Dopamine and serotonin interplay for valence-based spatial learning

Carlos Wert-Carvajal[1,2], Melissa Reneaux[1], Tatjana Tchumatchenko [*2,3], and Claudia Clopath[*1]

[1]Bioengineering Department, Imperial College London, London SW7 2AZ, United Kingdom
[2]Theory of Neural Dynamics Group, Max Planck Institute for Brain Research, Frankfurt 60438, Germany
[3]Institute of Experimental Epileptology and Cognition Research, Life and Brain Center, University of Bonn Medical Center, Bonn 53127, Germany

March 4, 2021

Abstract

Dopamine and serotonin are important modulators of synaptic plasticity and their action has been linked to our ability to learn the positive or negative outcomes or valence learning. In the hippocampus, both neuromodulators affect long-term synaptic plasticity but play different roles in the encoding of uncertainty or predicted reward. Here, we examine the differential role of these modulators on learning speed and cognitive flexibility in a navigational model. We compare two reward-modulated spike time-dependent plasticity (R-STDP) learning rules to describe the action of these neuromodulators. Our results show that the interplay of dopamine (DA) and serotonin (5-HT) improves overall learning performance and can explain experimentally reported differences in spatial task performance. Furthermore, this system allows us to make predictions regarding spatial reversal learning.

1 Introduction

The interplay between dopamine (DA) and serotonin, or 5-hydroxytryptamine (5-HT), regulates cognitive functions underpinning decision-making. However, its behavioral consequences and integration into a control system remain elusive (Barnes and Sharp, 1999; Dayan and Huys, 2015). DA has long been characterized as a plasticity modulator that encodes uncertainty or prediction error in reinforcement learning (Schultz et al., 1997; Sutton and Barto, 2018). Unlike other neuromodulators, such as acetylcholine or noradrenaline (Frémaux and Gerstner, 2015), the role of 5-HT is less clear and it has been hypothesized to contribute to aversive processing, analogous to the function of DA in positive or appetite-driven rewards (Rogers, 2011; Cools et al., 2011; Crockett et al., 2012; Cohen et al., 2015; Fischer and Ullsperger, 2017). 5-HT also regulates traits of the social brain such as depression and impulsiveness (Rogers, 2011; Dalley and Roiser, 2012). In the hippocampus, which is critical for spatial memory formation (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978), DA and 5-HT have been studied in the context of valence-based learning (Fischer and Ullsperger, 2017; Fernandez et al., 2017; Schmidt et al., 2017; Waider et al., 2019). Even if true opponency

*Correspondence: clopath@imperial.ac.uk (CC), tatjana.tchumatchenko@brain.mpg.de (TT)

is not well-established (Daw et al., 2002; Boureau and Dayan, 2011), evidence suggests that the antagonistic effects of DA and 5-HT can explain neural activity during reward-driven learning (Crockett et al., 2012; Cohen et al., 2015; Matias et al., 2017). Notably, DA has been shown to induce long-term potentiation (LTP) in reward-guided navigation (Brzosko et al., 2015; Palacios-Filardo and Mellor, 2019), and 5-HT long-term depression (LTD) for some receptor-specific hippocampal areas (Kemp and Manahan-Vaughan, 2004, 2005; Berumen et al., 2012; Wawra et al., 2014; Lecouflet et al., 2020). However, 5-HT could also produce LTP or metaplasticity regulation (Wang and Arvanov, 1998; Hagena and Manahan-Vaughan, 2017; Teixeira et al., 2018).

Motivated by these experimental findings, we present a mathematical model of valence-based learning in the hippocampus which details the antagonistic roles of DA and 5-HT for long-term synaptic plasticity. To this end, we use available biological data describing the dynamics of both neuromodulators and present a stable neoHebbian three-factor learning rule (Frémaux and Gerstner, 2015; Gerstner et al., 2018; Zannone et al., 2018) characterising their effect in synapses. By evaluating and optimizing two spike time-dependent plasticity (STDP) rules during forward learning in a navigational task, we find that 5-HT increases training performance. Finally, we show that the proposed interplay of 5-HT and DA resembles behavioral evidence and can shape the adaptation in reversal learning (Matias et al., 2017).

2 Results and discussion

The valence system that we propose for DA and 5-HT contributions highlights the functional importance of the competition between timing-dependent long-term potentiation (t-LTP) and depression (t-LTD) during rewarding and punishing reinforcement cues. We followed a navigational hippocampal model (Foster et al., 2000; Vasilaki et al., 2009; Frémaux et al., 2013; Brzosko et al., 2017) with a feed-forward network of presynaptic place cells and a layer of postsynaptic action neurons (***Figure 1A***; Materials and methods). In reward-modulated spike timing-dependent plasticity (R-STDP) weight change is a function of the firing difference and the action of a neuromodulator (***Figure 1B***). We used previously reported data describing the STDP window for DA in the hippocampus (Brzosko et al., 2015, 2017) and assumed that LTD-inducing effects of 5-HT can be captured by an anti-causal learning window, as shown in cortical 5-HT_{2C} receptors (He et al., 2015; ***Figure 1B***). We found that varying STDP windows for 5-HT and DA produced similar navigational outcomes through an equal decay which we chose for further analysis (***Figure S1***). Temporal discrimination of neural activity leading up to the reinforcement signal was achieved through an eligibility trace (Gerstner et al., 2018), also known as proto-weight, for which there is evidence for DA (Brzosko et al., 2017) and 5-HT (He et al., 2015; ***Figure 1C***). The eligibility trace of 5-HT, adapted from the neocortex, presents slower dynamics than that of DA, in the hippocampus, for an equal Hebbian response (He et al., 2015; Saylor et al., 2019), which implies that distal predictive neural activity is less persistent under the modulation of the latter (***Figure 1C***). For the navigational task we employed a Morris water maze (MWM) where the agent has to find a hidden platform in a water arena (Vorhees and Williams, 2006;***Figure 1D***). Water is considered a punishing or stress-inducing cue (Harrison et al., 2009), which we hypothesized to cause 5-HT release (Karabeg et al., 2013), and subsequent arrival to the reward zone in the corner of the maze produces an increased dopaminergic response (Frémaux et al., 2013).

We considered two R-STDP learning rules to model DA and 5-HT. Firstly, we implemented sequential weight change (SWC), inspired by sequentially neuromodulated plasticity (sn-Plast) (Brzosko et al., 2017; Zannone et al., 2018), in which DA produces t-LTP and 5-HT induces t-LTD during exploration. As sn-Plast, SWC is outcome-dependent and relies on the assumption that either potentiation or depression occurs after a long delay, which allows decoupling rewarding and non-rewarding trials (***Figure 1E***; Materials and methods). In the MWM task, SWC produces mutually exclusive t-LTP, when the agent arrives at the

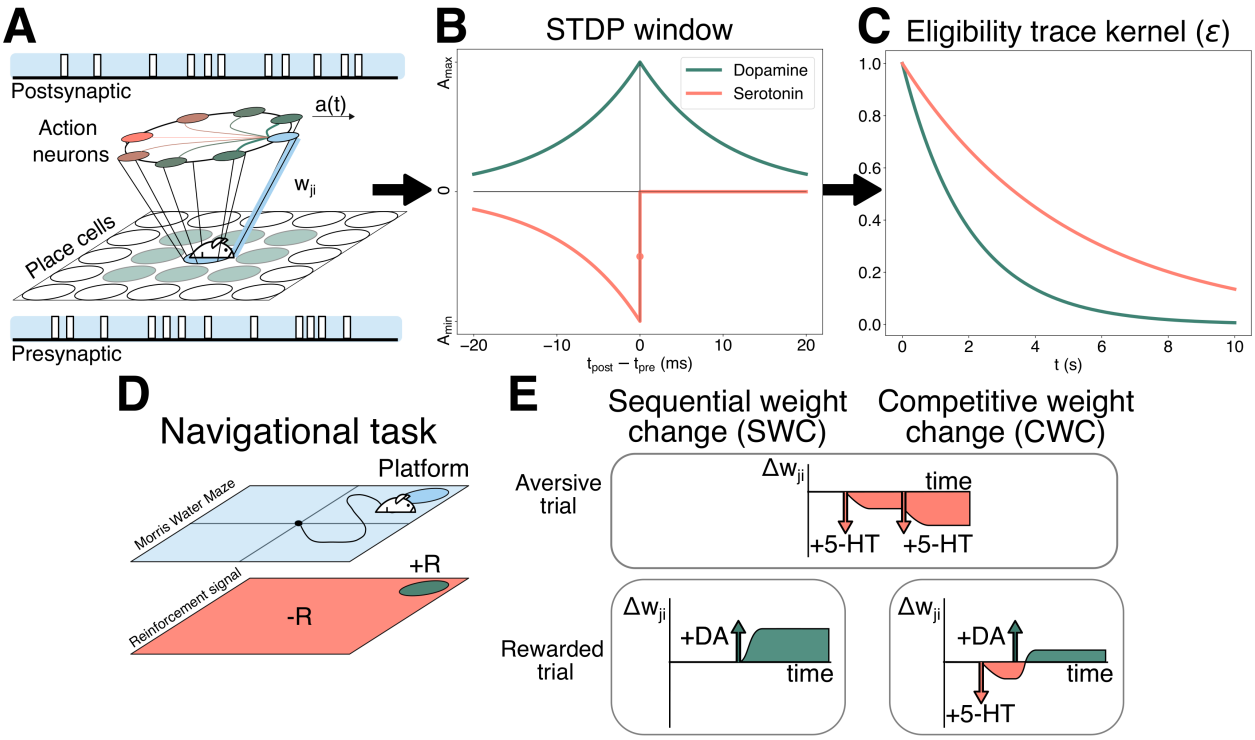

Figure 1: Schematics illustrating the learning model details used in the study. **(A)** The navigational model uses a feed-forward network between Gaussian receptive fields and neurons performing action selection $a(t)$ through "winner-takes-all" connectivity. Forward synaptic weights w_{ji} between neuron pairs (blue) are updated through a R-STDP rule. **(B)** STDP window used for DA and 5-HT. Both neuromodulators presented equal decays and opposite modulation with 5-HT presenting exclusively anti-causal depression. **(C)** Kernel of an exponentially decaying eligibility trace with biologically-derived time constants. **(D)** Reward representation derived from water aversion in the Morris water maze (MWM) task. Negative rewards are linked to unsuccessful trials, whereas dopaminergic activation is achieved at platform arrival. **(E)** Strictly aversive trials are processed equally by SWC and CWC. Upon reaching a reward, weight updates for SWC are solely driven by DA. Thus, SWC ensures all connections are either potentiated or depressed at the end of a trial. In CWC, rewarded trials also include aversive cues leading up to the reward. For equations and specific values in **A-C** see Materials and methods.

reward site, and t-LTD, upon a punishing trial. This is equivalent to consider that DA dominates over 5-HT if the positive valence item is located, which nullifies the contribution of the stress-inducing cue. Conversely, unrewarded trials solely present 5-HT modulation which, in the case of the MWM, is performed at the end-of-trial. Additionally, in opposition to continuous depression (Zannone et al., 2018), we adapted SWC to reflect the presence of an eligibility trace for 5-HT (Materials and methods).

