Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 4296356, 15 pages
http://dx.doi.org/10.1155/2016/4296356

Research Article

A Cognitive Model Based on Neuromodulated Plasticity

Jing Huang,[1,2] Xiaogang Ruan,[1] Naigong Yu,[1] Qingwu Fan,[2] Jiaming Li,[2] and Jianxian Cai[1]

[1]*Institute of Artificial Intelligence and Robotics, Beijing University of Technology, Beijing 100124, China*
[2]*Pilot College, Beijing University of Technology, Beijing 101101, China*

Correspondence should be addressed to Jing Huang; aiandrobot@163.com

Received 6 March 2016; Revised 17 July 2016; Accepted 22 September 2016

Academic Editor: Leonardo Franco

Associative learning, including classical conditioning and operant conditioning, is regarded as the most fundamental type of learning for animals and human beings. Many models have been proposed surrounding classical conditioning or operant conditioning. However, a unified and integrated model to explain the two types of conditioning is much less studied. Here, a model based on neuromodulated synaptic plasticity is presented. The model is bioinspired including multistored memory module and simulated VTA dopaminergic neurons to produce reward signal. The synaptic weights are modified according to the reward signal, which simulates the change of associative strengths in associative learning. The experiment results in real robots prove the suitability and validity of the proposed model.

1. Introduction

Associative learning can be divided into two types: classical conditioning and operant conditioning [1]. As a basic type of learning, associative learning has been studied a lot. Many computational models have been presented. Some are about classical conditioning [2–5]. The stimulus in these models is assumed to have a weight to measure how strongly it predicts the reward. The bigger the weight is, the closer the stimulus is to the reward. Others are about operant conditioning and reinforcement learning which originates from the former [6–10]. A universal frame for both kinds of conditioning is much less studied. In the present study, we try to set up a model to regard the two aspects of associative learning as a whole and explain them in a common way.

Although the two categories are distinguished in some aspects (e.g., the reward does not depend on the actions chosen by the animal in classical conditioning while it does in operant conditioning), they still have many common features [11]. Both of them are concerned with how animals find the causal relationship between reward and the corresponding signs, for example, some stimulus or their actions. Meanwhile, both of them describe how stimulus is associated with response. Given a stimulus S, the animal tries a response *R*. In classical conditioning, if S tends to predict the appearance of reward (e.g., food), the connection is strengthened [12]. While in operant conditioning, if the result is positive, the connection between S and *R* is strengthened, otherwise it is weakened [13].

In essence, associative learning is not a prerogative of human being. Many researches have suggested that even organisms with rather simple neural systems can have such abilities and establish the association between stimulus and response in classical [14] or operant conditioning [15] way. These findings indicate that relatively simple neural network can have the function of associative learning.

At macroscopic level, associative learning is a process during which human beings and animals discover relationships between stimuli, actions, and outcomes. However, at neural level, associative learning is related to synapses' ability to change their strength in signal transmission, which is called synaptic plasticity.

Synaptic plasticity is considered as a prime mechanism for learning and memory. Such idea is firstly studied by Hebb [16] and then gathers a broad consensus among researchers [17–19]. These studies revealed that there is an important link between local plasticity and macrolevel behavioral learning [20]. The synaptic changes of particular pathways in sensorimotor system could lead to the behavioral changes. Meanwhile, multiple researches suggest that synaptic plasticity is

often affected by neuromodulators like dopamine [21–25]. Neuromodulation may involve associative learning and work as a type of synaptic gating mechanism. Therefore, synaptic plasticity modulated by neuromodulators is considered to play an important role in conditioning behavior learning [20]. Driven by these findings, we try to construct a model based on synaptic plasticity with neural modulation and apply it to explain associative learning. Here, the synaptic plasticity is artificial and represented by changing the network's connective weights according to learning mechanism.

Another problem is how to represent the weights. According to Yang et al.'s research [26, 27], neurons in the lateral intraparietal area (LIP) may involve simple decision-making, which is similar to action selection in associative learning. Such decision-making may be done in the form of a log likelihood ratio (log LR) in the neural system. Pfeiffer et al. adopted the conclusion and presented a brief frame for neural modulated plasticity [28, 29].

