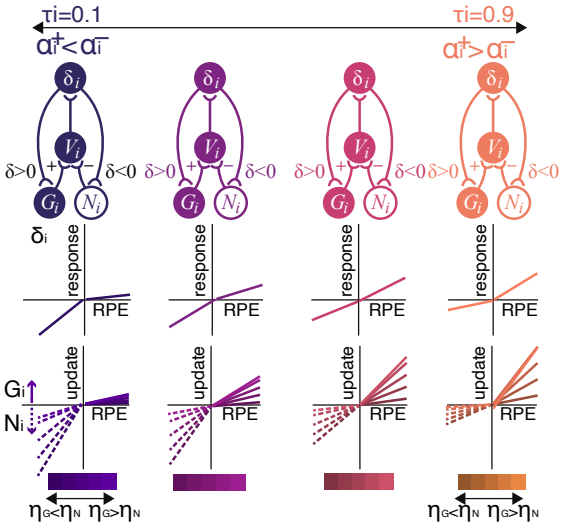

Figure 3: **A** Schematic of DRL with SPNs populations. The dopamine population (δ_i) presents a distribution of asymmetric scaling factors (τ_i, mid-row). The scaled positive or negative RPEs are used to update G_i or N_i. There is a global rescaling of the updates (η_G,η_N) at the SPN level (Bottom). The value predictions (V_i) are the pairwise difference between G_i and N_i (Top).

Whether these anatomical architectures and the recently found plasticity rules are compatible with RL models remain unclear.

Therefore, we sought to examine whether we can obtain new insights by taking into account these biological findings. We implemented a DRL model with separate populations of value predictors corresponding to D1- and D2-SPNs, which store the quantities G_i and N_i, respectively. Mimicking dopamine's effect on potentiation in D1- and D2-SPNs, G_i or N_i *increase* their estimates if an RPE is positive or negative, respectively, with the learning rates defined by η_G and η_N(Fig 3, Eq 6). For simplicity, we assume η_G and η_N to be equal for all neurons within each population. In addition, η_G and η_N scale only the *update* process and are independent of the RPE scaling factors (α_i^+, α_i^-).

$$G_i \leftarrow G_i + \eta_G \cdot \hat{\delta}_i, \text{ if } \delta_i > 0, \text{ where } \hat{\delta}_i = \alpha_i^+ \cdot |\delta_i|$$
$$N_i \leftarrow N_i + \eta_N \cdot \hat{\delta}_i, \text{ if } \delta_i < 0, \text{ where } \hat{\delta}_i = \alpha_i^- \cdot |\delta_i| \quad (6)$$

$$V_i = G_i - N_i \quad (7)$$

$$V_i = \frac{(\tau_i/(1-\tau_i)\cdot\phi/(1-\phi)\cdot/(1-p)\cdot r)}{(\tau_i/(1-\tau_i)\cdot\phi/(1-\phi)\cdot p/(1-p)+1)} \quad (8)$$
$$\text{where } \phi = \eta_G/(\eta_N+\eta_G)$$

We can show that the value (V_i) can be obtained as the difference between G_i and N_i (*10*) (Eq 7). For a Bernoulli distribution (of reward r with probability p, and 0 otherwise) a given V_i will converge to the expectile defined by Eq 8.

This model preserves various properties in the previous models including the equivalence of the simple update equation (Eq 6) with a gradient descent that minimizes RPEs (δ_i), yet more closely reflects the plasticity rules found in the brain.

3.1 Different affinities of dopamine receptors induces biases in value learning with changes in baseline dopamine

The efficacy of synaptic plasticity is modulated by a number of factors (*13*). Here we explore the role of baseline dopamine levels, whose changes are implicated in MDD. Importantly, D1- and D2-type dopamine receptors (D1r and D2r, respectively) have different affinities to dopamine: high in D2r and low in D1r (EC50 affinity constant is 1 µM for D1r and 10 nM for D2r (*14*)). This means that at the normal dopamine level (approx. 60nM), the receptor occupancy (i.e. the fraction of receptors binding their ligand, dopamine) is high in D2rs and low in D1rs (*14*) (Figure 4a). The receptor's dose-occupancy relationship in both types is sigmoidal but shifted with one another with respect to dopamine level.

