Parametrization of Neuromodulation in Reinforcement Learning

December 10, 2016

Ryan Golden, Marley Rossa, and Tunmise Olayinka

Abstract

Neuromodulation has been implicated in the induction and modulation of synaptic plasticity, and thus provides a potential vehicle for local reinforcement learning within a defined neural network. Specifically, the neuromodulator dopamine has been shown to encode reward-prediction error, the amount that an agent's received reward deviates from its expected reward. Furthermore, dopamine has been demonstrated to modulate synaptic plasticity based on novelty or uncertainty. Additional evidence exists supporting the parallel role of acetylcholine in the induction of plasticity via a gating mechanism, as well as the modulation of learning rates within discrete neural networks. We endeavor to understand the calculus underlying these interactions, and propose the possibility that neuromodulators act combinatorially to modulate synaptic plasticity. We evaluate the validity of our model with respect to parameters such as reward prediction error encoding, learning rate, and inducibility to spiking, conditioned on single layer neural networks with a bipartite graph connectivity scheme. Within this model we contrast the results of combinatorial modulation with that of classical reward-based training on canonical reinforcement learning tasks. We hope to extend our investigation by further quantifying the effect of potentially correlated activity amongst the stated parameters, particularly regarding task performance. Through iterative regression on the linearly separable as well as the nonlinear interactions of these neuromodulators, we ultimately hope to gain a more complete understanding of their collective dynamics and interactions as well as their evoked synaptic changes underlying reinforcement learning.

1 Introduction

The operant study of behaving animals under scheduled stimuli has long been a mainstay in the investigation of animal cognition. In these experiments, as an animal performs a desired action in response to a stimulus, it begins to learn and understand the relationship between the stimulus, its response, and

the reward. As it does so, it changes its behavior, re-orienting its attention in order to gain further reward [1]. This observation—that animals alter behavior to maximize their reward—has been fundamental to the study of animal cognition. The stimulus-response-reward dynamic can be recapitulated at the synaptic level, wherein repetitive stimulation of the postsynaptic neuron programmatically changes its activity, in turn causing the rewarded action—postsynaptic spiking—to become more or less likely [2]. Since this discovery, the algorithms with which an agent seeks to maximize reward, collectively called reinforcement learning (RL), have been iteratively examined at increasingly exacting levels of analysis, from the behavioral to the cellular.

Within a neuron, the RL is thought to be manifested via plasticity, i.e., the dynamic facilitation of synaptic activity by prior stimulation [2]. In this regime, activity within a synapse directly leads to the promotion of synaptic participators, the recruitment of which in turn leading to increased activity. As this learning paradigm occurs without regards to an overarching or known trajectory, it is an unsupervised learning rule, a learning rule that requires no known target, or error therefrom [3]. Classical experiments in synaptic long-term depression (LTD) and potentiation (LTP) and have typically characterized local synaptic potentiation as a weighted function of the presynaptic and the postsynaptic activity in the form of rate-dependent or spike-timing-dependent plasticity (STDP). Learning via STDP depends upon drawing inferences from the relative timing of the presynaptic and postsynaptic spikes [4, 5]. However, our increasing understanding of RL has brought with it the caveat that, in these models, full recapitulation of in vivo learning rules require more than the participation of local dynamics –the important modulatory role of neurotransmitters must be acknowledged.

The predominating neuromodulators in the central nervous system include dopamine (DA), acetylcholine (ACh), serotonin (5-HT), and norepinephrine (NE). As their name suggests, they modulate the synaptic activity of their target neurons, and are characteristically spatially and temporally distributed throughout the brain and its corresponding circuits [3]. The richness of understanding regarding their spatiotemporal localization and downstream behavioral effects has provided an exciting basis for the idea that, more than simply modulating arousal, these neuromodulators specifically and precisely regulate global dynamics via the implementation of identifiable distributed learning modules in the brain. Numerous studies have proposed that because these neurotransmitters have been shown to modulate STDP, they may thus serve as a basis for a computational theory underlying reinforcement learning and acquisition of goal-directed behaviors [2, 6]. As such, the present study contributes to the growing body of research that seeks to better characterize the complex role of neuromodulatory networks in generating and influencing learned reward-seeking and goal-directed behaviors.