Inspired by the researches above, we present a cognitive model based on modulated synaptic plasticity. The focus of this study is to apply the model to explain the two kinds of associative learning in a unified way. Moreover, as memory plays a fundamental and important role in learning and cognition [30], we add memory module in our model. To find out how memory works in high-level cognitive activities, a number of computational models for memory have been proposed. Many of them focus on two challenging problems: defining the nature of working memory storage and the relationship between working memory and long-term memory. The representative work includes levels of processing [31], parallel-distributed processing [32], models involving hippocampal area of human brain [33], and information processing [34]. Considering the universality, popularity, and influence in history, we adopt information processing model in our work.

This paper is organized as follows. In Section 2, we explain the architecture of the model. In Section 3, we present the working algorithm for the model. In Section 4, we analyze the convergence of the learning mechanism. In Section 5, we reproduce both classical conditioning experiment and operant conditioning experiment in real robots. We also introduce the details about the experiment settings and the structure of the networks and analyze the results of the experiments in this section. The paper ends with concluding remarks in Section 6.

2. The Architecture of the Model

The architecture of our model is shown in Figure 1. The relationship between stimulus and response is modeled as the mapping from perception to motor, represented by the information stream from sensory module to action module. Meanwhile, as mentioned above, memory plays an important role in cognition. Therefore, 3-layer memory module is added in the model. Learning mechanism here refers to the rule of changing the synaptic weights between working memory and action module, which makes the model self-learning and self-organized. VTA dopaminergic neurons are also simulated here to represent the neuromodulation and to produce the reward signal for the learning mechanism.

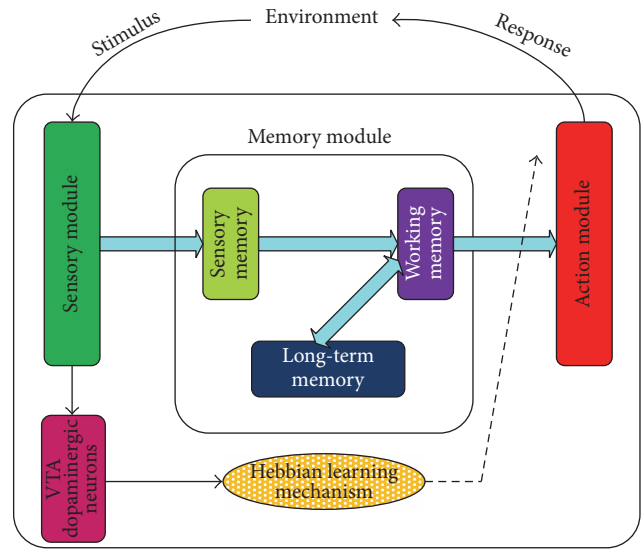


Figure 1: The architecture of the cognitive model. The model includes 3 main modules: sensory module, action module, and memory module. Thus, the interactions between agents and environment, that is, stimulus and response, are transferred as the information flows from sensory module to action module, signifying the sensorimotor system. Besides, VTA dopaminergic neurons and learning mechanism work together as the learning system to modify the synaptic weights between working memory and action module. All the above, the agent and the environment, compose a close-loop system.

2.1. Sensory Module. The sensory module represents sensors in animals or robots. It collects and receives stimulus from environment, which will soon be transmitted to sensory memory. Its output also provides the unit of VTA dopaminergic neurons, helping judge whether there is a reward or not.

2.2. Memory Module. Memory is important in cognition. Here, we adopt the three-layer architecture to describe the memory module. They are sensory memory, working memory, and long-term memory.

Sensory Memory. Sensory memory stores sensory information from sensory module just long enough to transfer it to next memory unit: working memory. Its function is to provide a snapshot of agents' overall sensory experience and retain the impressions after the original stimulus has stopped.

Working Memory. Working memory, the second layer of the multistore memory model, receives the output from sensory memory. It plays the role not only of a bridge between sensory memory and long-term memory, but also of the key for learning and memory. Working memory processes the information from sensory memory to make it easy to handle, that is, memory coding, and delivers it to action module.

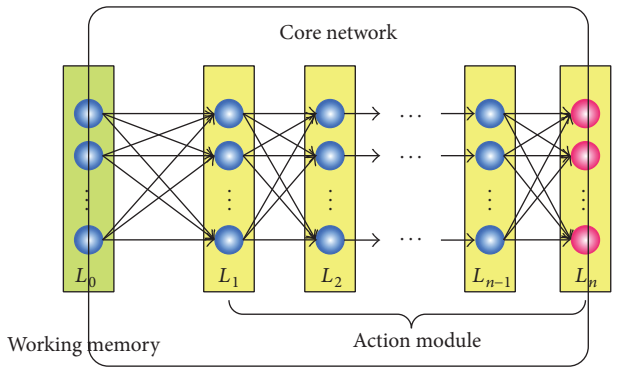


Figure 2: The structure of the core network. Working memory and action module are connected in a full connective way; that is, every neuron in the former layer is connected to all neurons in the next layer and so on. Neurons in working memory are ADALINE neuron [35] whose threshold value is 0, shown as blue ones in the figure. All the neurons of the action module except the ones in the last layer are the same. The neurons in last layer are perceptron neurons [36] whose output is discrete and easy to use, shown as red ones in figure.

