Research on Bio-inspired Self-balancing Control Based on LIF Network

Zhixin Yan, Jin Li*, Junbang Jiang, Shanmengdai Luo, Lifang Huang

School of Science, Hubei University of Technology, Wuhan, Hubei, P.R. China

Received: January 18, 2026

Revised: January 19, 2026

Accepted: January 23, 2026

Published online: January 23, 2026

To appear in: *International Journal of Advanced AI Applications, Vol. 2, No. 2 (February 2026)*

* Corresponding Author: Jin Li (10279323@qq.com)

Abstract. Human balance is a skill gradually established through a sensory-action-feedback loop, relying on repetitive training, trial-and-error mechanisms, and the dynamic plasticity of synaptic connections. In this process, sensory signals are continuously transmitted to the central nervous system, where stable motor paths are formed through learning, enabling action reuse without complex calculations. Inspired by this mechanism, this paper proposes a balance learning method based on brain-like spiking neural networks and dopamine-modulated synaptic plasticity for self-learning control of the classic inverted pendulum system. The method connects the one-hot encoded sensory neuron group with motor neurons and utilizes a reward-driven synaptic weight update mechanism to gradually master the stable control of the inverted pendulum without the need for prior models or training data. Unlike traditional control algorithms such as PID or LQR, this approach features biological realism, strong adaptability, and self-organizing behavior, providing a new perspective on bio-inspired learning strategies for artificial intelligence in continuous control tasks.

Keywords: *Spiking Neural Network; Dopamine-modulated Synaptic Plasticity; Autonomous learning; Reward*

1. Introduction

In traditional control engineering, control loops typically consist of several key modules: the internal and external state perception modules, the control decision module, and the system dynamic model module [1, 3]. The working principle of a controller is to predict the future expected state based on the system's current state and the acquired environmental information, and then generate control actions accordingly, ultimately driving the system to achieve the desired behavior. In terms of control methods, model-based control relies on accurate modeling and simulation of the physical process to predict system behavior, while model-free control does not require an explicit dynamic model [4, 14], instead optimizing the control strategy through continuous interaction with the environment.

In recent years, machine learning techniques, particularly deep reinforcement learning (Deep Reinforcement Learning, DRL), have been widely applied in control tasks such as industrial process control, autonomous driving decision-making, and robotic operations, due to their powerful ability to learn complex strategies in high-dimensional state spaces [7, 12]. However, although traditional artificial neural networks (ANNs) mimic the connections of biological neurons in structure, their computational units are essentially continuous numerical mappings. This fundamentally differs from the time-dependent computational mechanisms that biological neural systems rely on, which depend on spike transmission [13, 18-23]. To bridge this gap, Spiking Neural Networks (SNNs), as the "third generation of neural networks," have been proposed. SNNs use spike trains in the time domain to transmit information, more accurately simulating the way signals are transmitted between biological neurons [9, 10]. The advantage of SNNs lies not only in their ability to encode information in time, but also in their event-driven sparse activation mechanism, which significantly improves energy efficiency, making them more suitable for embedded control scenarios with limited resources.

To enable SNNs to learn effective control strategies, researchers have developed various reward-modulated synaptic plasticity mechanisms. For example: R-STDP (Reward-modulated STDP): Combines the spike-timing differences (STDP) of pre- and post-synaptic spikes with external reward signals to achieve fine-tuning optimization of the strategy.RM-STDP: Builds upon R-STDP by introducing a weight-dependent multiplicative modulation factor to enhance the stability of the training process and the generalization ability of the strategy [9, 24-27].TD-STDP: Introduces the temporal difference error from reinforcement learning into the synaptic learning process and uses an eligibility trace mechanism to address the reward delay issue.

Although mechanisms such as R-STDP, DA-STDP, and TD-STDP have initially established a connection between synaptic plasticity and environmental rewards, they still have limitations in terms of biological realism, effective handling of delayed rewards, and adapting to dynamic task feedback. R-STDP mainly controls and amplifies the synaptic update based on instantaneous reward signals, making it difficult to effectively cope with situations where reward signals are significantly delayed [16, 17]. The DA-STDP model only establishes a weight update mechanism between pre- and post-synaptic spikes and fails to capture delayed rewards that appear several seconds after the behavior [28-32].

In contrast, DE-STDP (Dopamine-Eligibility STDP) shows greater potential in terms of biological plausibility and mechanism consistency [8, 33]. This mechanism uses dopamine (DA) concentration as a dynamic modulation factor and introduces the "eligibility trace" variable, coupling the local plasticity of STDP with the global reward signal reflected by dopamine concentration, giving synaptic weight changes "causal controllability" over time. This not only naturally simulates the core function of dopamine in reward-driven learning in biological neural systems, but also eliminates the need for external TD error calculation modules. The key feature of DE-STDP lies in its temporally separated weight update mechanism: STDP determines the possible direction of weight change based on spike timing differences (eligibility trace). The reward gating is then executed, with dopamine signals deciding whether these preset changes are actually implemented. This "trace-reward" pairing mechanism aligns with the time-scale differences between plasticity events and reward signals in biological systems [11, 15]. This two-stage regulation strategy makes DE-STDP advantageous in tasks involving sparse reinforcement signals, significant reward delays, or the need for local plasticity adjustments.