A number of theoretical studies have investigated the hypothesis that reward-modulated STDP could be the neuronal basis for reward learning [6, 7, 8]. Within these models, DA represents the global learning signal for prediction of rewards, and ultimately, the reinforcement of behavior. Additionally,

ACh regulates the balance between memory storage and renewal, thus indirectly encoding the bases for saliency and novelty, and in ultimately informing our understanding of attention [9]. It is thought to regulate the degree with which we differentially weight our short-term and long-term predictions of reward, with long-term expectations of reward typically being reduced in a process known as temporal discounting.

In this study, we investigate possible consequences of linear and nonlinear interactions of the neuromodulators DA and ACh, which encode our meta-parameters of interest. Ultimately, we hope to gain a more complete understanding of whether and how neuromodulators may interact to more effectively govern the synaptic changes underlying reinforcement learning.

Dopamine has long been thought to encode a reward-prediction error which could be used to modulate synaptic plasticity based on novelty or surprise [10, 11]. Additionally, some evidence exists supporting the hypotheses that both DA and ACh play a role in inducing plasticity (via a gating mechanism), and modulating the learning rate of neural networks [8, 11]. Therefore, it is possible that the neuromodulators act in a combinatorial fashion to modulate synaptic plasticity.

By adding additional meta-parameters to encode the reward-prediction error (RPE), learning rate, and inducibility to a spiking neural network with random connectivity, we can compare how this combinatorial modulation compares to more classical reward-based networks on classical reinforcement learning tasks. We also hope to extend our investigation to include the effect of how correlated activity among meta-parameters affects task performance. This would be a method of studying possible consequences of linear and nonlinear interactions of the neuromodulators encoding these meta-parameters.

2 Methods

In the following, the network model, task, calculation of reward, and reward-modulation of STDP is based on the work of Fremaux et al. (2010).

2.1 Neuron Model & Network Architecture

Each network was modeled as a fully connected bipartite graph with 50 input nodes and 5 output nodes. The inputs were given by independent Poisson processes with a 10 Hz firing rate. The outputs were modeled as point neurons governed by a Spike-Response Model (SRM) (1-3), where $V_{threshold} = 16\ mV$, $\tau_m = 20\ ms$, $\tau_s = 5\ ms$, $\Delta V = 1\ mV$, and $V_{reset} = -5\ mV$. Additionally, for each output neuron the membrane potential was used to approximate the instantaneous firing rate of the SRM neuron according to (4), where $\rho_0 = 60\ Hz$. The instantaneous firing rate could then be used as the rate parameter for an inhomogeneous Poisson process that determined the spike-train of the output neuron. This last step was done to ensure an element of stochasticity was

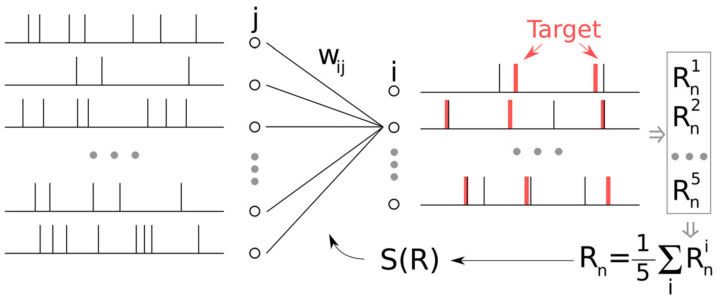


Figure 1: **The model network**. The (sum of) the corresponding post-synaptic output spike trains are compared with the (sum of) the 5 target trains, the difference of which yields a neuron specific estimate of the reward ${R^i}_n$. The summation of ${R^i}_n$ with respect to each neuron N_i yields the global reward for the network, per trial.