The second learning rule that we examined was competitive weight change (CWC), based on competitive reinforcement learning (Huertas et al., 2016; ***Figure 1D***). In contrast to SWC, eligibility traces perform in-trial opposition (***Figure S2***), defined mathematically as an addition of DA and 5-HT contributions (He et al., 2015; Materials and methods), in which the weight change is determined by the balance between reward-trace pairs. Hence, activated synaptic weights are not guaranteed to be either potentiated or depressed after the reinforcement signal is introduced (***Figure S3***). For some configurations, CWC may yield depressive effects upon reaching the platform if DA does not counterbalance 5-HT. Experimental data from serotonergic and dopaminergic neurons suggests differential activity, since 5-HT is tonically released as a reponse to a long-term punishment, and DA has a greater phasic response towards a rewarding event (Boureau and Dayan, 2011; Cohen et al., 2015). Hence, we represented transient 5-HT as a step function active until the positive valence item is found, which in turn triggers a one-second constant DA signal (Cohen et al., 2015). Consequently, for this navigational task, successful trials include both modulation by DA and 5-HT, whereas unrewarded ones involves exclusively the latter. Fundamentally, SWC and CWC diverge in the timing of the weight update, either end-of-trial or in-trial, and in the characterization of DA and 5-HT as alternate or additive.

Both models were systematically parametrized through grid search to optimize performance. SWC had a better efficiency than CWC in successful simulations over successive trials (***Figure 2A***) and accumulated successful episodes (***Figure 2B***). In both cases, the addition of 5-HT as a t-LTD inducer improved the rewarding outcomes, especially for CWC, which may worsen its learning efficiency with time for particular configurations (***Figure S4***). The poor performance of DA-only CWC can be explained by the fitting of the reward function and the parameters employed, which cause the saturation of weights from non-predictive paths. Compared with SWC, CWC does not optimize the path distance, as measured by the time to reach the reward (***Figure 2C***). Moreover, CWC increases the median distance to the center with time for both conditions (Gehring et al., 2015;***Figure 2D***). The systematic depression of synapses in central place cells, which causes the agent to move near the edges, can explain the apparent incoherence between the latency time and the median distance of CWC with 5-HT. Such a phenomenon is due to the dynamic competition between traces, which can produce depression in neurons activated early under certain conditions. In terms of exploration, we computed the Kullback–Leibler divergence (KL) between the first reward distribution of both conditions (Zannone et al., 2018). The divergence in CWC (KL(+5-HT|-5-HT) = 0.021) and SWC (KL(+5-HT|-5-HT) = 0.015) indicates exploration remains unaltered between both cases (***Figure S5***). For SWC, we also considered a constant action of the punishment, instead of a phasic rewarding response at the end of the trial, without improvements in performance (***Figure S6***). Likewise, in CWC, the efficiency was lowered when a complete phasic response was introduced (***Figure S7***). In conclusion, although limited by dynamical and encoding assumptions, SWC with 5-HT has a better performance than CWC and a dopamine-only learning rule.

The biological viability of DA and 5-HT modeling through SWC and CWC was assessed against a study by Teixeira et al. (2018), which showed that optogenetic inhibition and activation of serotonergic neurons modified learning abilities of mice without significantly affecting locomotion. We aimed at replicating these results by imposing three intervals of the simulation time increasing or omitting the serotonergic signal (***Figure 3A***), considered as a doubling and absence of the punishment, correspondingly. Notably, inhibition of 5-HT in CWC caused a performance reduction although overactivation yielded no variation (***Figure 3A.i***). In contrast, SWC worsened the number of rewarding trials under either condition (***Figure 3A.ii***).

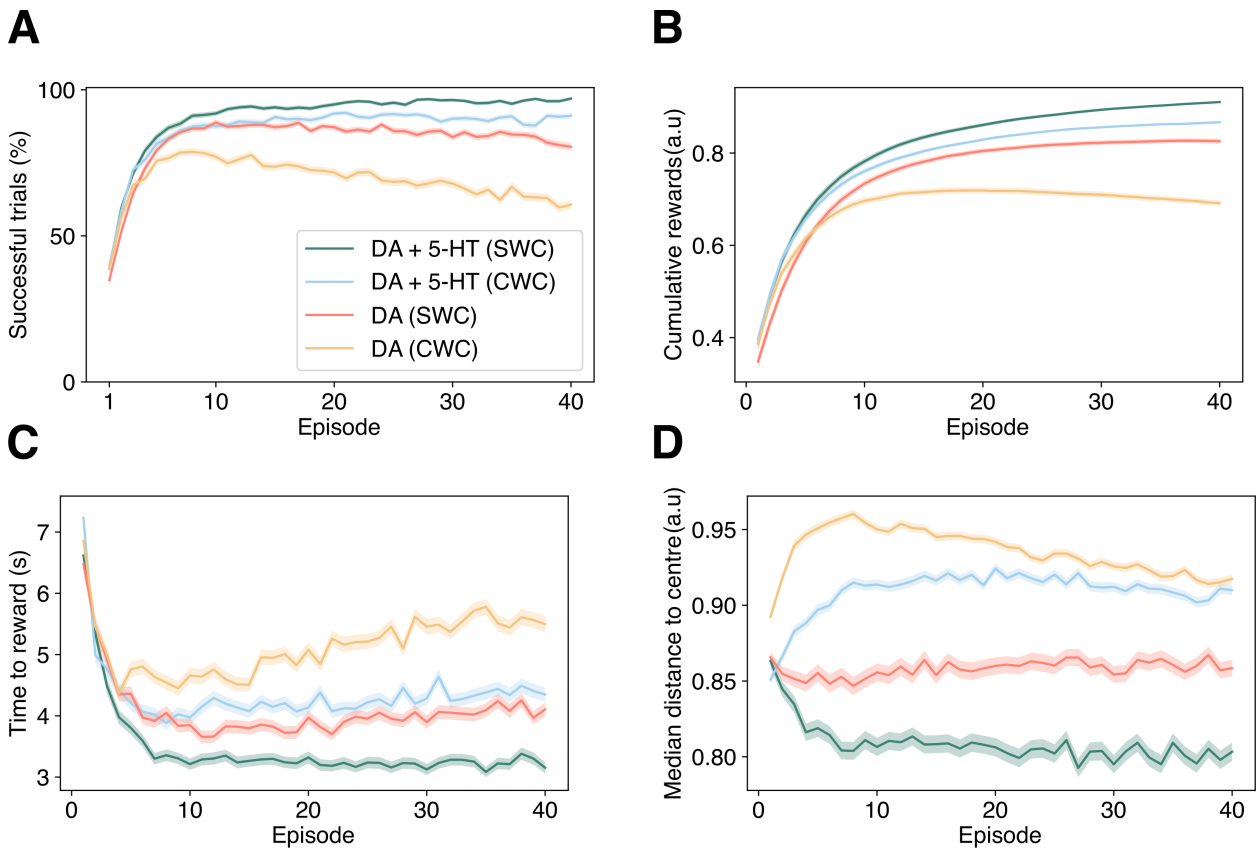

Figure 2: Inclusion of 5-HT improves learning under both R-STDP rules, with enhanced learning for SWC. **(A)** Learning curve of the percentage of successful simulations in each trial. The differences in means between the control (DA) and the addition of 5-HT (DA+5-HT)are significant on the last trial ($p < 0.01$, two-tailed Student's t-test). **(B)** Cumulative relative number of successful trials averaged over the simulations. **(C)** Average latency time to the reward in positive valence trials. Changes in the time to the reward for the final trial are significant between the different conditions in each rule ($p < 0.01$, two-tailed Student's t-test). **(D)** Average median distance to the center as measured in spatial memory tests. The shaded ranges correspond to the standard error of the mean (SEM) in M=1000 simulations. See Materials and methods for the value of parameters.