Unlike current mainstream control methods based on reinforcement learning or deep neural networks, this study emphasizes exploring the synaptic learning rules and biological information processing mechanisms achievable by the nervous system itself, and focuses on the possibility of efficient, unsupervised balance learning in low-dimensional state spaces. The research not only validates the practical feasibility of DE-STDP in dynamic control tasks but also provides theoretical foundations and potential technical pathways for promoting brain-like computational paradigms in practical control systems.

2. Methodology

2.1 Network Structure

To achieve reinforcement learning control for the inverted pendulum system, this study constructs a two-layer spiking neural network consisting of an input layer and an output layer. The network structure is simple, with clear connections, providing good biological interpretability and hardware deployment potential.

The input layer consists of 24 Leaky Integrate-and-Fire neurons, which receive discretized encoded information of the environment's state. Specifically, the system's four-dimensional state variables (cart position, cart velocity, pole angle, and angular velocity) are discretized into several intervals and mapped to the 24 neurons using one-hot encoding. This ensures the unambiguous transmission of state information and the capability for spike-based expression. The output layer contains 2 neurons, each representing one of the two discrete control actions (applying force to the left or applying force to the right). The network uses a fully connected structure, meaning each neuron in the input layer is synaptically connected to all neurons in the output layer.

To reduce computational complexity and enhance the biological plausibility of neuron behavior, this study adopts the classic Leaky Integrate-and-Fire model for neuron modeling [37-39]. In this model, each neuron contains only one state variable—its membrane potential $V(t)$, and its dynamic behavior follows the differential equation:

$$\frac{dV}{dt} = -\frac{V(t) - V\_rest}{\tau\_m} + \frac{I\_syn(t) + I\_ext(t)}{C\_m}$$

In this model, V_rest represents the resting potential, τ_m is the membrane time constant, and C_m is the membrane capacitance. $I\_ext(t)$ represents the externally injected current, primarily coming from the state perception input. $I\_syn(t)$ is the total synaptic current, triggered by synaptic inputs from within the network. When the membrane potential $V(t)$ exceeds the threshold voltage V_th, the neuron is considered to fire a spike and undergoes a potential reset followed by a refractory period [4].

This network architecture fully integrates the fundamental characteristics of biological neural systems, while maintaining high engineering feasibility, providing a solid foundation for subsequent control learning based on reward-modulated spiking plasticity rules.

2.2 State Discretization and One-Hot Encoding

The spikes generated by the input neurons are used to encode the observation states of the inverted pendulum system. Each observation variable of the system (including the cart position x、velocity v、pole angle θ and angular velocity ω) is mapped to an integer index according to the following rule[32]：

$$id\_obs=\begin{cases} 0, & obs \le obs_{min} \\ floor(\frac{x\text{-}x_{min}}{\Delta x}), & obs_{min} < obs < obs_{max} \\ N_{states,obs}\text{-}1, & obs \ge obs_{max} \end{cases}$$

In this context， Δx is the width of each interval, and obs_{min} and obs_{max} are the discretization limits for the variable. The total number of discrete states for each variable is given by：$N_{states,obs}=ceil(\frac{x\text{-}x_{min}}{\Delta x})$， The combination of the four observation variables forms a complete state $(id_x, id_v, id_\theta, id_\omega)$，The total number of states in the system is:

$$N_{states,total} = N_{states,x} * N_{states,v} * N_{states,\theta} * N_{states,\omega}$$

To achieve a unique representation for each state, each group of states is encoded by a set of n_{input} input neurons. Therefore, the total number of neurons in the input layer of the SNN is:

$$N_{input\ neurons} = N_{states,total} * n_{input}$$

When a specific state is input, only the n_{input} neurons corresponding to that state will spike, while all other neurons remain silent. This method is a classic example of one-hot encoding [30,34], which is commonly used in machine learning to represent categorical variables. For the discretization of the angle θ：

the central balanced region [-π/12, π/12]（equivalent to [-15°, 15°]）is divided into 10 subintervals；

The other unbalanced regions (such as [-π/2, -π/12] and [π/12, π/2]) are divided into coarser subintervals.

This type of "sparse-dense-sparse" partitioning helps to enhance the system's resolution in the critical balanced region, thereby improving control performance.

2.3 Reward Function Design

Intuitively, the reward function should reflect the core objective of the control task, which is to maintain the pole in the upright position. Since the control outcome depends on the action selected and executed in the current state of the system, when an action guides the system toward a direction more favorable for achieving this goal, it should be assigned a positive reward. To enhance the Spiking Neural Network (SNN) controller's responsiveness to system dynamics, various reward functions are designed based on the evolution of the state. As the reward function progresses from R_1 to R_2, the perceptual variables introduced become more complex, and the feedback mechanism transitions from a single physical quantity to a composite trend judgment. This allows the system to become more sensitive to "balance tendency" during the training process [35,40]. The second reward function R_1 is based on the trend of angular velocity changes between two time steps.

$$R_1(\omega_{old},\omega_{new}) = \begin{cases} 1, & \omega_{old} * \omega_{new} < 0 \\ 1, & |\omega_{new}| > |\omega_{old}| \\ -1, & otherwise \end{cases}$$

In this context, the first term checks whether the direction of the angular velocity has reversed, which indicates that the system is attempting to correct the existing rotational trend. The second term encourages a reduction in angular velocity, reflecting the control action's effect in suppressing the rotation amplitude. If neither of these conditions is met, the action is considered ineffective, and a punitive reward of -1 is applied to the system.