present, so that each trial could be considered independent.

$$\begin{aligned} &\tau_m \frac{dV}{dt} = K - V + g \\ &\textit{Upon spiking, } V, K \rightarrow V_{reset} \\ &\textit{Upon receiving a spike from neuron } i, \; g \rightarrow g + w_i. \end{aligned} \tag{1}$$

$$\tau_s \frac{dg}{dt} = -g \tag{2}$$

$$\tau_m \frac{dK}{dt} = -K \tag{3}$$

$$\rho = \rho_0 e^{\frac{V - V_{threshold}}{\Delta V}} \tag{4}$$

Plasticity Model

Plasticity in the model was governed according to the STDP rule (6) presented in Song & Abbott (2001), and used to update an eligibility trace according to (5), where $\tau_e = 500\ ms$, $A_+ = 0.188\ mV$, $A_- = -0.094\ mV$, $\tau_+ = 20\ ms$,

$\tau_- = 40\ ms$, and $\Delta_t = t_{post} - t_{pre}$ (the times of the postsynaptic and presynaptic spikes). Additionally, η is a parameter that modulates the learning rate, and was used to model the effect of ACh. At the beginning of training $\eta = 1$; it is decremented by 0.005 on every subsequent trial (during the simulations wherein only DA modulation was modeled, η was held constant at 1).

$$\tau_e \frac{de_{ij}}{dt} = -e_{ij} + \eta STDP \tag{5}$$

$$STDP = \begin{cases} A_+ e^{\frac{\Delta_t}{\tau_+}} & if \Delta_t \geq 0 \\ A_- e^{\frac{\Delta_t}{\tau_-}} & else \end{cases} \tag{6}$$

2.2 Spike-time learning task

Every repeated simulation consisted of a consecutive series of 250 trials of 1 second duration. For each trial, the input layer neurons were presnted with predetermined input spike-trains generated from inhomogenous Poisson processes. The reward contribution from each synapse was calculated as in (7), where N_i and N_i* are the current and target spike counts respectively. The local contributions were then averaged as in (8), where N is the number of output neurons, in order to attain a global reward, and ensure that the network is not implementing supervised learning. Additionally, a running average of this global reward was calculated as in (9) to use as a baseline value for DA signaling. The reward signal $S(R_n)$, was then computed as in (10), and could be thought of as representing phasic DA signaling. The reward signal was presented at the end of each trial, and caused a change in the synaptic weights according to (11), which were constrained to range between 0 and 1. Following each trial the eligibility traces were set to 0 to simulate long intervals between trials.

$$R_n^i = abs(N_i - N_i*)/max(N_i - N_i*) \tag{7}$$

$$R_n = (\frac{1}{N} \sum_i R_n^i) \tag{8}$$

$$\overline{R_{n+1}} = \overline{R_n} + \frac{R_n - \overline{R_n}}{N} \tag{9}$$

$$S(R_n) = R_n - \overline{R_n} \tag{10}$$

$$\Delta w_{ij} = S(R_n)e_{ij}(T) \tag{11}$$

As determined by the value of the delivered reward, the network updates its synaptic weights accordingly at the end of each trial where time = T, and i and j are specific to the input and output neurons, respectively.

To generate the input spike-trains, target spike counts, and target synaptic weights, the synaptic weights were randomly set and Poisson input was generated. These input spike-trains were saved to become the input stimulus for subsequent trials, the synaptic weights were saved to become the target weights, and the spike-trains produced by the output neurons was used to calculate the target spike counts. This was done to ensure that the target output would be learnable by the network.

3 Results

Grossly, our results validated our basic expectations: that is, a spike-time dependent learning network, under a reward- and/or novelty- modulated synaptic learning rule, can succeed in the semi-supervised learning of a given task; in this case a target spike train generated from feeding a fixed input to a reference network. In repeated simulations, the dynamics of reward-modulated (Dopamine) and reward-and-novelty (Dopamine & Acetylcholine) modulated learning showed a consistent statistically significant difference over the course of the learning trials($p = 0.0494$, repeated paired two-sample t test) (Fig. 2). This indicates that these networks employed distinct stratagems to solve the same problem. However, several results were surprising, and contravened our prior expectations.