Meanwhile, working memory records the statistics of each action with reward or without reward, which offers data for learning mechanism. All results for each action selection, that is, the numbers of times of reward or no reward for each serial action, will be saved in working memory. For instance, suppose there is an action chain: $a_1 a_2 \cdots a_n$ $(n \geq 1)$, in which a_1 is the first action while a_n is the last one. $R_{1,2,\ldots,n}$ records the number of times of reward after the action chain is selected, while $\overline{R}_{1,2,\ldots,n}$ represents the number of times of no reward. Both $R_{1,2,\ldots,n}$ and $\overline{R}_{1,2,\ldots,n}$ will be saved in working memory.

Finally, working memory communicates with long-term memory for accumulating and taking advantage of learning experience. Every time when learning starts, it loads the last time learning result from long-term memory and stores new learning result at the end of learning.

Long-Term Memory. Long-term memory along with working memory and sensory memory constitutes the complete memory mechanism. The main function of long-term memory is to save the learning result, which represents the experience accumulated through the interaction of agents with the environment. Every time when learning starts, the long-term memory is retrieved and loaded to working memory for new learning. When learning ends, the result is saved in long-term memory.

2.3. Action Module. Action module represents effector or neurons related to actions. Its input comes from working memory, while its output represents the expression of actions. The action module along with working memory, especially the connections between them, is the core of the whole model. The structure of the core network is shown in Figure 2.

As illustrated in Figure 2, action module consists of multiple layers of neurons. Each layer represents one time of action selection while each neuron represents one action. In fact, the network in action module indicates the action chains learned.

The actions will be chosen in winner-take-all way. In other words, the action with the biggest connective weight will be chosen.

Suppose at time t the input vector for a neuron in action module is $\mathbf{x}(\mathbf{t}) = (x_1(t), x_2(t), \ldots, x_n(t))^{\mathrm{T}}$, and the corresponding weight vector is $\mathbf{w} = (w_1, w_2, \ldots, w_n)^{\mathrm{T}}$. If the neuron is not in the last layer, its output is calculated as follows:

$$o(t) = \sum_{i=1}^{n} w_i x_i(t) = \mathbf{w}^{\mathrm{T}} \mathbf{x}(\mathbf{t}). \tag{1}$$

Otherwise, its output is calculated as follows:

$$o(t) = \begin{cases} 0 & u(t) < b \\ 1 & u(t) \geq b. \end{cases} \tag{2}$$

The symbol b in formula (2) signifies the threshold value, or bias, of the neuron. And $u(t) = \sum_{i=1}^{n} w_i x_i(t) = \mathbf{w}^{\mathrm{T}} \mathbf{x}(\mathbf{t})$.

2.4. VTA Dopaminergic Neurons. Reward signal is a crucial factor in associative learning. As mentioned above, many evidences show that neuromodulation plays an important role in mediating the reward signal. The ventral tegmental area (VTA) in midbrain area is believed to be the neural substrate of such modulation [37].

VTA is one of the most important dopaminergic areas. The best-developed current theory of dopaminergic function is the "reward prediction error" hypothesis that dopamine encodes the difference between actual and predicted rewards [38, 39]. The magnitude of phasic dopamine-neuron bursts quantitatively represents positive prediction errors [40].