R_2 builds upon R_1 by further considering the trend in the direction of the angle to improve the system's overall ability to judge the return to equilibrium. It is defined as follows:

$$R_2(\omega_{old},\omega_{new},\theta_{old},\theta_{new})=\begin{cases} R_1(\omega_{old},\omega_{new}), & \theta_{new} * \omega_{old} > 0 \\ 1, & \theta_{new} * \omega_{old} \le 0 \ and \ \theta_{new} * \omega_{new} < 0 \\ -1, & otherwise \end{cases}$$

The logic of this function emphasizes that when both the angular velocity and the angle direction point toward the "return to vertical" trend, a positive reward should be given; otherwise, a penalty is applied. Particularly in some cases, if the angle θ_{old} and the angular velocity ω_{old} have opposite signs, it indicates that the current angular velocity is actually decreasing the tilt angle, meaning the action itself has a positive effect. In such a case, simply using the "direction reversal or deceleration" criterion in R_1 is insufficient to accurately evaluate the system's evolution. Therefore, R_2 further introduces a check on the sign combination of θ_{new} and ω_{new}:if the signs of θ_{new} and ω_{new} are opposite, it indicates that the new state is still maintaining the ideal trend of "angular velocity correcting the angle," and a positive reward is given; otherwise, the action is considered detrimental to system balance, and a punitive reward of -1 is applied. Compared to R_1, R_2 can more accurately recognize the actual contribution of the agent's action to the "system's return to balance" and provides more directional feedback signals during the SNN learning process.

2.4 DE-STDP

Since the dynamics of intracellular processes triggered by STDP and dopamine (DA) are not yet fully understood, this paper proposes a simplified phenomenological model to characterize the basic mechanism by which DA regulates STDP plasticity. Referring to the method by *i* et al. (2004) [46], the paper uses two phenomenological variables to describe the state of each synapse: the synaptic weight (s) and the enzyme activity variable (c) closely related to synaptic plasticity, such as the autophosphorylation of CaMK-II (Lisman, 1989), oxidation reactions of PKC or PKA, or other slower biochemical processes. These processes together form the so-called "synaptic tag" [38-41].

The basic dynamics of the model are described as follows:

$$\dot{c}=-\frac{c}{\tau_c}+STDP(\tau)\delta(t-t_{pre/post})$$

Here, *(\delta(t))* is the Dirac delta function, which is triggered when the pre- or post-neuron fires at the times *(t_{\text{pre}})* or *(t_{\text{post}})*, causing the variable (*c*) to be updated according to the STDP curve (Figure 1b). To clarify the mathematical nature of the STDP mechanism, the following model function is used to describe the synaptic timing-dependent plasticity changes [2, 47]:

$$W(\Delta t)\begin{cases} A^{+}e^{(-\frac{\Delta t}{\tau^{+}})}, & \textit{if } \Delta t >0 \\ -A^{-}e^{(\frac{\Delta t}{\tau^{-}})}, & \textit{if } \Delta t <0 \end{cases}$$

$\Delta t=t_i\text{-}t_j$ represents the time difference between the postsynaptic and presynaptic neuron spikes, with A^{+} and A^{-} representing the maximum adjustment amplitudes for long-term potentiation (LTP) and long-term depression (LTD), respectively, and τ^{+} 、 τ^{-} being the corresponding time window constants. This function characterizes the update magnitude of the synapse at different time differences, reflecting the fundamental principles of STDP.

The accumulated "plasticity potential" of the variable ccc only influences the synaptic weight sss when the DA concentration d > 0, enabling synaptic strengthening or weakening. Therefore, *c(t)* is considered as the "plasticity trace" or "eligibility trace" of the synapse, a concept introduced by Houk, Adams, and Barto (1995) [43-46]. Additionally, the dynamics of DA are described by the following equation:

$$\dot{d}=-\frac{d}{\tau_d}+DA(t)$$

Here, τ_d is the dopamine (DA) uptake time constant, and DA(t) represents the DA input generated by dopaminergic neuron firing in brain structures such as the ventral tegmental area (VTA) and the substantia nigra compacta. In this study, τ_d = 0.01 s， s is set to reflect the rapid clearance of DA in physiological processes. To better simulate the phasic and tonic patterns of DA, and in line with the dopamine encoding logic shown in Figure 1, when the system receives a reward (reward = 1), *DA(t)* is set to 0.05 μM, corresponding to the phasic activation triggered by reward in Figure 1(a) or the activation after conditioned stimulus predicts a reward in Figure 1(b). In the absence of a reward or with a negative reward (reward = -1), DA(t) is maintained at a baseline level of 0.001 μM, corresponding to tonic inhibition during the reward absence shown in Figure 1(c). At the same time, the background DA concentration is incorporated into the STDP weight update mechanism, represented by the following formula:：

$$\dot{s}=c(d\text{-}d\_baseline)$$

Here, *d_baseline* = 0.005 μM represents the background DA level of the system. This mechanism makes the synaptic potentiation process more sensitive to increases in DA concentration, while it becomes less likely to produce reinforcement effects when the DA level is below the baseline, helping to suppress the phenomenon of false reinforcement.