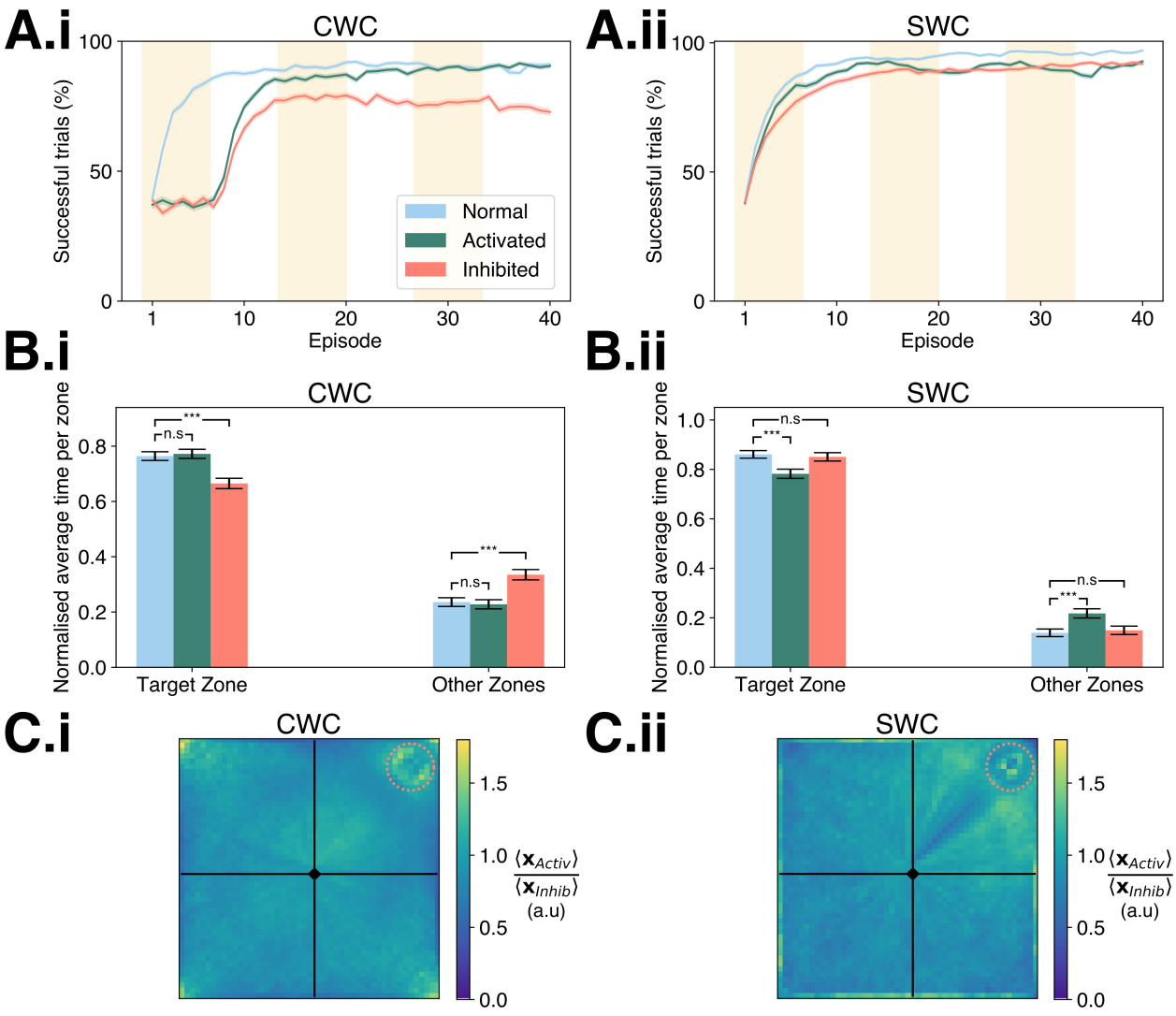

Figure 3: Overactivation and inhibition of 5-HT in CWC resembles behavioral data from optogenetic modulation of serotonergic neurons in the MWM. **(A)** Average percentage of successful simulations for (i) CWC and (ii) SWC. Episodes in yellow correspond to times in which optogenetic changes are introduced. **(B)** Bar plot of the average time in the target quadrant against the other zones for (i) CWC, which resembles real data observed in mice from Teixeira et al. (2018) and (ii) SWC. Statistical significance (two-sample Student's t-test with $p < 0.001$, ***) is shown for changes in some conditions under paired test . **(C)** Fold change between activated and inhibited averaged over time and simulation position histograms (50 bins per side) for **(i)** CWC and **(ii)** SWC. In a trial, the position is binned spatially then averaged across episodes and trials. The ratio of mean location between conditions is shown with the initial position (rhombus) and the reward location (dotted circle). Filled area and error bars in **A-B** correspond to SEM (M=1000).

These results highlight the importance of 5-HT as a compensatory mechanism of DA in CWC against a greater role in negative sampling for SWC. For both learning rules, sequential 5-HT inhibition decreased the residence time at the target platform in contrast with the control (***Figure 3B***), albeit only significantly for CWC (***Figure 3B.i***). Nevertheless, the increase in serotonergic response only resulted in a larger amount of time spent at the target zone for CWC (***Figures 3B.i-ii***). Hence, in time spent in the target quadrant, CWC replicated the changes observed empirically but SWC did not. A comparison between the position traces for activation and inhibition shows that 5-HT intensifies movement near the starting position and the corners of the maze in CWC (***Figure 3C.i***) and a more systematic navigation of the target quadrant in SWC (***Figure 3C.ii***), in which serotonergic amplification impedes the learning of the shortest path. Overall, these results suggests a greater biological feasibility of CWC in DA and 5-HT modeling. Nonetheless, additional evidence regarding the activity of serotonergic and dopaminergic neurons during individual trials could provide a more robust test for the model and corroborate the LTD contribution of 5-HT.

The described valence system has been evaluated in reversal learning, which involves punishment and reward switching (Matias et al., 2017). In this setting, t-LTP and t-LTD are disjointed (i.e., the agent is either rewarded or punished in the same trial), which reconciles SWC and CWC rules into the same system with phasic activity (***Figure S8***). As 5-HT operates under different timescales in reversal (Matias et al., 2017), we evaluated five learning rates to quantify the relative importance of t-LTD in relearning. For all rates, forward learning was successful (***Figure 4A***) although there was no recovery after inversion (***Figure 4B***), with high learning rates performing better. The lack of discrimination is observable in post-reversal synaptic weights compared to forward ones (***Figure 4C***). Additionally, negative valence simulations decreased for all learning rates (***Figure 4D***) although most simulations were neutral (***Figures 4E***), indicating the emergence of a non-decisive or metastable state in deliberation (Bakkour et al., 2019). Reduced selectivity is explained by low polarization in feed-forward weights, as measured by the coefficient of variation (***Figure 4F***), for all learning rates. Taken together, these results predict a role of 5-HT in aiding reversal learning and present testable conditions in open field navigation under rewarding and punishing cues. However, this model does not explain positive reward encoding of 5-HT observed in conditioning trials (Matias et al., 2017), which suggests a more complex interplay between neuromodulators in valence learning.

In summary, our plasticity model of interacting DA and 5-HT contributions provides mechanistic insights into their role in hippocampal-dependent spatial navigation. Future extensions of this model could explore combinations of aversive and attractive states coded at the level of circuits.

3 Materials and methods

Hippocampal-dependent spatial navigation is modeled through a one-layer network based on a navigational actor-critic system (Frémaux et al., 2013; Brzosko et al., 2017). Location, encoded through the spiking rate of place cells, serves as the input. The output layer is composed of action neurons, which determine the preferred movement of the agent by their firing rate.

3.1 Place cells

The spiking activity of place cells represents two-dimensional positional information. These are modeled through an inhomogeneous Poisson process with maximum spiking activity $\bar{\lambda}^{pc} = 400Hz$. The squared norm of the deviation between the Cartesian location of the agent $\mathbf{x}(t)$ and the center of the place cell i is used for the calculation of the rate as follows:

$$\lambda_i^{pc}(\mathbf{x}(t)) = \bar{\lambda}^{pc} \exp\left(-\frac{\|\mathbf{x}(t) - \mathbf{x}_i\|^2}{\sigma^2}\right), \tag{1}$$

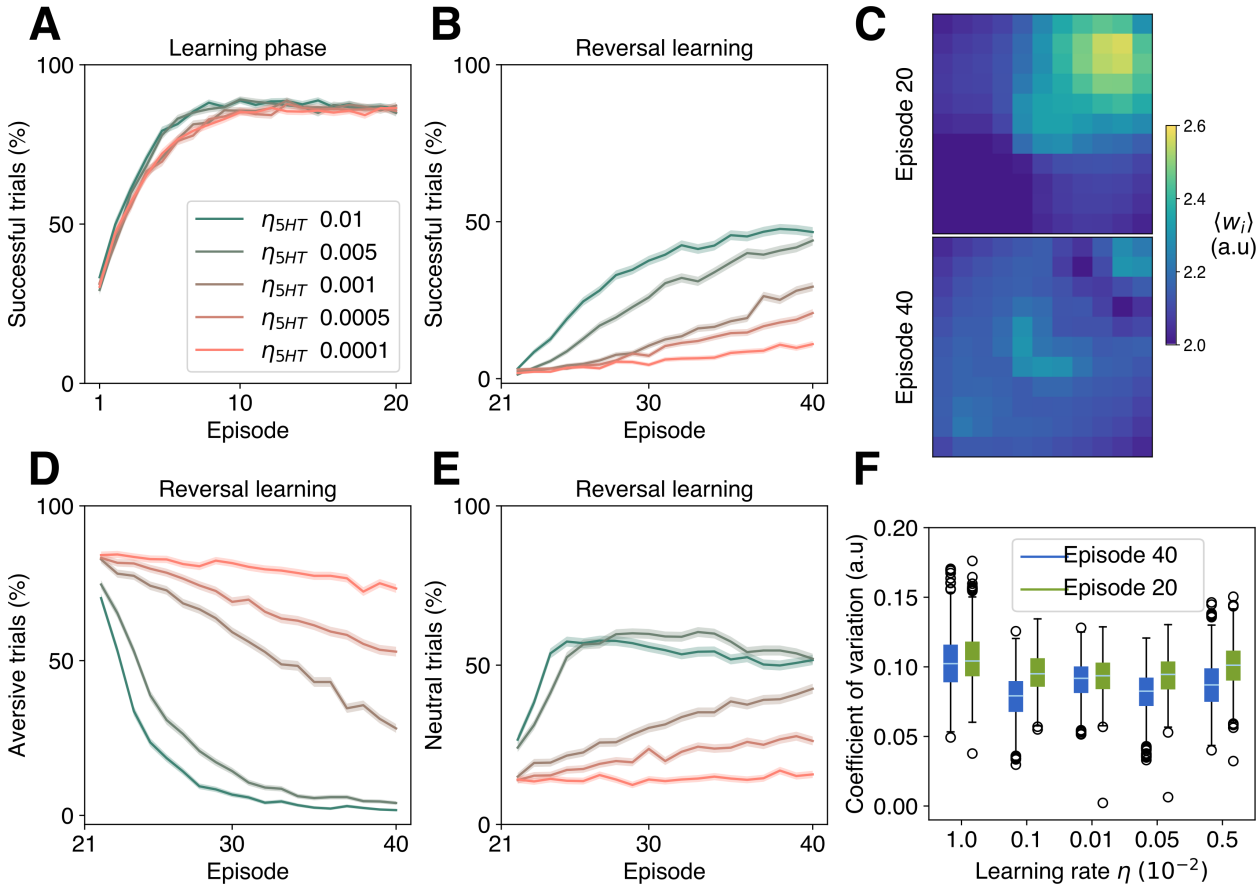

Figure 4: Reversal learning in an open field improves under 5-HT modulation for different learning rates with a high preference for a neutral state. **(A)** Learning curve depicted as the average percentage of successful simulations per trial for five 5-HT learning rates before reversal. **(B)** Average successful simulations after reversal. **(C)** Average place cell weights before (trial 20) and after reversal (trial 40) for $\eta_{5HT} = 0.01$. **(D)** Average punishing or negative simulations after reversal. **(E)** Average latency time to the reward in successful simulations. **(F)** Distribution of the mean coefficient of variation (CV) of synaptic weights before and after inversion. This corresponds to the ratio of the standard deviation to the mean of weights in each place cell. Lower CV values after inversion imply a decrease in dispersion or polarization of synaptic weights, as shown in **(C)**. In curves, the shaded region corresponds to SEM (M=1000).