The idea can be expressed in the following formula where δ represents the dopamine signal, R represents the actual reward, R' represents the predicted reward, and k is the positive coefficient:

$$\delta = k \times \left(R - R'\right). \tag{3}$$

In this work, agents are supposed to have no expectations about reward, that is, $R' = 0$. Thus, the dopamine signal can be regarded as proportion to actual reward; that is,

$$\delta = k \times R. \tag{4}$$

To calculate the dopamine signal, we introduce the concept *ideal degree* (ID) in this model. It is a numeric value decided by specific applications and describes how ideal the status agents perceive is. The bigger the ideal degree is, the better the corresponding status is. We assume that the ideal degree will increase if agents get reward and it will decrease if not. Based on such an assumption, a reward is regarded as the function of ideal degree. Suppose at present time t that the status perceived is s, and its ideal degree is ID(s). After the serial action a is executed, the status transfers to s', whose

ideal degree is $\mathrm{ID}(s')$. Then, the reward function R can be defined as follows:

$$R = \frac{2}{1 + e^{-\Delta \mathrm{ID}}} - 1, \tag{5}$$

where $\Delta\mathrm{ID} = \mathrm{ID}(s') - \mathrm{ID}(s)$. Then, we can calculate the dopamine signal according to formula (4) and (5).

As formula (3) shows, R is a sigmoid function with the value domain $(-1, 1)$. The function is on the symmetry of origin and monotonically increasing. When $\Delta\mathrm{ID} > 0$, that is, $\mathrm{ID}(s') > \mathrm{ID}(s)$, then $R > 0$, indicating that the ideal degree increases and agents get reward by selecting the action chain. Moreover, since it is monotonically increasing, the more ideal degree increases, the bigger R becomes. The extreme case is that R will approach 1 in case $\Delta\mathrm{ID}$ approaches infinite. On the contrary, if $\Delta\mathrm{ID} < 0$, then $R < 0$, indicating that agents have not been rewarded and the status is worsening. The more the ideal degree decreases, the less R becomes. When $\Delta\mathrm{ID} \to -\infty$, $R \to -1$. A special case is $\Delta\mathrm{ID} = 0$, that is, $\mathrm{ID}(s') = \mathrm{ID}(s)$; then $R = 0$. The case is regarded as unrewarded.

3. Working Algorithm of the Model

Working algorithm describes how the model works and the input is transformed to the output. We introduce a new concept system entropy as a measure of convergence in the working algorithm. *System entropy* (denoted as SE), like entropy in information theory, is calculated as follows:

$$\mathrm{SE} = -\sum_{i=1}^{m_1} p_i \log p_i, \tag{6}$$

where p_i represents the probability of the selected action a_i.

Obviously, system entropy signifies the degree of self-organization. The less it is, the higher the degree of self-organization is. When system entropy approaches its minimum, the model or the working algorithm has converged. We use SE or the learning times as the ending condition for the system.

The whole algorithm is as shown below.

Step 1 (initialization). Retrieve long-term memory and load its content to working memory.

Set $nr_i = 0$ and $\overline{nr}_i = 0$ $(i = 1, 2, \ldots, m_1)$, where nr_i represents the number of times of being rewarded for action a_i and $\overline{nr}_i$ represents the number of times of not being rewarded.

Set the connective weights between neurons in working memory and action module $w_{ji} = 0$ $(j = 1, 2, \ldots, m_0, i = 1, 2, \ldots, m_1$, m_0 and m_1 are, resp., the number of neurons in working memory and action module).

Calculate the *system entropy* according to formula (6) where $p_i = 1/m_1$; that is, agents select action randomly at the beginning.

Step 2 (select action in WTA (winner-take-all) way). Choose the action a_i with the maximum corresponding weight.

Update the number of being selected for the action a_i: $N_i = N_i + 1$.

Update the probability of each selected action as follows:

$$p_i = \frac{N_i}{\sum_{j=1}^{m_1} N_j}. \tag{7}$$

Update the system entropy SE.

Step 3 (observe the response from the environment, judge whether the action is rewarded, and then get the output of VTA dopaminegic neurons). Get the new perceived information through sensory module.

Update the representation of the sensory information in sensory memory and working memory.

Calculate the dopamine signal according to formula (4) and (5).

For each action a_i of the action chain being learned, update nr_i and $\overline{nr}_i$ as follows.

If action a_i results in reward,

$$nr_i = nr_i + 1. \tag{8}$$

Otherwise,

$$\overline{nr}_i = \overline{nr}_i + 1. \tag{9}$$

Step 4 (adjust the weights related). For each weight w_{ji} related to the action sequence being learned, the following happens.