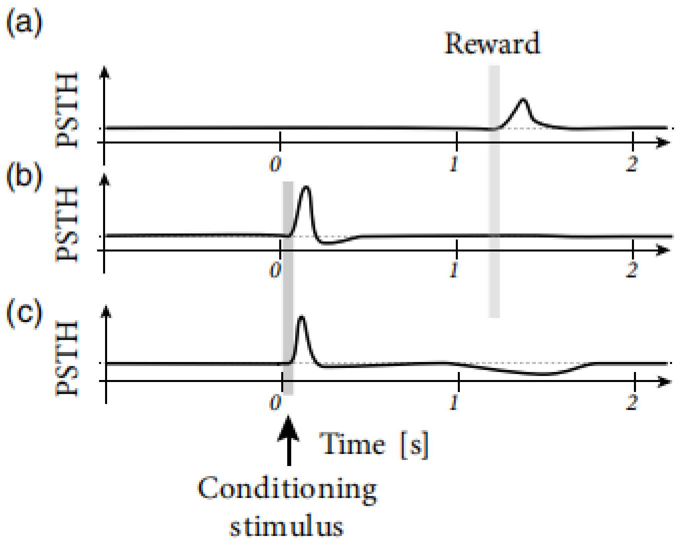


Figure 1. dopamine reward rule

In the inverted pendulum control system, the learning and reward mechanism is similar to the dopamine response logic shown in Figure 1. When the system successfully maintains balance, it corresponds to the reward activation in Figure 1(a), where dopamine activity in the neurons increases, reinforcing the successful balancing action. As the system learns, if the inverted pendulum has already learned the relationship between specific control signals and successful balance, these signals become conditioned stimuli, similar to the situation in Figure 1(b), where neurons respond to the conditioned stimulus in advance, without waiting for the reward to arrive. Eventually, when the system can predict the reward through the conditioned stimulus, the neuron's response becomes more stable, as shown in the trough in Figure 1(c), indicating that the system has learned how to efficiently and automatically maintain balance, without relying on every reward feedback. This learning process makes the inverted pendulum system more independent, enabling it to maintain balance more stably.

In summary, the model reasonably integrates the millisecond-scale synapse-specific STDP with the second-scale behavioral feedback in terms of timescale differences, as reflected in the dopamine encoding of reward timing in Figure 1. Although there is currently no direct experimental evidence to prove or disprove this model, it provides a clear, testable theoretical framework for exploring the regulatory mechanism of DA in STDP.

3. Results

3.1 Experimental Environment

The Cart-Pole system is one of the most classic control problems in reinforcement learning and is widely used to evaluate the performance of various control algorithms. In recent years, many studies based on Spiking Neural Networks (SNNs) have also used this system as a platform for algorithm testing [35,42]. This task can be described as follows: a cart and a rod connected by a hinge form the system, with the rod being able to rotate only in the plane perpendicular to the ground. The cart (Fig. 2) moves along a frictionless horizontal track, and the control agent must choose an action in each frame: apply a force to the left or to the right. The chosen action will affect the dynamics of the entire system, with the control objective being to keep the rod upright for as long as possible without becoming unstable.

In the MuJoCo simulation environment, decisions are made every 16 milliseconds. The observed system state includes:

The position of the cart: x, in meters;

The velocity of the cart: $v = \frac{dv}{dt}$, in meters per second;

The angle of the rod: θ, in radians (usually referenced to the vertical direction);

The angular velocity of the rod: $\omega = \frac{d\theta}{dt}$, in radians per second.

The simulation will terminate when any of the following conditions are triggered:

Rod tilt: The absolute value of the rod's angle exceeds 15°.

Cart out of bounds: The position of the cart exceeds the track boundaries of -2.0 meters to 2.0 meters.