The firing rate of the Poisson process exponentially decays with the Euclidean distance between the agent and the center of the place cell. A total of 121 place cells, equally separated by a distance of $\sigma = 0.4$ a.u, were distributed on a square of length side 4 a.u (Brzosko et al., 2017).

3.2 Action neurons

To model action neurons, a zero-order Spike Response Model (SRM_0) was used (Gerstner, 1995) , in which the membrane potential is represented as

$$u_j(t) = \sum_j w_{ji}^{feed} \sum_{t_i^f \in \mathcal{F}_i, t_i^f > \hat{t_j}} \epsilon(t - t_i^f) + \sum_{k,k\neq j} w_{jk}^{lat} \sum_{t_k^f \in \mathcal{F}_k, t_k^f > \hat{t_j}} \epsilon(t - t_k^f) + \chi\Theta(t - \hat{t_j}) \exp\left(\frac{\hat{t_j} - t}{\tau_m}\right). \quad (2)$$

In the feed-forward network, the action neuron j receives an excitatory postsynaptic potential (EPSP) $\epsilon(t - t_i^f)$, from place cell i, for firing times in the set $\mathcal{F}_i$, after the last spike of action neuron $\hat{t_j}$ and under a synaptic efficiency w_{ji}^{feed}. Similarly, action neurons k, of the lateral connectivity network, are connected with synaptic weights w_{jk}^{lat} and their spike arrival times are contained in set $\mathcal{F}_k$. The EPSP kernel is

$$\epsilon(t) = \frac{\epsilon_0}{\tau_m - \tau_s}(e^{\frac{-t}{\tau_m}} - e^{\frac{-t}{\tau_s}})\Theta(t), \quad (3)$$

where $\Theta(t)$ is the Heaviside step function, the membrane constant $\tau_m = 20$ ms and the rising time $\tau_s = 5$ ms. Eq. 2 considers a scale factor for the refractory effect $\chi = -5$mV and Eq. 3 as well with $\epsilon_0 = 20$ mV·ms.

The spiking activity is determined by an inhomogeneous Poisson process with rate $\lambda_j(u_j(t))$, which is formulated as

$$\lambda_j(u_j(t)) = \lambda_0 \exp\left(\frac{u_j(t) - \theta}{\Delta u}\right), \quad (4)$$

in which the maximum rate is $\lambda_0 = 60$ Hz, the potential threshold is $\theta = 16$ mV and Δu is a voltage window for spike emission that determines the degree of randomness. The dynamics are simplified if the resting potential is assumed to be 0V (Zannone et al., 2018).

The instantaneous firing rate of an action neuron $\rho_j(t)$ is obtained by filtering the spiking activity $Y_j = \sum_{t_j^f \in \mathcal{F}_j} \delta(t - t_j)$ with the kernel $\gamma(t) = \frac{e^{\frac{-t}{\tau_\gamma}} - e^{\frac{-t}{\nu_\gamma}}}{\tau_\gamma - \nu_\gamma}\Theta(t)$, where $\tau_\gamma = 50$ms and $\nu_\gamma = 20$ ms.

Each action neuron k represents a preferred direction of movement $\mathbf{a}_k$ and interacts with other angle-encoding action neurons through a lateral connectivity (Frémaux et al., 2013). The lateral synaptic weight dynamics produce a "N-winner-takes-all" arrangement by which N_{action} neurons compete for the preferential angle. Hence, the connectivity between neural units k and k' is modeled to inhibit opposite directions and excite similar ones as

$$w_{kk'} = \frac{w_-}{N_{action}} + w_+ \frac{f(k, k')}{\sum_{k'} f(k, k')}, \quad (5)$$

inhibitory and excitatory weights are $w_- = -300$ and $w_+ = 100$ and f is the lateral connectivity function, which reaches a maximum for $k = k' \pm 1$ and decreases monotonically towards zero for $k = k'$. Concretely, in this case, $f(k, k') = (1 - \delta_{k,k'}) \exp(\zeta \cos(\theta_k - \theta_{k'}))$ decreases exponentially for increasingly dissimilar angles $\theta_{k'}$ and is scaled by a factor $\zeta = 20$. These parameters were tuned for a population of $N_{action} = 40$ with $\theta_k = \frac{2k\pi}{N_{actor}}$ (Zannone et al., 2018). Thus, action vectors were $\mathbf{a}_k = a_0(\sin(\theta_k), \cos(\theta_k))^T$ with $a_0 = 0.08$.

The action resulting from the spiking activity of the network is coded through a population vector as

$$\mathbf{a}(t) = \frac{1}{N_{action}} \sum_k \rho_k(t)\mathbf{a}_k, \quad (6)$$

which is weighted by the filtered spiking activity of neurons $\rho_k(t) = (Y_k \circ \gamma)(t)$. Thus, the action at each time step $\mathbf{a}(t)$ is computed as the average of the action vectors with the predicted instantaneous activity of actor neurons. The inertia of movement is determined by the activity of the network with the maximum velocity being limited by a_0.

3.3 Navigational setting

The square S delimiting the two-dimensional plane, serves as a boundary condition for the position $\mathbf{x}(t)$ and the movement of the agent $\mathbf{a}(t)$. This is formulated as

$$\Delta\mathbf{x}(t) = \begin{cases} \mathbf{a}(t) & \forall \mathbf{x}(t+1) \in S \mid S = \{[-2,2] \times [-2,2]\} \\ \mathbf{d}(t) & \text{else.} \end{cases} \tag{7}$$

The bouncing vector $\mathbf{d}(t)$ corresponds to the displacement of the agent in the direction of the normal vector to S or $\mathbf{n}_S$, which is defined as $\mathbf{d}(t) = d_0 \mathbf{n}_S(\mathbf{x}(t))$. The bouncing distance is set as $d_0 = 0.01$.

In the Morris water maze, the rewarding platform was positioned at $r_c = (1.5, 1.5)$ with a radius $r_r = 0.3$. In reversal learning, the reward is initially maintained at its position and the punishment is placed at $p_c = (-1.5, -1.5)$ with radius $p_r = 0.3$, for trials 1 to 20. After reversal, both elements switch position, $r_c = (-1.5, -1.5)$ and $p_c = (1.5, 1.5)$, with unvaried radii, for trials 21 to 40. In all instances, the initial position of the agent was $\mathbf{x}(t = 0) = (0, 0)$ and the maximum trial time was $T_{max} = 15$ s (Zannone et al., 2018). If the reward or, in reversal learning, punishment is reached before T_{max}, the trial is ended, and place cells are deactivated. In sequential weight change, the weight update occurs at $t = T_{rew} + 300$ ms, to replicate consummatory behavior. Per contra, in competitive weight change, weights are updated continuously until the DA signal is no longer active $t = T_{rew} + T_{DA}$. Activity is reset between trials.

3.4 Sequential weight change (SWC)

The sequentially neuromodulated plasticity (sn-Plast) rule from Brzosko et al. (2017) was adapted to match the empirical evidence available for 5-HT in He et al. (2015). Hence, instead of presenting an online depression mediated by acetylcholine, the adjusted SWC update introduces an eligibility trace ϵ_{5HT} for the depressor, which permits the temporal discrimination of neural activity leading up to an aversive cue.

The weight update is determined by the spike-timing-dependent plasticity (STDP) window of each neuromodulator $W(s)$, filtered by its eligibility trace kernel ϵ, and an outcome-dependent signal A that . As such,

$$\Delta w_{ji}(t) = \eta A((\sum_{t_i^f \in \mathcal{F}_i^{pc}} \sum_{t_j^f \in \mathcal{F}_j^a} W(t_j^f - t_i^f)) \circ \epsilon)(t), \tag{8}$$

with differentiated learning rates η for 5-HT and DA $\eta = \begin{cases} \eta_{DA} & +\text{DA} \\ \eta_{5HT} & +\text{5-HT} \end{cases}$. SWC assumes either DA or 5-HT act at the end-of-trial. Thus, in the MWM, DA is released if the agent reaches the reward while, in an unrewarded trial, water aversiveness causes 5-HT stimulation. This representation holds for reversal learning, outcomes are exclusive as well. Hence, the valence signal A is conditioned by whether the trial was rewarding or punishing and it is defined as $A = \begin{cases} 1 & +\text{DA} \\ -1 & +\text{5-HT} \end{cases}$. The firing times t_j^f and t_i^f of action neuron j and place cell i are contained in the sets $\mathcal{F}_j^a$ and $\mathcal{F}_i^{pc}$.

The STDP window for DA was preserved as $W_{DA}(s) = e^{\frac{-s}{\tau}}$, where $\tau = 10$ ms. However, for 5-HT (He et al., 2015), it was transformed into an asymmetric STDP curve as

$$W_{5HT}(s) = \begin{cases} A_{5HT} e^{\frac{-s}{\tau}} & \text{if } s > 0 \\ \frac{1}{2} A_{5HT} & \text{if } s = 0 \\ 0 & \text{if } s < 0. \end{cases} \tag{9}$$

The eligibility trace kernel for 5-HT and DA is formulated as

$$\epsilon(t) = e^{\frac{t}{\tau_e}} \Theta(t), \tag{10}$$

with $\tau_e = \begin{cases} 2 \text{ s} & \text{for DA} \\ 5 \text{ s} & \text{for 5-HT} \end{cases}$. The value of the time constant for 5-HT comes from *in vivo* experiments (He et al., 2015). The decay disparity among traces elicits a differential response between predictive neural "trails". In other terms, place cells encoding for the same path suffer a greater weight change, in absolute terms, through serotonergic action than with DA modulation.