If corresponding action a_i results in reward,

$$\begin{aligned} w_{ji} &= \ln \frac{nr_i + 1}{\overline{nr}_i} = \ln \frac{nr_i}{\overline{nr}_i} \left(1 + \frac{1}{nr_i}\right) \\ &= w_{ji} + \ln \left(1 + \frac{1 + e^{-w_{ji}}}{nr_i + \overline{nr}_i}\right). \end{aligned} \tag{10}$$

Otherwise,

$$\begin{aligned} w_{ji} &= \ln \frac{nr_i}{\overline{nr}_i + 1} = -\ln \frac{\overline{nr}_i}{nr_i} \left(1 + \frac{1}{\overline{nr}_i}\right) \\ &= w_{ji} - \ln \left(1 + \frac{1 + e^{w_{ji}}}{nr_i + \overline{nr}_i}\right). \end{aligned} \tag{11}$$

Step 5 (judge whether the action module should be changed). If reward has not been observed after given times learning, a layer will be added in action module, signifying the action chain should be more complicated. The number of the neurons in the new layer is the number of the actions allowed to be selected.

Step 6 (judge whether the learning has come to the end). If SE is low enough or the learning times have exceeded the maximum limit, then end the algorithm; otherwise, get back to Step 2.

4. Convergence Analysis of the Learning Mechanism

The learning mechanism in the model is shown in formula (10) and (11). We modify them before analysis in a briefer form.

Let t represent the learning times and A represent the action sequence to be learned, in which a_i is an action ($i = 1, 2, \ldots, n$). When $t \to \infty$, $(nr_i + \overline{nr}_i) \to \infty$, then, $(1 + e^{w_{ji}})/(nr_i + \overline{nr}_i) \to 0$, $(1 + e^{-w_{ji}})/(nr_i + \overline{nr}_i) \to 0$. According to *L'Hospital rule*, $\ln(1 + x)$ is the equivalent infinitesimal to x when $x \to 0$. Therefore, we can get the following formula by such an equivalent substitution:

$$\Delta w_{ji} = w_{ji}(t+1) - w_{ji}(t) = \begin{cases} \ln\left(1 + \dfrac{1 + e^{-w_{ji}}}{nr_i + \overline{nr}_i}\right) = \dfrac{1 + e^{-w_{ji}}}{nr_i + \overline{nr}_i}, & \text{if receiving reward} \\ -\ln\left(1 + \dfrac{1 + e^{w_{ji}}}{nr_i + \overline{nr}_i}\right) = -\dfrac{1 + e^{w_{ji}}}{nr_i + \overline{nr}_i}, & \text{if not receiving reward.} \end{cases} \tag{12}$$

Let $\mu = 1/(nr_i + \overline{nr}_i)$; then formula (12) can be transformed into the following equation:

$$\Delta w_{ji} = \begin{cases} \mu\left(1 + e^{-w_{ji}}\right), & \text{if receiving reward} \\ -\mu\left(1 + e^{w_{ji}}\right), & \text{if not receiving reward.} \end{cases} \tag{13}$$

Obviously, $\mu > 0$, so it can be regarded as the learning rate.

According to formula (13), when agents receive reward, $\Delta w_{ji} = \mu(1 + e^{-w_{ji}}) > 0$, so the weight w_{ji} between layers will increase continuously, which indicates that the correlation between the action sequence A and the reward is increasing, too. Thus, the probability of the corresponding actions being selected is also increasing. In short, those synaptic weights related to the actions, that is, more likely to bring reward, will be strengthened so that the agents will more probably choose the actions.

On the contrary, if agents do not receive reward, $\Delta w_{ji} < 0$, then the weight w_{ji} between layers will decrease continuously. Therefore, the whole process can be described like the following: if selecting those actions that are less likely to result in reward, the related synaptic weights will decrease. Then, the actions will be less likely chosen.

Another question is whether there is limitation for the change of synaptic weights. In fact, the change of synaptic weights is bounded in our model, which is in accordance with the biological fact and suggests the convergence of the model.

Let $E(\Delta w_{ji})$ represent the expected value of Δw_{ji}. When $t \to \infty$, we can obtain the following formula (14) based on formula (13):

$$E\left(\Delta w_{ji}\right) = p \cdot \mu\left(1 + e^{-w_{ji}}\right) - q \cdot \mu\left(1 + e^{w_{ji}}\right), \tag{14}$$

where p represents the probability of being rewarded while q represents the probability of not being rewarded. Obviously,

$$p = \frac{nr_i}{nr_i + \overline{nr}_i}, \qquad q = \frac{\overline{nr}_i}{nr_i + \overline{nr}_i}. \tag{15}$$