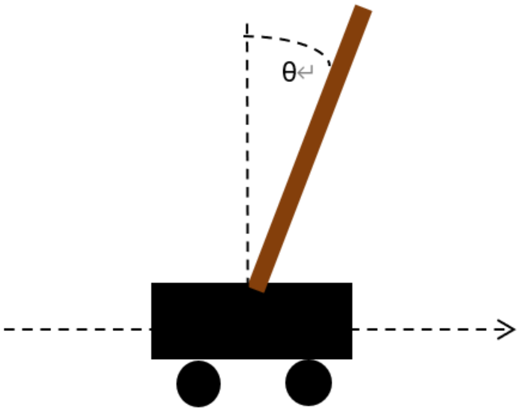


Figure 2. Cart and pole

3.2 Experimental Plan

In this experiment, the initial network weights are set to small random values, and the Spiking Neural Network (SNN) learns online through continuous interaction with the environment. The system's learning objective is to continuously keep the rod within a specified angle threshold range, i.e., in the "balanced state," for each episode until the cart exceeds the track boundary, which is considered a successful episode. The training process consists of 200 episodes. To evaluate the model's stability and generalization ability within a local time window, this paper introduces a sliding window success rate metric. Specifically, it is defined as the proportion of episodes within a sliding window of fixed length (20 episodes) where the number of balanced steps exceeds 7000 steps. This metric is considered the probability of "success" within the window. It dynamically reflects the phase effectiveness of the strategy and the stability improvement during the convergence process. To comprehensively evaluate the performance of different STDP mechanisms, all employing the reward function defined in R_2 ,the experiment compares the training performance of three plasticity rules: R-STDP (basic version), DA-STDP (with dopamine signal), and DE-STDP (with error and dopamine signal).

3.3 Experimental Results and Analysis

3.3.1 Evolution of Balance Steps During Training

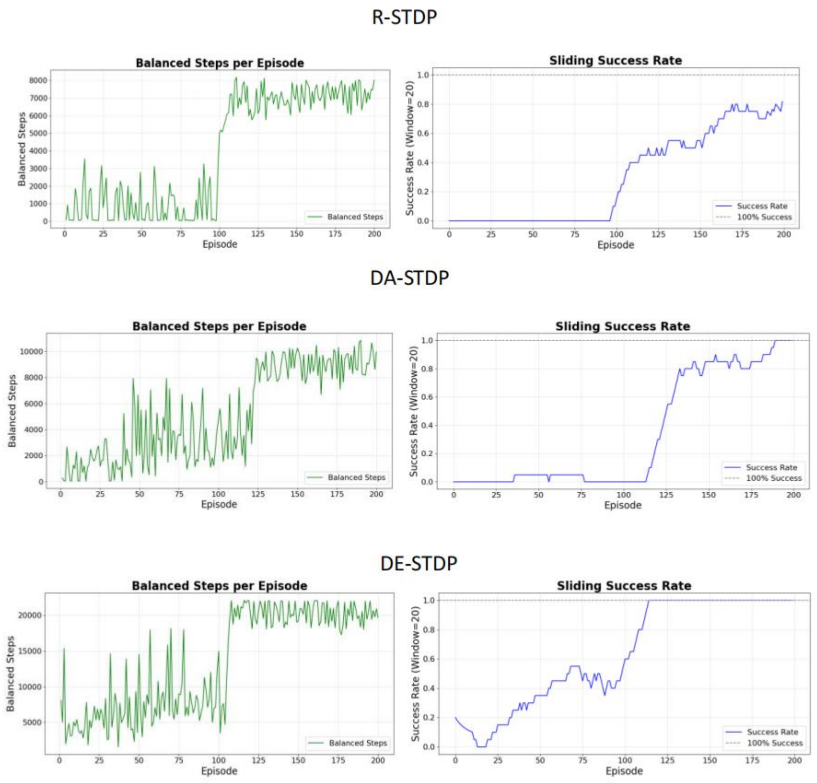


Figure 3. the comparison of performance for the three different STDP mechanisms

Fig. 3 shows the evolution of the number of balance steps per episode during the training process under three different STDP learning rules. R-STDP exhibits a significant training delay, with a notable improvement occurring only after around the 100th episode. In contrast, DA-STDP and DE-STDP quickly converge around the 110th episode, with DE-STDP demonstrating a strong learning capability in the early stages and maintaining the highest stability after convergence.