Namely, we expected a gross decrease in the rate of convergence of the dual Dopamine & Acetylcholine network in comparison to the Dopamine-only network, coupled with an increased variance in the convergence of the Dopamine, reward-modulated network. These expectations were directly in line with our prior understanding of the cholinergic neuromodulation of novel stimuli. We predicted that cholinergic modulation—as modeled here by a decreased learning rate η, would in turn decrease the rate of convergence by directly decreasing the magnitudes of the trial-to-trial candidate weight changes e_{ij}. Additionally, we predicted that with more stable exploration of the solution space, the dual Dopamine & Acetylcholine network would converge, albeit slowly, to asymptotically higher rewards. This was not the case. Curiously, pure reward-based Dopamine modulation demonstrated consistently higher terminal rewards (Fig), and thus consistently closer targeting, than its dual-regime Dopamine & Acetylcholine counterpart. In the latter, simultaneous regulation of learning by both

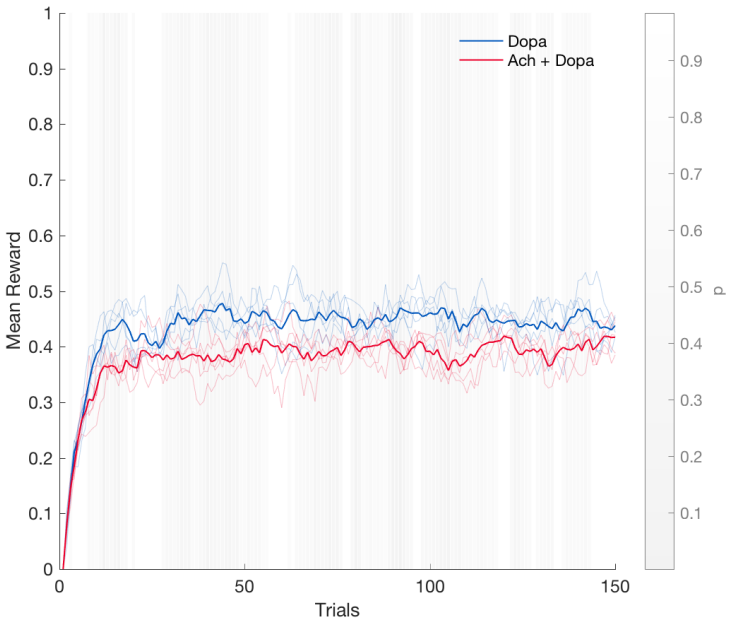


Figure 2: **Average reward over trials**. Learning curve of the mean evolution of the trial-average reward during repeated episodes of learning ($n = 5$), under either reward-based (Dopamine) or reward-and-novelty (Dopamine & Acetylcholine) based modulation. Strict reward-based modulation (Dopamine) results in consistently and significantly higher rewards ($p = 0.0494$, repeated paired two-sample t test).