3.5 Competitive weight change (CWC)

In competitive weight change (CWC), inspired by competitive reinforcement learning (CRL) from Huertas et al. (2016), the eligibility traces, or proto-weights, of the two modulators are under a dynamic competition for the upgrade of the synaptic weight. Proto-weights T are the result of filtering the Hebbian term $W(s)$ for each neuromodulator with the eligibility trace kernels ϵ defined in Eq. 10. The equation results in

$$T(t) = (W(s) \circ \epsilon)(t), \tag{11}$$

with the STDP windows for DA and 5-HT defined as in SWC. In integral terms, and for a particular connection between neurons j and i, Eq. 11 becomes $T_{ji}(t) = ((\sum_{t_i^f \in \mathcal{F}_i^{pc}} \sum_{t_j^f \in \mathcal{F}_j^a} W(t_j^f - t_i^f)) \circ \epsilon)(t)$.

The weight uptade is expressed as the dynamic competition between traces, which corresponds to the difference between t-LTP and t-LTD proto-weights or

$$\dot{w_{ji}}(t) = \eta(R_{DA}(t)T_{DA} - R_{5HT}(t)T_{5HT}), \tag{12}$$

where T_{DA} and T_{5HT} are the proto-weights of the neurotransmitters and R_{DA} and R_{5HT} models the reinforcement response for each neuromodulator with learning rate η. Hence, the condition for t-LTP or t-LTD becomes

$$\begin{aligned} R_{DA}(t)T_{DA} > R_{5HT}(t)T_{5HT} &\implies \dot{w_{ji}}(t) > 0 \; (LTP) \\ R_{DA}(t)T_{DA} < R_{5HT}(t)T_{5HT} &\implies \dot{w_{ji}}(t) < 0 \; (LTD) \end{aligned} \tag{13}$$

The integral version of Eq. 12 is,

$$\Delta w_{ji}(t) = \eta\Big(R_{DA}(t)((\sum_{t_i^f \in \mathcal{F}_i^{pc}} \sum_{t_j^f \in \mathcal{F}_j^a} W_{DA}(t_j^f - t_i^f)) \circ \epsilon_{DA})(t) - R_{5HT}(t)((\sum_{t_i^f \in \mathcal{F}_i^{pc}} \sum_{t_j^f \in \mathcal{F}_j^a} W_{5HT}(t_j^f - t_i^f)) \circ \epsilon_{5HT})(t)\Big), \tag{14}$$

where the reinforcement signals are modeled as Heaviside step functions $\Theta(t)$. Specifically, $R_{DA}(t) = R_{DA}(\Theta(t - T_{rew}) - \Theta(t - T_{DA} - T_{rew}))$ and $R_{5HT}(t) = R_{5TH}(\Theta(t) - \Theta(t - T_{rew}))$. Accordingly, 5-HT is active until the reward is attained or the end of the trial at time, if the agents has been unsuccessful,

which corresponds to time T_{rew}. DA acts for a time after the arrival of the agent to the positive reinforcement site, which was assumed to be $T_{DA} = 1$ s in accordance with biological studies involving long-term observation of the dopamine-serotonin interplay (Cohen et al., 2015).

The definition from Eq. 12 In opposition to the formulation by Huertas et al. (2016), the eligibility traces of LTP and LTD were not such that could be used for the representation of cues through their dynamic equilibrium. Alternatively, at the end of the simulation, the potentiation or depression of flagged synapses (i.e., neural connections that have been active during the trial) is not assured and may not be consistent with the outcome of the trial. Notably, in our MWM implementation, the condition for weight potentiation after reaching the platform requires that

$$\Delta w_{ji}(t = T_{trial}) > 0 \iff \int_0^{T_{trial}} R_{DA}(t)T_{DA}(t) > \int_0^{T_{trial}} R_{5HT}(t)T_{5HT}(t), \tag{15}$$

in which T is the proto-weight of each neuromodulator and R the reinforcement signal. It should be noted that this equation depends on the time of the trial T_{trial} and, therefore, the sign of weight update values can vary between trials of different duration. This is equivalent, in integral terms, to Eq. 13.

3.6 Parameter values

The configuration of each model was optimized through grid search parametrization. In particular, we optimized the amplitudes of the STDP widows, A_{DA} and A_{5HT}, the reward magnitudes for CWC R_{DA} and R_{5HT}, and the learning rates η. Sweeps of these values were first conducted by orders of magnitude and then with fine tuning around good estimates. The best models were selected by the proportion of successful simulations at the final trial.

Table 1: Parameters used for SWC and CWC models. These parameters correspond to the best models with regard to the percentage of successful simulations in the last trial.

Model	A_{DA}	A_{5HT}	η	w_{min}	w_{max}	R_{DA}	R_{5HT}	τ_{STDP}	$\tau_\epsilon(DA)$	$\tau_\epsilon(5HT)$
CWC	1	0.01	0.0001	1	3	1	1	10 ms	2 ms	5 ms
SWC	1	1	0.01	1	3	-	-	10 ms	2 ms	5 ms

The code, written in Python, will be available after publication in ModelDB.

4 Acknowledgments

This work was funded by BBSRC (BB/N013956/1 and BB/N019008/1), EPSRC (EP/R035806/1), German Research Foundation (CRC 1089), "la Caixa" Foundation (LCF/BQ/EU19/11710071), Max Planck Society, Simons Foundation (564408) and Wellcome Trust (200790/Z/16/Z).

References

Bakkour A, Palombo DJ, Zylberberg A, Kang YH, Reid A, Verfaellie M, Shadlen MN, Shohamy D. The hippocampus supports deliberation during value-based decisions. elife. 2019; 8:e46080.

Barnes NM, Sharp T. A review of central 5-HT receptors and their function. Neuropharmacology. 1999; 38(8):1083–1152. doi: 10.1016/S0028-3908(99)00010-6.

Berumen LC, Rodríguez A, Miledi R, García-Alcocer G. Serotonin receptors in hippocampus. The Scientific World Journal. 2012; 2012. doi: 10.1100/2012/823493.

Boureau YL, Dayan P. Opponency revisited: Competition and cooperation between dopamine and serotonin. Neuropsychopharmacology. 2011; 36(1):74–97. doi: 10.1038/npp.2010.151.

Brzosko Z, Schultz W, Paulsen O. Retroactive modulation of spike timingdependent plasticity by dopamine. eLife. 2015; 4(OCTOBER2015):e09685. doi: 10.7554/eLife.09685.

Brzosko Z, Zannone S, Schultz W, Clopath C, Paulsen O. Sequential neuromodulation of hebbian plasticity offers mechanism for effective reward-based navigation. eLife. 2017; 6:e27756. doi: 10.7554/eLife.27756.

Cohen JY, Amoroso MW, Uchida N. Serotonergic neurons signal reward and punishment on multiple timescales. eLife. 2015; 2015(4):e06346. doi: 10.7554/eLife.06346.

Cools R, Nakamura K, Daw ND. Serotonin and dopamine: Unifying affective, activational, and decision functions. Neuropsychopharmacology. 2011; 36(1):98–113. doi: 10.1038/npp.2010.121.

Crockett MJ, Clark L, Apergis-Schoute AM, Morein-Zamir S, Robbins TW. Serotonin modulates the effects of pavlovian aversive predictions on response vigor. Neuropsychopharmacology. 2012; 37(10):2244–2252. doi: 10.1038/npp.2012.75.

Dalley JW, Roiser JP. Dopamine, serotonin and impulsivity. Neuroscience. 2012; 215:42–58. doi: 10.1016/j.neuroscience.2012.03.065.

Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Networks. 2002; 15(4-6):603–616. doi: 10.1016/S0893-6080(02)00052-7.

Dayan P, Huys Q. Serotonin's many meanings elude simple theories. eLife. 2015; 4:e07390. doi: 10.7554/elife.07390.

Fernandez SP, Muzerelle A, Scotto-Lomassese S, Barik J, Gruart A, Delgado-García JM, Gaspar P. Constitutive and Acquired Serotonin Deficiency Alters Memory and Hippocampal Synaptic Plasticity. Neuropsychopharmacology. 2017; 42(2):512–523. doi: 10.1038/npp.2016.134.

Fischer AG, Ullsperger M. An update on the role of serotonin and its interplay with dopamine for reward. Frontiers in Human Neuroscience. 2017; 11:484. doi: 10.3389/fnhum.2017.00484.

Foster DJ, Morris RG, Dayan P. A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus. 2000; 10(1):1–16.

Frémaux N, Gerstner W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. Frontiers in Neural Circuits. 2015; 9(JAN2016):85. doi: 10.3389/fncir.2015.00085.

Frémaux N, Sprekeler H, Gerstner W. Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons. PLoS Computational Biology. 2013; 9(4):e1003024. doi: 10.1371/journal.pcbi.1003024.

Gehring TV, Luksys G, Sandi C, Vasilaki E. Detailed classification of swimming paths in the Morris Water Maze: Multiple strategies within one trial. Scientific Reports. 2015; 5:14562. doi: 10.1038/srep14562.

Gerstner W. Time structure of the activity in neural network models. Physical Review E. 1995; 51(1):738. doi: 10.1103/PhysRevE.51.738.

Gerstner W, Lehmann M, Liakoni V, Corneil D, Brea J. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Frontiers in Neural Circuits. 2018; 12:53. doi: 10.3389/fncir.2018.00053.

Hagena H, Manahan-Vaughan D. The serotonergic 5-HT4 receptor: a unique modulator of hippocampal synaptic information processing and cognition. Neurobiology of learning and memory. 2017; 138:145–153.

Harrison FE, Hosseini AH, McDonald MP. Endogenous anxiety and stress responses in water maze and Barnes maze spatial memory tasks. Behavioural Brain Research. 2009; 198(1):247–251. doi: 10.1016/j.bbr.2008.10.015.