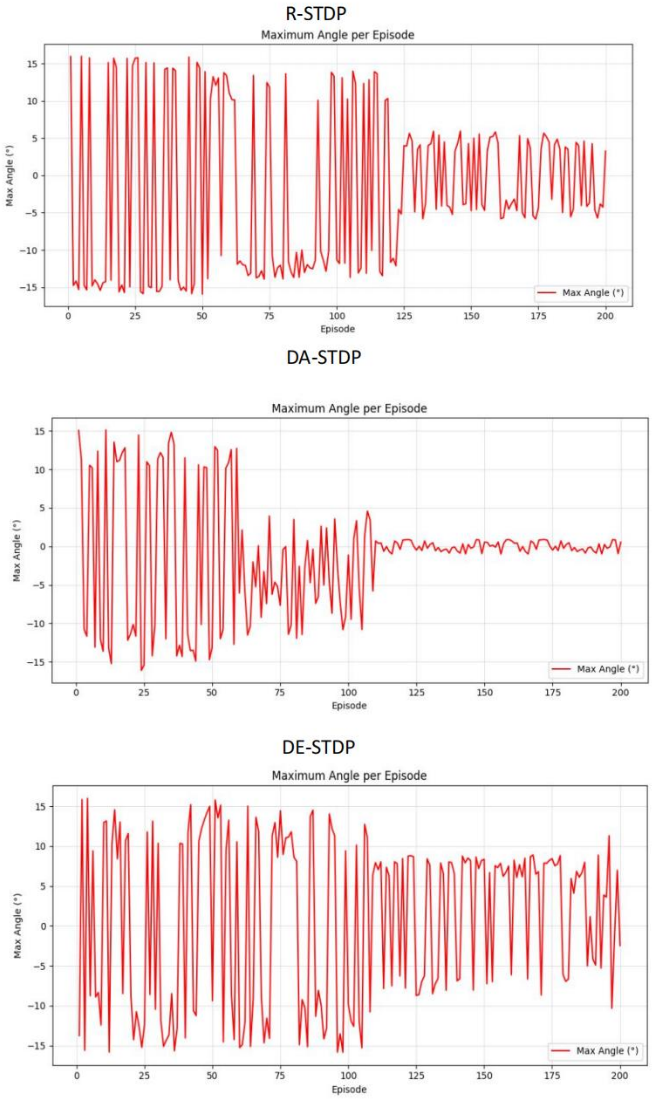
As shown in the figure, under DE-STDP modulation, the number of balance steps in the SNN during the CartPole task evolves over the course of training. In the initial phase (approximately the first 100 episodes), the SNN struggles to maintain the rod's stability, demonstrating a clear exploration phase. However, as training progresses, the synaptic connections are gradually optimized under DA modulation, and the system's balancing ability improves significantly. DE-STDP outperforms both R-STDP and DA-STDP in terms of convergence speed and stability, while DA-STDP shows a higher success rate and better sustained balance ability compared to R-STDP in the later stages.

3.3.2 Evolution of Maximum Angle During Training

This experiment uses the "maximum angle per episode" as a core observation metric to compare the training performance of R-STDP, DA-STDP, and DE-STDP in reinforcement learning tasks. By analyzing the fluctuations of the maximum angle over 200 episodes, the convergence and stability of different mechanisms are evaluated. From the experimental curves, the performance differences among the three STDP mechanisms are significant: R-STDP remains within a large oscillation range of -15° to 15° throughout the 200 episodes, with the system continuously cycling between "exploration and loss of control." This occurs because it relies solely on the temporal correlation between pre- and post-synaptic neurons, without considering "reward delay" or "error feedback," leading to an inability to establish a stable "action-reward" relationship. Its variance is 112.39, indicating large fluctuations.

DA-STDP, through dopamine encoding of the "reward prediction error," shows phase-wise convergence. The fluctuations in the first 50 episodes are similar to R-STDP, but after the 75th episode, the oscillation amplitude gradually decreases. After the 125th episode, it stabilizes between -5° and 10°. Although there is some convergence, due to the unresolved "temporal mismatch between actions and delayed rewards," there is still some fluctuation in the later stages. Its variance is 116.38, with reduced volatility compared to R-STDP.

DE-STDP performs the best. There is some fluctuation in the first 50 episodes, but after the 75th episode, the oscillation amplitude rapidly narrows. After the 125th episode, it stabilizes between -5° and 5°, and approaches 0°, achieving stable angle control. Its variance is 55.62, indicating a more stable learning process. Overall, R-STDP performs the worst due to the lack of adaptation to reward delay, DA-STDP shows improvement but with limited convergence, and DE-STDP excels in both convergence speed and stability, providing a more efficient STDP-based reinforcement learning framework.


R-STDP

DA-STDP

DE-STDP

Figure 4. shows the comparison of performance for the three different STDP mechanisms, illustrating the fluctuations of the maximum angle over 200 episodes.

3.4 Summary

This paper presents and implements a biologically-inspired phenomenological modeling approach focused on dopamine-modulated, time-dependent synaptic plasticity mechanisms, aiming to explain how delayed rewards at the behavioral level can lead to adjustments in synaptic strengths at the neural synapse level. The model draws from the ideas proposed by

Izhikevich et al., with the core concept being the introduction of two synaptic variables: synaptic weight (s) and the eligibility trace variable (c). The model is biologically grounded, combining the weight potential change (STDP rule) triggered by spikes with the delay mechanism of reward signals. This method is particularly suited to address a common issue in reinforcement learning — the delay of rewards relative to the timing of neural firing behaviors.

Additionally, the DA signal in the model is expressed in both baseline and phasic forms, with the sensitivity of weight adjustments under different DA concentrations enhancing the system's ability to differentiate environmental feedback and avoid erroneous reinforcement. This strategy effectively resolves the insensitivity to delayed rewards found in traditional STDP models, offering enhanced learning stability and biological plausibility. In conclusion, this approach provides a reasonable and experimentally testable modeling framework for synaptic learning mechanisms in neuromorphic reinforcement learning, especially suited for adaptive behavioral learning systems in delayed reinforcement scenarios.

Acknowledgements

Guiding project of Scientific Research Plan of Hubei Provincial Department of Education, Design of Application-Specific Integrated Circuit for Voice Command Recognition Based on Compute-in-Memory Architecture (No. B2020046); Cooperation Agreement on the Project of Design and Process Manufacturing Technology of 3D Stacked Code Flash Memory Chip (No. 2021802); PhD Research Foundation Project of Hubei University of Technology, Research on Sun Tracking Control and System State Monitoring of Butterfly Concentrating Photovoltaic Based on Deep Learning Image Analysis (No. 00185).