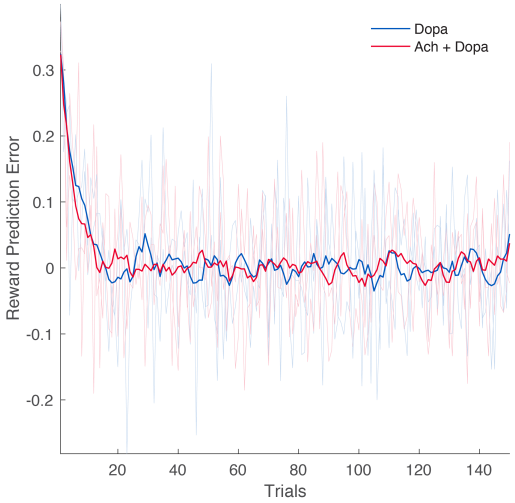


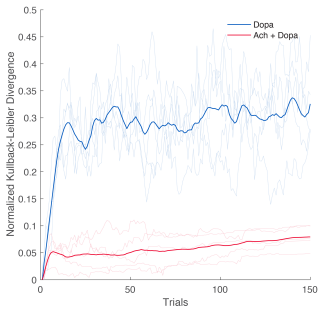
Figure 3: **Reward Prediction Error**. The difference between the actually delivered reward and the reward prediction for the stimulus. Corresponds to the convergence of an unsupervised network to a learned task.

reward and novelty lead to surprisingly poorer targeting, and lower reward. The reward prediction error (RPE) within both conditions nevertheless vanished to zero (Fig. 3), critically indicating that both regimes, the network successfully learned the task.

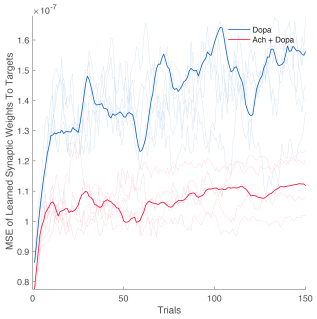
4 Discussion

The focus of our study was a biologically-grounded evaluation of a generic class of synaptic learning rules, under two sources of regulation: reward-modulated (R-STDP) and reward & novelty-modulated learning (RN-STDP). In either case, the fundamental paradigm was to elaborate upon an underlying unsupervised Hebbian learning rule, STDP, in the learning of a spike-rate based task-pattern. In the purely reward-modulated case, the time-dependent signal S(R) is a monotonic function of the understood reward R. In a learning regime modulated by both reward and novelty however; S is a function of both, with the latter implicitly encoded by changes in the learning rate. The learning rate, η controls the speed of learning, and in our simplified model, novelty-modulation was represented by the exponential attenuation of this rate over trials.

First we explain the discrepancies outlined in the results in several ways. We believe that for RN-STDP, the faster convergence to a lower-rewarding state is indicative of asymptotic residency in a local maximum. That is to say, it



(a) Kullback-Leibler Divergence



(b) Mean Squared Error from Target

Figure 4: **Kullback-Leibler Divergence and Mean Squared Error from target**. **A. KL Divergence** of the Dopamine & Dopamine + Acetylcholine learned network weights from target weights, during a learning session. The KL divergence is a constrained measure of the difference of the implicit probability distributions governing the profiles of synaptic weights. Strict reward-based modulation leads to significance divergence of the network from the target distribution of weights, along with trial-to-trial instability. In contrast, combined modulation shows both remarkably attenuated network divergence, along with a corresponding stability of the synaptic weights. **B. MSE** Like the K-L, the MSE is a quantatitive measure of the difference between two sets of synaptic weights. In that vein, reward-modulated Dopamine show consistent divergence and instability in network weights as compared to combined modulation.

appears that the RN-STDP network converges rapidly, but not optimally, to a locally maximal and stable value. In retrospect, this still coheres with our model of ACh as a modifier of the learning rate: for in our simulations, we presented identical stimuli to a learning network over the course of 250 trials. In this regime, successive re-presentation of the same stimulus is intuitively expected to transmit increasingly diminishing information to the network. In this context, one would expect a relative decrement in the ceiling of maximal possible reward, as the network receives diminished returns with each successive stimulus. In contrast, preservation of the learning dynamics in R-STDP leads to persistent and robust exploration of the solution space, as stimulus re-presentation without a decaying learning rate is sufficient to allow the network to dynamically escape local minima and explore alternative solutions. This is in turn reflected by decreased stability and increased ringing of the R-STDP network around the determined reward, especially in comparison to that of the RN-STDP network (Fig).