He K, Huertas M, Hong SZ, Tie XX, Hell JW, Shouval H, Kirkwood A. Distinct Eligibility Traces for LTP and LTD in Cortical Synapses. Neuron. 2015; 88(3):528–538. doi: 10.1016/j.neuron.2015.09.037.

Huertas MA, Schwettmann SE, Shouval HZ. The role of multiple neuromodulators in reinforcement learning that is based on competition between eligibility traces. Frontiers in Synaptic Neuroscience. 2016; 8(DEC):37. doi: 10.3389/fnsyn.2016.00037.

Karabeg MM, Grauthoff S, Kollert SY, Weidner M, Heiming RS, Jansen F, Popp S, Kaiser S, Lesch KP, Sachser N, Schmitt AG, Lewejohann L. 5-HTT Deficiency Affects Neuroplasticity and Increases Stress Sensitivity Resulting in Altered Spatial Learning Performance in the Morris Water Maze but Not in the Barnes Maze. PLOS ONE. 2013 10; 8(10). `https://doi.org/10.1371/journal.pone.0078238`, doi: 10.1371/journal.pone.0078238.

Kemp A, Manahan-Vaughan D. Hippocampal long-term depression and long-term potentiation encode different aspects of novelty acquisition. Proceedings of the National Academy of Sciences. 2004; 101(21):8192–8197. `https://www.pnas.org/content/101/21/8192`, doi: 10.1073/pnas.0402650101.

Kemp A, Manahan-Vaughan D. The 5-hydroxytryptamine4 receptor exhibits frequency-dependent properties in synaptic plasticity and behavioural metaplasticity in the hippocampal CA1 region in vivo. Cerebral Cortex. 2005; 15(7):1037–1043. doi: 10.1093/cercor/bhh204.

Lecouflet P, Roux CM, Potier B, Leger M, Brunet E, Billard JM, Schumann-Bard P, Freret T. Interplay between 5-HT4 Receptors and GABAergic System within CA1 Hippocampal Synaptic Plasticity. Cerebral Cortex. 2020 09; 31(1):694–701. `https://doi.org/10.1093/cercor/bhaa253`, doi: 10.1093/cercor/bhaa253.

Matias S, Lottem E, Dugué GP, Mainen ZF. Activity patterns of serotonin neurons underlying cognitive flexibility. eLife. 2017; 6:e20552. doi: 10.7554/eLife.20552.

O'Keefe J, Dostrovsky J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research. 1971; 34(1):171–175. doi: 10.1016/0006-8993(71)90358-1.

O'Keefe J, Nadel L. The Hippocampus as a Cognitive Map. Oxford: Clarendon Press; 1978. `http://hdl.handle.net/10150/620894`.

Palacios-Filardo J, Mellor JR. Neuromodulation of hippocampal long-term synaptic plasticity. Current Opinion in Neurobiology. 2019; 54:37–43. doi: 10.1016/j.conb.2018.08.009.

Rogers RD. The roles of dopamine and serotonin in decision making: Evidence from pharmacological experiments in humans. Neuropsychopharmacology. 2011; 36(1):114–132. doi: 10.1038/npp.2010.165.

Saylor RA, Hersey M, West A, Buchanan AM, Berger SN, Nijhout HF, Reed MC, Best J, Hashemi P. In vivo hippocampal serotonin dynamics in male and female mice: Determining effects of acute escitalopram using fast scan cyclic voltammetry. Frontiers in Neuroscience. 2019; 13:362. doi: 10.3389/fnins.2019.00362.

Schmidt SD, Furini CRG, Zinn CG, Cavalcante LE, Ferreira FF, Behling JAK, Myskiw JC, Izquierdo I. Modulation of the consolidation and reconsolidation of fear memory by three different serotonin receptors in hippocampus. Neurobiology of Learning and Memory. 2017; 142:48–54. doi: 10.1016/j.nlm.2016.12.017.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275(5306):1593–1599. doi: 10.1126/science.275.5306.1593.

Sutton RS, Barto AG. Reinforcement learning: An introduction. Cambridge, MA, USA: MIT press; 2018.

Teixeira CM, Rosen ZB, Suri D, Sun Q, Hersh M, Sargin D, Dincheva I, Morgan AA, Spivack S, Krok AC, Hirschfeld-Stoler T, Lambe EK, Siegelbaum SA, Ansorge MS. Hippocampal 5-HT Input Regulates Memory Formation and Schaffer Collateral Excitation. Neuron. 2018; 98(5):992–1004. doi: 10.1016/j.neuron.2018.04.030.

Vasilaki E, Frémaux N, Urbanczik R, Senn W, Gerstner W. Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. PLoS Comput Biol. 2009; 5(12):e1000586.

Vorhees CV, Williams MT. Morris water maze: Procedures for assessing spatial and related forms of learning and memory. Nature Protocols. 2006; 1(2):848–858. doi: 10.1038/nprot.2006.116.

Waider J, Popp S, Mlinar B, Montalbano A, Bonfiglio F, Aboagye B, Thuy E, Kern R, Thiel C, Araragi N, Svirin E, Schmitt-Böhrer AG, Corradetti R, Lowry CA, Lesch KP. Serotonin deficiency increases context-dependent fear learning through modulation of hippocampal activity. Frontiers in Neuroscience. 2019; 13:245. doi: 10.3389/fnins.2019.00245.

Wang RY, Arvanov VL. M100907, a highly selective 5-HT2A receptor antagonist and a potential atypical antipsychotic drug, facilitates induction of long-term potentiation in area CA1 of the rat hippocampal slice. Brain research. 1998; 779(1-2):309–313.

Wawra M, Fidzinski P, Heinemann U, Mody I, Behr J. 5-HT4-receptors modulate induction of long-term depression but not potentiation at hippocampal output synapses in acute rat brain slices. PLoS ONE. 2014; 9(2):e88085. doi: 10.1371/journal.pone.0088085.

Zannone S, Brzosko Z, Paulsen O, Clopath C. Acetylcholine-modulated plasticity in reward-driven navigation: a computational study. Scientific Reports. 2018; 8(1):1–20. doi: 10.1038/s41598-018-27393-2.

Supplemental Information

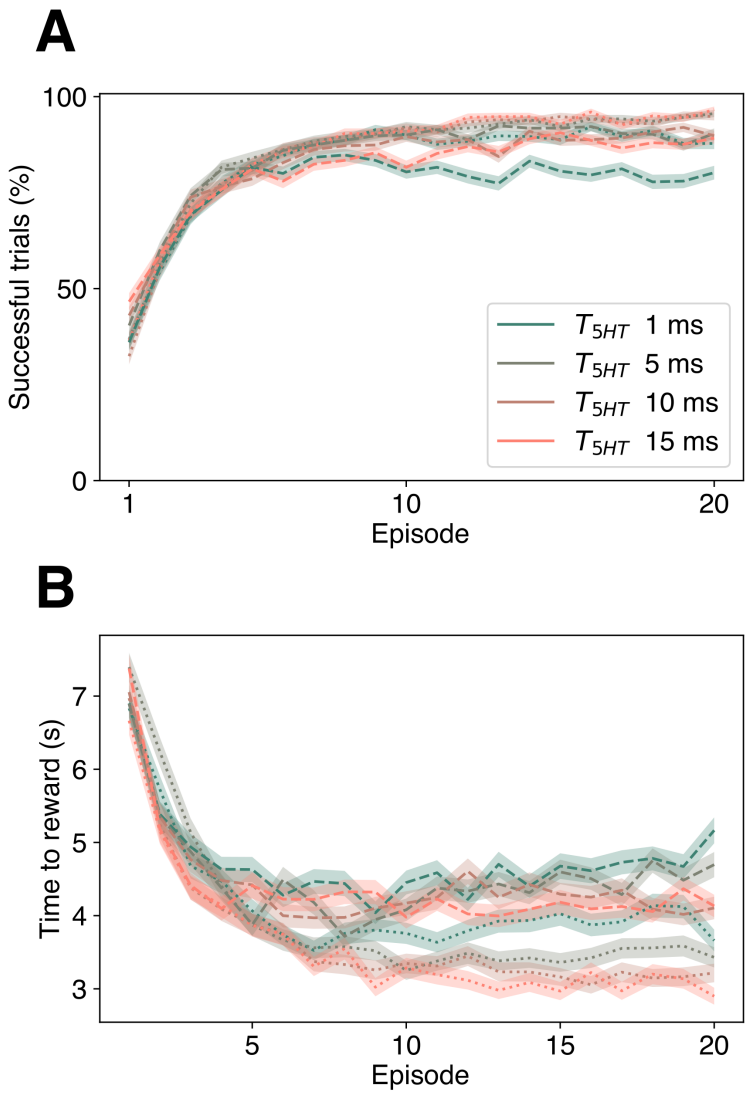

Figure S1: STDP decay value of 5-HT does not hinder learning. **(A)** Learning curve depicted as the percentage of successful trials along episodes. Neither SWC (dotted) nor CWC (dashed) altered significantly their performance for different decays of the exponential kernel of the STDP window (Eq. 9, Materials and methods). **(B)** Latency time. Greater variation can be seen among the times to reach the reward and shortest path optimization. Higher time constants for 5-HT-dependent STDP facilitate greater distance minimization. Filled areas correspond to SEM (M=1000).