ORCID:

Jin Li - https://orcid.org/0000-0002-4615-2574

Junbang Jiang - https://orcid.org/0009-0008-8914-5658

References

[1] Baydin, A. G., Pearlmutter, B. A., Syme, D., Wood, F., & Torr, P. (2022). Gradients wit hout backpropagation. *arXiv preprint arXiv:2202.08587*.

[2] Burms, J., Caluwaerts, K., & Dambre, J. (2015). Reward-modulated Hebbian plasticity a s leverage for partially embodied control in compliant robotics. *Frontiers in neuroroboti cs*, *9*, 9.

[3] Barto, A. G. (2019). Reinforcement learning: Connections, surprises, and challenge. *AI Magazine*, *40*(1), 3-15.

[4] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zarem ba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540.

[5] Chen, W. (2022). Neural circuits provide insights into reward and aversion. Frontiers in Neural Circuits, 16, 1002485.

[6] Chevtchenko, S. F., Bethi, Y., Ludermir, T. B., & Afshar, S. (2024, June). A neuromorph ic architecture for reinforcement learning from real-valued observations. In 2024 Internat ional Joint Conference on Neural Networks (IJCNN) (pp. 1-10). IEEE.

[7] Barto, A. G., Sutton, R. S., & Anderson, C. W. (2012). Neuronlike adaptive elements tha t can solve difficult learning control problems. IEEE transactions on systems, man, and c ybernetics, (5), 834-846.

[8] Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. Current opinion in neurobiology, 10(6), 732-739.

[9] Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). Neuronal dynamics: From single neurons to networks and models of cognition. Cambridge University Press.

[10] Gewaltig, M. O., & Diesmann, M. (2007). Nest (neural simulation tool). Scholarpedia, 2 (4), 1430.

[11] Jitsev, J., Abraham, N., Morrison, A., & Tittgemeyer, M. (2012, September). Learning fr om delayed reward und punishment in a spiking neural network model of basal ganglia w ith opposing D1/D2 plasticity. In International Conference on Artificial Neural Network s (pp. 459-466). Berlin, Heidelberg: Springer Berlin Heidelberg.

[12] Mathews, M. A., Camp, A. J., & Murray, A. J. (2017). Reviewing the role of the efferent vestibular system in motor and vestibular circuits. Frontiers in Physiology, 8, 552.

[13] Markov, B., & Koprinkova-Hristova, P. (2024, September). Reinforcement Learning Co ntrol of Cart Pole System with Spike Timing Neural Network Actor-Critic Architecture. In International Conference on Artificial Intelligence: Methodology, Systems, and Appli cations (pp. 54-63). Cham: Springer Nature Switzerland.

[14] Shim, M. S., & Li, P. (2017, May). Biologically inspired reinforcement learning for mob ile robot collision avoidance. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 3098-3105). IEEE.

[15] Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP an d dopamine signaling. Cerebral cortex, 17(10), 2443-2452.

[16] Akl, M., Sandamirskaya, Y., Ergene, D., Walter, F., & Knoll, A. (2022, July). Fine-tunin g deep reinforcement learning policies with r-stdp for domain adaptation. In Proceedings of the International Conference on Neuromorphic Systems 2022 (pp. 1-8).

[17] Bing, Z., Meschede, C., Huang, K., Chen, G., Rohrbein, F., Akl, M., & Knoll, A. (2018, May). End to end learning of spiking neural network based on r-stdp for a lane keeping v ehicle. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 4 725-4732). IEEE.

[18] Han, Z., Chen, N., Xu, J., & Li, W. (2021). Research on intelligent control of inverted pe ndulum based on BP neural network. Experimental Technology and Management, 38(0 6), 101-106.

[19] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal poli cy optimization algorithms. arXiv preprint arXiv:1707.06347.

[20] Dev, A., Chowdhury, K. R., & Schoen, M. P. (2024, May). Q-learning based Control for Swing-up and Balancing of Inverted Pendulum. In 2024 Intermountain Engineering, Tec hnology and Computing (IETC) (pp. 209-214). IEEE.

[21] Shianifar, J., Schukat, M., & Mason, K. (2024, June). Optimizing Deep Reinforcement L earning for Adaptive Robotic Arm Control. In *International Conference on Practical Ap plications of Agents and Multi-Agent Systems* (pp. 293-304). Cham: Springer Nature Swi tzerland.

[22] Bhourji, R. S., Mozaffari, S., & Alirezaee, S. (2024). Reinforcement learning DDPG–PP O agent-based control system for rotary inverted pendulum. *Arabian Journal for Science*

and Engineering, *49*(2), 1683-1696.

[23] Dawane, M. K., & Malwatkar, G. M. (2025). Theoretical and experimental implementation of PID and sliding mode control on an inverted pendulum system. *Bulletin of Electrical Engineering and Informatics*, *14*(2), 920-930.

[24] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529-533.