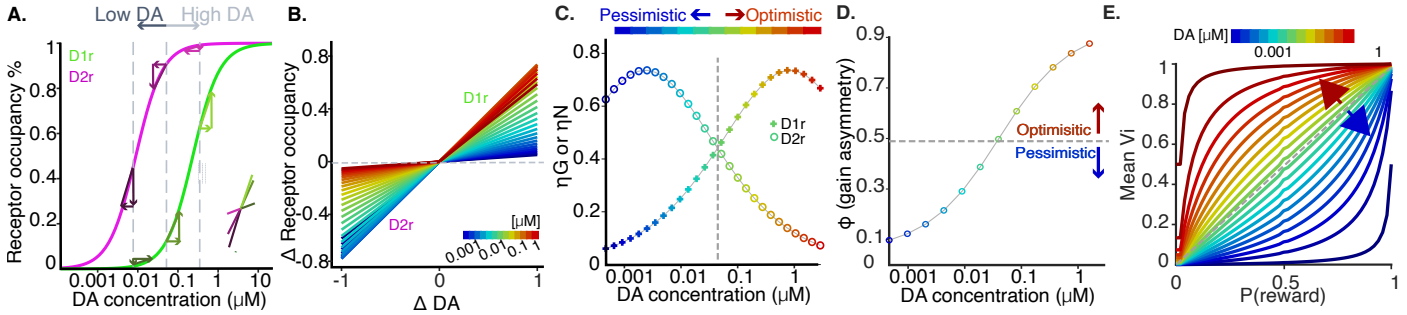

Figure 4: **A** Dose-occupancy curves for D1r and D2r. Insert: Change in D1r and D2r occupancy with a transient increase or decrease in dopamine for three dopamine baseline levels. **B** Change in receptor occupancy for a given dopamine transient (ΔD) depends on baseline dopamine (colorbar).**C** Receptors' sensitivity (slope of curves in B) as a function of baseline dopamine. **D** Asymmetric factor for updates gains (ϕ)) as a function of baseline dopamine.

Given the sigmoidal relationship between dopamine concentration and receptor occupancy, the slope of the dose-occupancy curve (here we define this as receptor sensitivity) changes depending on the starting dopamine level (Fig 4b). That is, a given phasic change in dopamine leads to different magnitudes of change in receptor occupancy depending on the baseline level. Furthermore, given the different affinities of D1r and D1r, an increase or decrease in the baseline dopamine level can lead to opposite effects on their sensitivity. For example, a moderate increase in baseline dopamine can make D1r more sensitive while D2r less sensitive to a phasic increase or decrease in dopamine, respectively (Fig 4a).

Finally, based on previous results (*11*, *15*), we assume that the efficacy of synaptic potentiation scales with the product of three factors: the magnitude of dopamine transients (i.e. RPEs), the pre-synaptic activity, and the post-synaptic activity. Thus, the impact of dopamine transients on the efficacy of plasticity depends on the receptor sensitivity at a given moment: keeping fixed the pre-synaptic activity, the potentiation in D1- and D2-SPNs will be proportional to $\eta_G \cdot \delta$ and $\eta_N \cdot \delta$, respectively, where now η_G and η_N are the receptor sensitivities controlled by the baseline dopamine level.

Taken together, these results indicate that the baseline dopamine level can cause asymmetries in the value updates for positive versus negatives RPEs, via the η_G and η_N scaling factors (summarized by ϕ in Fig 4d). Therefore changes in baseline dopamine can lead to biases in value learning (Fig 4e).

4 Baseline dopamine changes account for biases in value in basal ganglia circuitry impairments

4.1 Biases in value learning can arise from the changes in dopamine activity related to mood disorders

These results raise the possibility that an overall decrease in the baseline firing of dopamine neurons may cause depressive-like symptoms. Just by assuming that this leads to a decrease in baseline dopamine level in the striatum, our model readily predicts that learning from negative outcomes will be emphasized over learning from positive outcomes (Fig 4). As in patients with MDD, RL agents learning in these conditions present enhanced risk-aversive behavior, pessimistic outcome expectations, and increased sensitivity to losses compared to gains.

It has been reported that spontaneous activity of dopamine neurons decreases in animal models of depression (*16*). In addition, decreased dopamine firing is consistent with an increased burst firing of lateral habenula (Lhb) neurons of depressive mice (*17*), as an increased activity in LHb neurons can suppress dopamine neurons. Further to this, some depression symptoms such as anhedonia can be ameliorated by optogenetic activation of dopamine neurons (*16*).

4.2 Baseline dopamine explains optimistic biases induced by habenula lesions

To test our model, we examined experimental data obtained in our previous study, that tested the effect of habenular lesions on dopamine neurons' firing and on reward-seeking behavior (*18*). Here, head-fixed mice were trained in a task

in which odor cues predicted different probabilities of reward delivery (10%, 50%, 90%), and then underwent habenula (n=5) or sham (n=7) lesions. After lesions, mice exhibited an elevated reward-seeking behavior (anticipatory licking) in anticipation of uncertain rewards, suggesting optimistic biases in reward expectation (Fig 5a). A similar bias was observed in the normalized cue-evoked dopamine responses after lesions (Fig 5b), although the responses were smaller in absolute magnitudes.

We next examined which aspects of dopamine activity can explain the optimistic biases in cue responses and behavior. First, our analysis showed that dopamine responses were largely consistent with the pattern of activity predicted by DRL, replicating the findings in Dabney et al. (*9*) in control and lesioned mice. However, the detailed analysis using learning rate parameters extracted from the data showed that neither a classic RL nor a DRL model could explain the optimistic biases in cue responses and behavior; although the τ_i distribution tiled a wider range between 0 and 1 after lesions, the mean did not change (Fig 5c-d), indicating a lack of bias between α_i^+ and α_i^- at the population level. Thus, contrary to the previous proposal made in the previous study (*18*), changes in the phasic responses (and the resulting scaling factors τ_i) do not explain the optimistic biases in cue responses and behavior.