However, with R-STDP, consistent overshooting and undershooting of the predicted reward (due to a preserved η), allows for greater average reward over time for the Dopamine network relative to the Dopamine & Acetylcholine network. The explanation, faster convergence of acetylcholine-modulated networks to stable but lower maxima, is further illustrated by the quantification of the trial-to-trial differences of the synaptic weights of the two networks. Both the Kullback–Leibler (K-L) divergence, as well as computation of the relative mean squared error (MSE) were employed to measure this. The latter metric more precisely measures the distance of a single synapse from its target, as opposed to the K-L, which is an aggregate measure of the mutual information between the target weights and the learned distribution of weights, from trial to trial. Both approaches demonstrate that the final distribution of synaptic weights of the RN-STDP network more closely resemble that of the target network, in addition to further illustrating the notable trial-to-trial stability in its learned weights. In contrast, the R-STDP network demonstrates consistently higher divergence of its synaptic profile with that of the target weights, with visible, persistent volatility. Indeed, the striking divergence of the final distribution of weights of the Dopamine network with that of the target indicates that the Dopamine network found a solution quite distinct—in synaptic weight-space—from the target weights. Overall, we posit that it is precisely this decreased stochasticity in the RN-STDP network that prevents it from robustly exploring far-flung distributions.

Altogether, our studies highlighted several key results. Firstly, it illustrated that these highly expressive neural networks are indeed functionally redundant, with the dynamics of the R-STDP network indicating that a neural network is capable of producing congruent outputs from dissimilar synaptic weight distributions. Our work also revealed that novelty, as represented here, is a double-edged sword. That is to say, decreased novelty appears to prevent robust exploration but increase stability. A jocund anthropomorphic interpretation of our observations is that during a simulation, the R-STDP network 'fixates' on obtaining more reward, accepting any and all stimuli as salient,

while the RN-STDP network becomes 'bored', settling on a satisfactory network state, lacking the motivation to explore more 'creative' solutions. However, the biological and physiological processes underlying our observed results remain to be explored.

5 Future Aims

Our investigation has proffered preliminary promising results on the role of cholinergic learning-rate modulation in reward-modulated reinforcement learning. In particular, it revealed that the overt behavioral phenotypes that inform how we learn from novel stimuli may be reproducible at the cellular scale. Namely, acquisition of goal-directed behaviors through reward-based reinforcement is reflected in synaptic dynamics, and allows for mechanistic explanation of learnability of reward-modulated and novelty-modulated tasks. However, as our prior discussion highlighted, a more biophysically relevant model is necessary for the sophisticated elucidation of learning programs not reproducible by STDP or unsupervised RL alone, such as non-local dynamics believed to governed by neurotransmitter neuromodulation. For example, the crude approach with which we model novelty prevents the extrication of network cholinergic dynamics from that of the learning rate, thus precluding us from potentially understanding the nature of their underlying correlation. One possible workaround would be to model directly model increase in acetylcholine concentration, e.g. by an increase in K+ leak channel and AMPA conductance, as has been previously demonstrated experimentally. In addition, significant modification of the schedule of stimulus presentation is likely warranted, as our current pattern of input currently constrains to us the investigation of *non-novelty*, as that is the only reasonable interpretation one can make with our current design, founded on the continued re-presentation of identical input. Potential solutions include interleaving the relevant stimulus with random stimuli, as well as simultaneously training to multiple targets, which would help abrogate the linear dependence of novelty with the number of trials.

The last of our immediate aims is to develop more sophisticated reward criteria in our pooled estimation of the global reward. Currently, the model only considers the firing rate, averaged and evaluated over a single 1s trial. However, a criterion that better takes the effects of spike-timing on reward estimation is needed, as our method fails to register the relative pre- and post-synaptic spike-timing. Intra-trial correlation of spike-times allows for the more robust causal transduction of pre-synaptic firing patterns, and thus, overall network computation. In addition, the method with which dopaminergic neurons may actually calculate the stimulus-specific RPE (ssRPE)—and thus relevant candidates for the success signal—remains unknown. The lack of an endogenous mechanism for the ssRPE is a major obstacle to the *in vivo* validation of reward-modulated STDP. This internal reward predictor, termed a *critic*, is required for the solution of the aforementioned credit-assignment problem of local rewards, and in turn, the computation of a global reward. In this model, the critic learns the