[25] Wang Haiyi.(2021). 神经网络自适应控制在倒立摆系统中的应用研究［Research on the Application of Neural Network Adaptive Control in Inverted Pendulum System］. Hebei University of Science and Technology. https://doi.org/10.27107/d.cnki.ghbku.2021.000658.[in Chinese]

[26] Li Xinda.(2017). 倒立摆系统的神经网络控制研究［Research on Neural Network Control of Inverted Pendulum System］. Science & Technology Vision. (08), 163-164. https://doi.org/10.19694/j.cnki.issn2095-2457.2017.08.116. [in Chinese]

[27] Sboev, A., Vlasov, D., Rybka, R., Davydov, Y., Serenko, A., & Demin, V. (2021). Modeling the dynamics of spiking networks with memristor-based STDP to solve classification tasks. *Mathematics*, *9*(24), 3237.

[28] Hasegan, D., Deible, M., Earl, C., D’Onofrio, D., Hazan, H., Anwar, H., & Neymotin, S. A. (2021). Multi-timescale biological learning algorithms train spiking neuronal network motor control. *bioRxiv*, 2021-11.

[29] Fernández, J. G., Ahmad, N., & van Gerven, M. (2025). Noise-based reward-modulated learning. *arXiv preprint arXiv:2503.23972*.

[30] Vlasov, D. S., Rybka, R. B., Serenko, A. V., & Sboev, A. G. (2024). Spiking Neural Network Actor–Critic Reinforcement Learning with Temporal Coding and Reward-Modulated Plasticity. *Moscow University Physics Bulletin*, *79*(Suppl 2), S944-S952.

[31] Yang, Z., Guo, S., Fang, Y., & Liu, J. K. (2022). Biologically plausible variational policy gradient with spiking recurrent winner-take-all networks. *arXiv preprint arXiv:2210.13225*.

[32] Liu, Y., & Pan, W. (2023). Spiking neural-networks-based data-driven control. *Electronics*, *12*(2), 310.

[33] Vlasov, D., Rybka, R., Sboev, A., Serenko, A., Minnekhanov, A., & Demin, V. (2022, September). Reinforcement learning in a spiking neural network with memristive plasticity. In *2022 6th Scientific School Dynamics Of Complex Networks And Their Applications (DCNA)* (pp. 300-302). IEEE.

[34] Rodriguez-Garcia, A., Mei, J., & Ramaswamy, S. (2024). Enhancing learning in spiking neural networks through neuronal heterogeneity and neuromodulatory signaling. *arXiv preprint arXiv:2407.04525*.

[35] Feng, H., & Zeng, Y. (2022). A brain-inspired robot pain model based on a spiking neural network. *Frontiers in Neurorobotics*, *16*, 1025338.

[36] Mozafari, M., Kheradpisheh, S. R., Masquelier, T., Nowzari-Dalini, A., & Ganjtabesh, M. (2018). First-spike-based visual categorization using reward-modulated STDP. *IEEE transactions on neural networks and learning systems*, *29*(12), 6178-6190.

[37] Fife, T. D. (2010). Overview of anatomy and physiology of the vestibular system. *Handbook of Clinical Neurophysiology*, *9*, 5-17.

[38] Goldberg, J. M. (2012). *The vestibular system: a sixth sense*. Oxford University Press, USA.

[39] Oteiza, P., & Baldwin, M. W. (2021). Evolution of sensory systems. *Current Opinion in Neurobiology*, *71*, 52-59.

[40] Latash, M. L., Levin, M. F., Scholz, J. P., & Schöner, G. (2010). Motor control theories and their applications. *Medicina (Kaunas, Lithuania)*, *46*(6), 382.

[41] Imai, T., Moore, S. T., Raphan, T., & Cohen, B. (2001). Interaction of the body, head, an d eyes during walking and turning. *Experimental brain research*, *136*(1), 1-18.

[42] Zeff, S., Weir, G., Hamill, J., & van Emmerik, R. (2022). Head control and head-trunk c oordination as a function of anticipation in sidestepping. *Journal of Sports Sciences*, *40* (8), 853-862.

[43] Thao, N. G. M., Nghia, D. H., & Phuc, N. H. (2010, October). A PID backstepping contr oller for two-wheeled self-balancing robot. In *International Forum on Strategic Technol ogy 2010* (pp. 76-81). IEEE.

[44] Houk, J. C., & Adams, J. L. (1995). 13 a model of how the basal ganglia generate and us e neural signals that. *Models of information processing in the basal ganglia*, *249*.

[45] Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP an d dopamine signaling. *Cerebral cortex*, *17*(10), 2443-2452.

[46] Izhikevich, E. M., Gally, J. A., & Edelman, G. M. (2004). Spike-timing dynamics of neu ronal groups. *Cerebral cortex*, *14*(8), 933-944.

[47] Akl, M., Sandamirskaya, Y., Ergene, D., Walter, F., & Knoll, A. (2022, July). Fine-tunin g deep reinforcement learning policies with r-stdp for domain adaptation. In *Proceedings of the International Conference on Neuromorphic Systems 2022* (pp. 1-8).