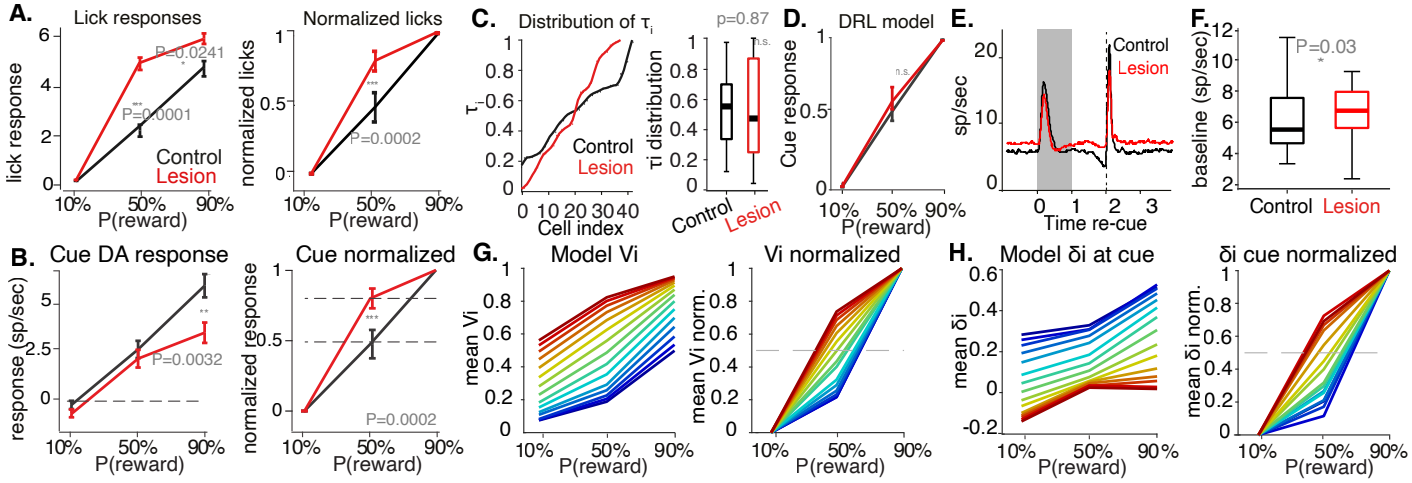

Figure 5: **A** Anticipatory licking to reward-predictive cues. **B** Average raw and normalized dopamine cue-evoked responses (n={45 control,44 lesion}) **C** τ_i distributions derived from dopamine outcome responses did not differ in their means. **D** DRL model with data-derived τ_i distributions did not explain optimistic biases in cue responses **E**. Average firing rates of all neurons (90% cue). **F** Distribution of baseline firing rates. **G** Model simulations of the task for a set of baseline dopamine levels (color in Fig 4). Increases in dopamine baseline lead to optimistic biases reflected in the mean and normalized mean of the V_i distribution. **H** δ_i cue responses decrease in magnitude with increases in dopamine baseline but normalized 50% cue response increases.

In addition to changes in phasic responses of dopamine neurons, an elevation in their baseline firing rates was observed after lesions (Fig 5e,f). This can increase striatal baseline dopamine levels, and, in turn, lead to a global asymmetry in value updates ($\eta_G > \eta_N$) in our model, as shown above. This results in a positive bias in the learned expected value and distribution, leading to the observed optimistic biases and changes in cue responses (Fig 5g,h). In addition, the seemingly unrelated decreases in magnitude of cue responses are reproduced in our model, and occur due to the increases in the pre-cue value baseline caused by learning under the optimistic regime. These results, together, indicate that changes in baseline firing of dopamine neurons, rather than changes in phasic responses, are a likely mechanism that led to optimistic biases in cue-evoked dopamine responses as well as reward-seeking behavior in habenula lesioned animals.

References

1. R. S. Sutton, A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
2. O. J. Robinson, H. W. Chase, *Computational Psychiatry* (2017).
3. W. T. Maddox *et al.*, *Cognition* (2012).
4. R. B. Rutledge *et al.*, *PNAS* (2014).
5. K. Baek *et al.*, *Scientific reports* (2017).
6. Y. Niv *et al.*, *Journal of Neuroscience* (2012).
7. O. Mihatsch, R. Neuneier, *Machine learning* (2002).
8. M. Rowland *et al.*, *ICML* (2019).
9. W. Dabney *et al.*, *Nature* (2020).
10. M. J. Frank, *Journal of cognitive neuroscience* (2005).
11. Y. Iino *et al.*, *Nature* (2020).
12. S. J. Lee *et al.*, *Nature* (2021).
13. S. J. Lee *et al.*, *Neuron* (2008).
14. M. E. Rice, S. J. Cragg, *Brain research reviews* (2008).
15. N. Frémaux, W. Gerstner, *Frontiers in neural circuits* **9**, 85 (2016).
16. K. M. Tye *et al.*, *Nature* (2013).
17. Y. Cui *et al.*, *Nature* (2018).
18. J. Tian, N. Uchida, *Neuron* (2015).