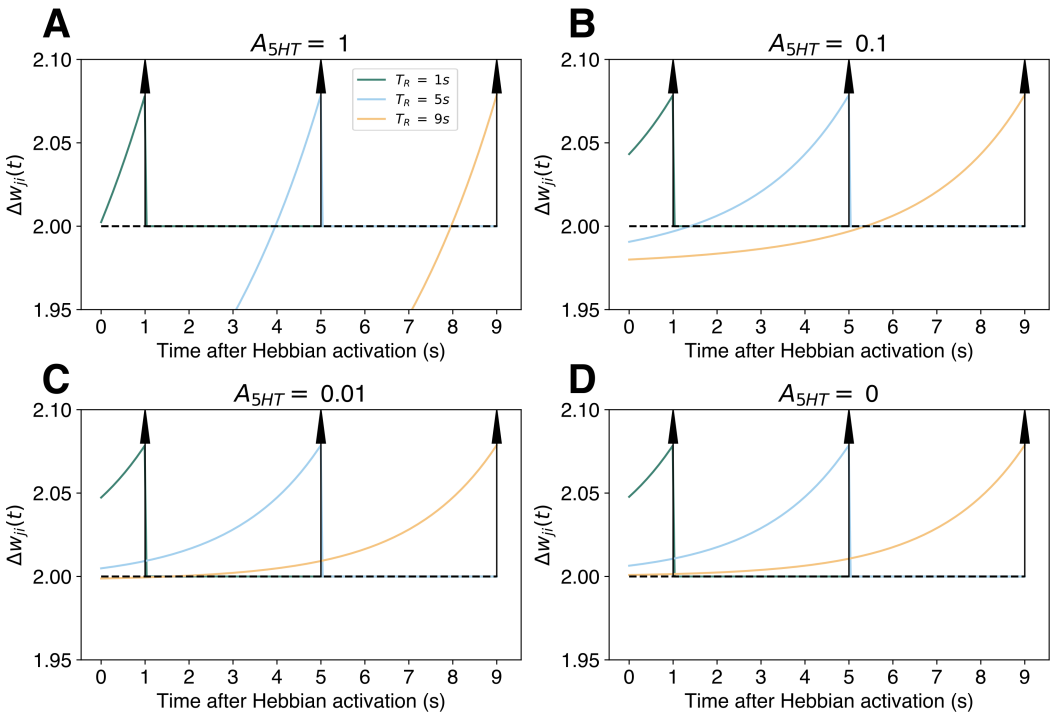

Figure S2: SWC performs sign-consistent updates across all weights. Weight change as a function of the time after a Hebbian activation $W(t_j^f - t_i^f)$ (Eq. 8), modeled as a Dirac delta function at $t = 0$ or $\delta(t)$. Three different rewarding times T_R, at 1s, 5s and 9s, and a punishment at the end of the episode are shown (arrows). With the weight update $\Delta w_{ji}(t)$ for each time after Hebbian activation plotted. The parameters were taken as specified in Materials and methods (Table 1) with the variation of the magnitude of the STDP window of 5-HT. Four different values **A-D** for serotonergic modulation are used (i.e. $H_{5HT}(t) = A_{5HT}\delta(t - t_{activation})$.

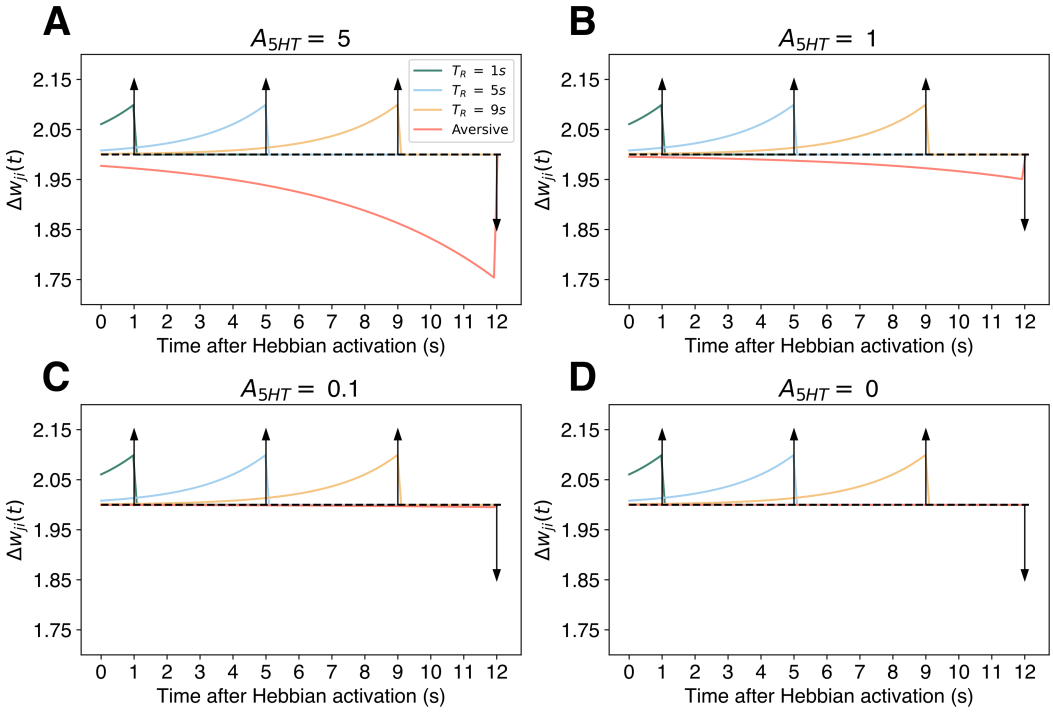

Figure S3: Potentiation of positive valence trials in CWC is not guaranteed for all neurons leading up to the reward. This case represents CWC for serotonin and dopamine with a Dirac delta function at $t = 0$ or $\delta(t)$ as the Hebbian term. Three rewarding time points T_R are shown (vertical arrows), which denote the time at which serotonergic and dopaminergic responses are switched (i.e. $R_{DA}(t) = \delta(t - T_R)$). The weight update $\Delta w_{ji}(t)$ is plotted against the time after the Hebbian activation. As described (Materials and methods), We maintained dopaminergic response for 1s as consummatory behavior. Four different values (**A-D**) for serotonergic modulation are used (i.e. $H_{5HT}(t) = A_{5HT}\delta(t - t_{activation})$). Depression occurs for all ranges of modulation (**A-C**) except for $A_{5HT} = 0$ (**D**).

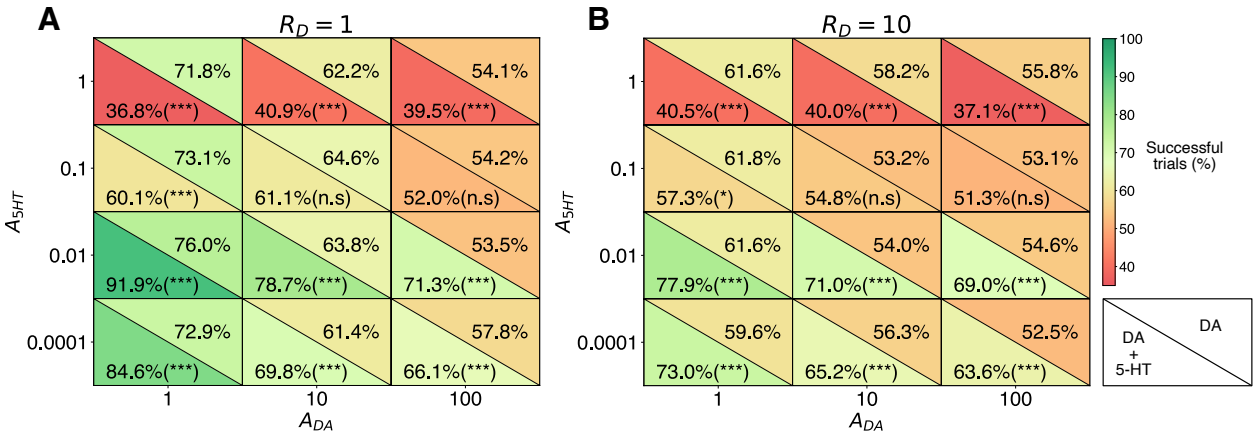

Figure S4: Manual search of the amplitudes of the STDP window for DA and 5-HT. Scores at each amplitude tuple correspond to the mean percentage of successful trials at episode 20 (M=1000) for runs with DA and 5-HT (bottom left) and DA-only (upper right). Two values of $R_D = \frac{R_{DA}}{R_{5-HT}}$ were used (Eq. 14, Materials and methods). Equal reward functions in strength (**A**) achieved better scores than a ten-fold difference (**B**). For some configurations, DA-only performs better than the competition between neuromodulators. However, the highest efficiencies in both tables correspond to a joint regulation. Changes in success rates between conditions were tested for statistical significance (two-sample Student's t-test with $p < 0.001$, ***)