state value function, and the actor—i.e. the network—uses this knowledge to learn how to respond, i.e., it learn the policy. However, the biological modes of computation with which this is endogenously achieved remain to be elucidated. An implicit corollary to this statement is that RN-STDP, as elaborated here, will never fully reproduce learning as seen in vivo, as the brain demonstrates a robust capacity for learning early in development, and in spite of an identifiable critic. Solutions to this problem include the use of an alternative rule to train the critic, prior to STDP. Other work has established the feasibility of separately training a critic to recognize and evaluate errors. Antecedent bootstrapping mechanisms that could implement such critic potentially include pure temporal difference learning, where adjacent states are assumed to be entirely consistent from trial-to-trial, and thus any deviation from this constraint allows for the computation of an error from that expectation. This error signal in turn allows the critic to learn the state value function for the network, and thus internally estimate anticipated reward.

Future models will invariably demand even more sophistication in order to more rigorously emulate STDP, in particular the multiplicative nonlinearities seen in studies investigating in vivo R-STDP facilitation. Namely, endogenous activation of D2 receptors, through independent activity, may differentially influence the expression and time course of spike-timing-dependent LTP and LTD. This disruption in the global registration of relative spike-times in turn prevents the canonical temporal decomposition of STDP as a successive sequence of multiplicative scaling events. Even more intriguingly, other studies have demonstrated that in some contexts, such as in the mammalian hippocampus, local increases in dopamine concentration may reverse the very paradigm of spike-time learning entirely, with the LTD component of STDP converted to LTP, while LTP, though remaining remaining facilitatory, becomes differentially timed with respect to its spike-time pre-post thresholds [5, 12]. Altogether, the robust elucidation of neuromodulator dynamics will likely demand models of iteratively increasing complexity, but in turn will allow us to better understand the extent to which the fundamental process that is learning is ultimately reducible.

References

[1] Fremaux N, Sprekeler H, Gerstner W (2010). Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity. Journal of Neuroscience, 30(40): 13326-13337.

[2] Kaelbling LP, Littman ML, Moore AW. (1996). Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, 4, 237-285.

[3] Doya K (2002). Metalearning and neuromodulation. Neural Netw. 15(4-6):495-506.

[4] Seol GH, Ziburkus J, Huang, SY, Song L, Kim IT, Takamiya K, Huganir RL, Lee H-K, Kirkwood A. (2007). Neuromodulators Control the Polarity of Spike-Timing-Dependent Synaptic Plasticity. Neuron, 55, 919–929. DOI 10.1016/j.neuron.2007.08.013.

[5] Song S, Abbott, LF. (2001). Cortical Remapping through Spike Timing-Dependent Plasticity. Neuron, 32, 1-20.

[6] Pawlak V, Wickens JR, Kirkwood A, Kerr JND. (2010). Timing is not everything: neuromodulation opens the STDP gate. Frontiers in Synaptic Neuroscience, 2(146), 1-14.

[7] Izhikevich EM (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex, 17(10), 2443-52.

[8] Fremaux N & Gerstner W (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. Front. Neural Circuits 9(85).

[9] Picciotto MR, Higley MJ, Mineur Y.S. (2012). Acetylcholine as a neuromodulator: cholinergic signaling shapes nervous system function and behavior. Neuron, 76(1): 116–129. doi:10.1016/j.neuron.2012.08.036.

[10] D'Ardenne K (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. PNAS 109(49):19900-19909.

[11] Gu Q (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. Neuroscience, 111(4), 815-35.

[12] Zhang J-C, Lau P-M, Bi G-Q. (2009). Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. PNAS, 106(31), 13028–13033. doi: 10.1073/pnas.0900546106.