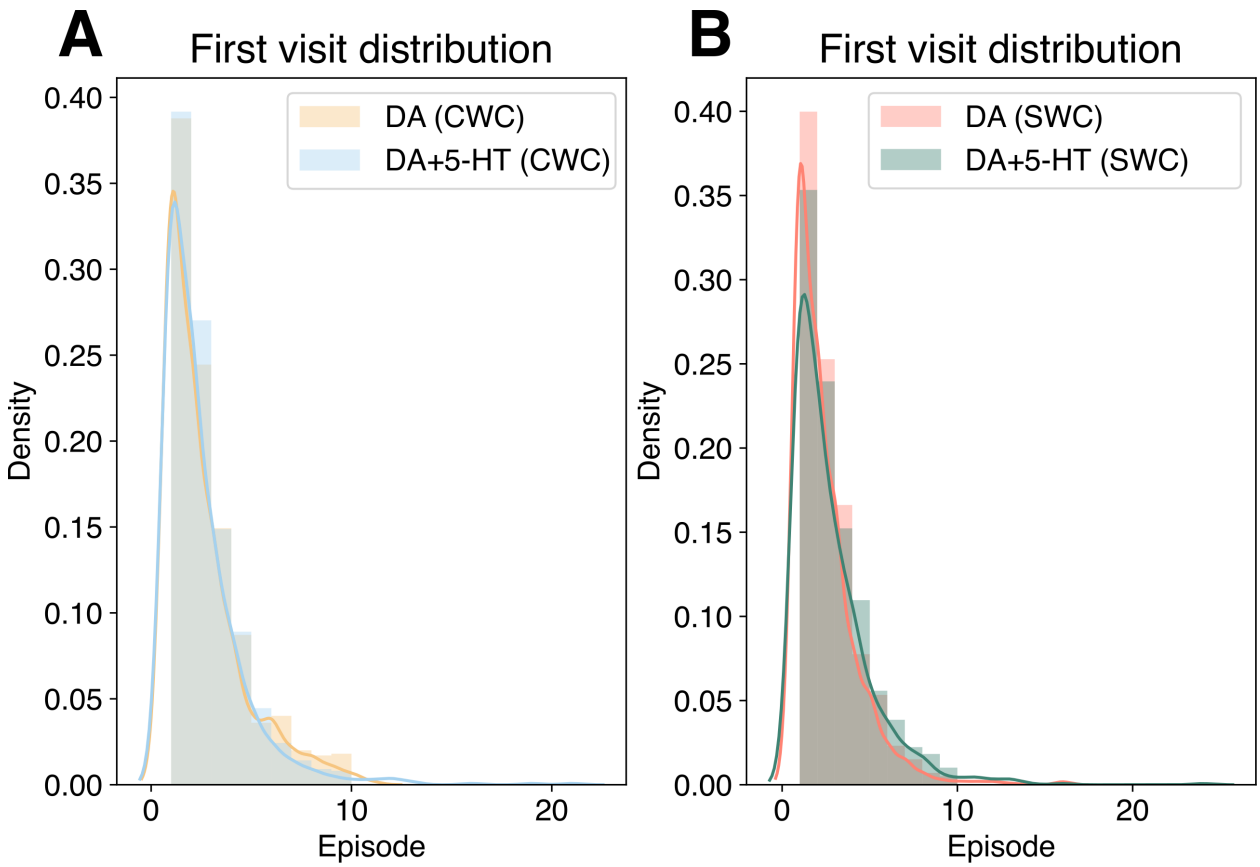

Figure S5: Addition of 5-HT does not alter explorative behavior, depicted as first visit distribution. **(A)** Comparison of the histograms of the first reward visit for 5-HT+DA and DA only in CWC (KL= 0.021). **(B)** Difference between the sample distribution of the first reward visit for 5-HT+DA and DA only in SWC (KL=0.015). Lines correspond to smoothing with kernel density estimation. Each conditioned was sampled for M=1000 simulations.

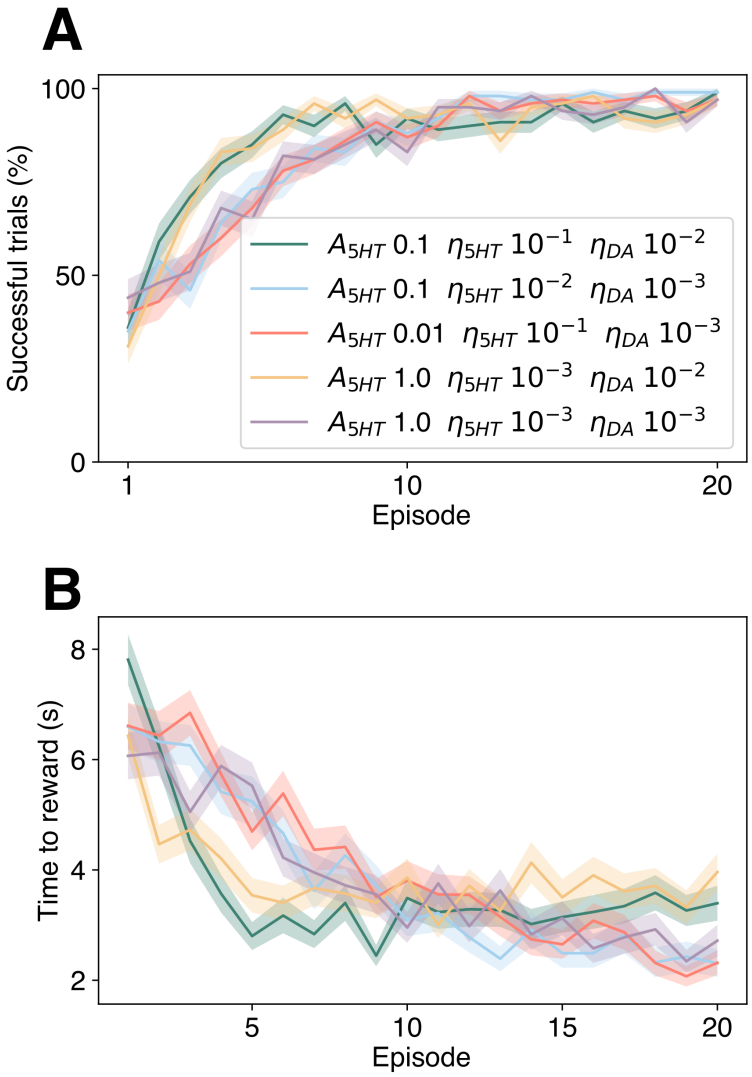

Figure S6: A step response of 5-HT in SWC does not impede learning in a MWM task. We tested a Heaviside function normalized to have the same maximum total response in a trial as the delta function formulation (i.e. $\int_0^{T_{max}} \kappa\Theta(t - T_{max})\,dt = \int_0^{T_{max}} \delta(t - T_{max})\,dt \implies \kappa = \frac{1}{T_{max}}$). **(A)** Learning curve of the best four STDP window combinations of nine tested. **(B)** Latency time for the best four Hebbian modulation formulations. In both cases, the average is shown with the filled area corresponding to the SEM (M=100 simulations). All other parameters were maintained constants.

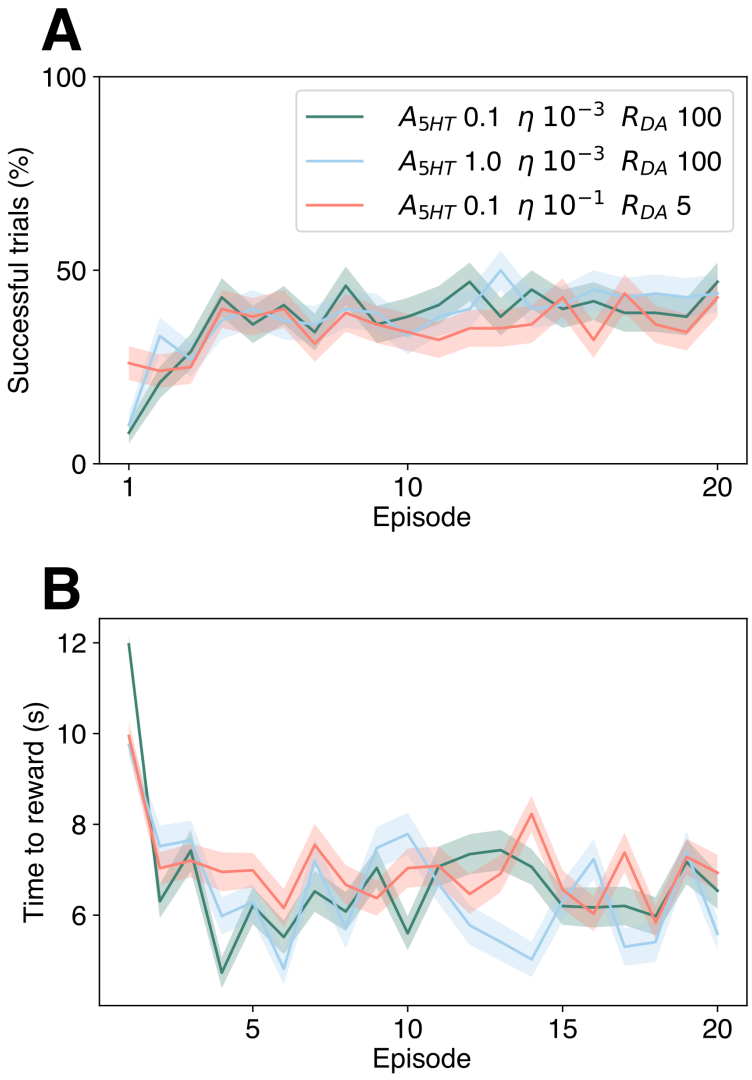

Figure S7: Dirac delta DA response hinders learning performance. We employed a Delta function formulation at the time of the reward with a posterior 300 ms continuous weight update. Thirty models were trained for different combinations of reward strength, STDP window of serotonin and learning rates. **(A)** Learning curve of the best three combinations in the percentage of successful trials. **(B)** Latency time for the best three configurations. In both cases, the line corresponds to average per test, and the shaded area corresponds to SEM (M=100 simulations).

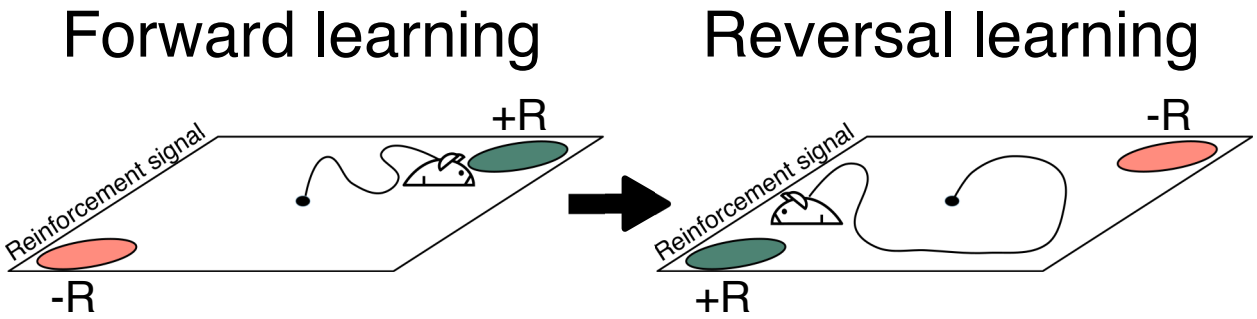

Figure S8: The task in reversal learning consists of an open-field (same characteristics as in Eq. 7; Materials and methods) with a reward (upper right) and a punishment (lower left). We assumed that the reward and the punishment are modeled with Dirac delta functions at the time the reinforcement signal is active T_{rew} (i.e. $R(t) = R\delta(t - T_{rew})$), with 300 ms of consummatory time for each neuromodulator (Materials and methods). If none of the items is reached in T_{max} the trial ends. In discrete form, the CWC rule transforms to

$$\Delta w_{ji}(t = T) = \eta_{DA} R(\sum_{t_i^f \in \mathcal{F}_i^{pc}} \sum_{t_j^f \in \mathcal{F}_j^a} W_{DA}(t_j^f - t_i^f)) \circ \epsilon_{DA}))(t = T),$$

which, with separated learning rates for each neurotransmitter, is identical to presenting different STDP window amplitudes. Thus, it is equivalent to SWC in this setting.