Substituting formula (15) into formula (14), we obtain another formula as follows:

$$\begin{aligned} E\left(\Delta w_{ji}\right) &= p \cdot \mu\left(1 + e^{-w_{ji}}\right) - q \cdot \mu\left(1 + e^{w_{ji}}\right) \\ &= \frac{nr_i}{nr_i + \overline{nr}_i} \cdot \mu \cdot \left(1 + \frac{\overline{nr}_i}{nr_i}\right) - \frac{\overline{nr}_i}{nr_i + \overline{nr}_i} \cdot \mu \\ &\quad \cdot \left(1 + \frac{nr_i}{\overline{nr}_i}\right) = \mu - \mu = 0. \end{aligned} \tag{16}$$

Therefore, when $t \to \infty$, $E(\Delta w_{ji}) = 0$, which means the synaptic weight w_{ji} will stop changing, neither increase nor decrease. Therefore, the boundation of weights is proved.

Besides, we can draw the same conclusion by analyzing the self-organization feature of the model. As mentioned above, we use the concept *system entropy* (SE) to describe the feature of self-organization. When SE decreases, the degree of self-organization increases; that is, the model is converging and the change of weights is becoming less.

Suppose there are n sequences of actions, among which A_i is the one with reward while other sequences A_j ($j = 1, 2, \ldots, n$, $j \neq i$) are those without reward. p_i represents the probability of being selected for A_i, while p_j is the probability of being selected for other sequence actions A_j ($j = 1, 2, \ldots, n$, $j \neq i$). Thus, we can get formula (17):

$$1 - p_i = 1 - \frac{N_i}{N} = \frac{N - N_i}{N}, \tag{17}$$

where N_i represents the number of times being selected for A_i and N is the total number of times for all action sequences, $N = \sum_{i=1}^{n} N_i$.

When $t \to \infty$, $N \to \infty$, as A_i more possibly results in reward, its number of times being selected will increase constantly while others will decrease; that is, $N_i \to \infty$ and $N_j \to 0$ ($j \neq i$).

Thus, $(N - N_i) \to 0$ when $N \to \infty$. Then, $1 - p_i = (N - N_i)/N \to 0$; that is, $p_i \to 1$ and $p_j \to 0$.

Therefore, we can get new system entropy as follows:

$$\mathrm{SE} = -p_i \log p_i - \sum_{j=1, j\neq i}^{n} p_j \log p_j = -1 * 0 - 0 = 0. \tag{18}$$

Formula (18) illustrates that SE will decrease to the minimum value when $t \to \infty$, which indicates that the system is self-organized.

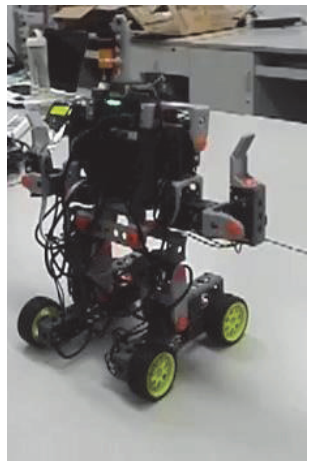


Figure 3: The real robot used in experiment I: *Cogbot I*. *Cogbot I* is a humanoid robot with 4 wheels. It has 1 infrared sensor and 1 camera. It is also equipped with a small LED screen, which can display the weights in the network.

5. Experiments Design and Analysis

To evaluate our model, we reproduce two classic animal experiments in associative learning. The first experiment is Pavlov's dog experiment, which is concerned with classical conditioning, while the other experiment is Thorndike's cat experiment, which is concerned with operant conditioning. We choose the two experiments because of their powerful influence and great fame in associative learning theory. Both experiments are reproduced in real robots, which echoed the embodied cognition.

5.1. Classical Conditioning Experiment: Pavlov's Dog Experiment. To study the mechanisms underlying the digestive system in animals, Pavlov and Anrep carried out a series of experiments [12]. In the most famous one, Pavlov and Anrep's dog experiment, he found that if a bell was sounded in very close association with dogs' meal for several times, the dogs learned to associate the bell sound with meal; that is, they would drool even if there is no food available. The phenomenon in the experiment is called classical conditioning, or Pavlovian conditioning. The procedure described above is acquisition of classical conditioning. Meanwhile, Pavlov and Anrep also found that if the dogs acquiring classical conditioning did not get food after the bell sounded for several times, the dogs would gradually forget the association between the bell sound and meal, that is, the extinction of classical conditioning. We reproduce both acquisition and extinction of classical conditioning in our model.

5.1.1. Experiment Design. We carry out the whole experiment (including acquisition phase and extinction phase) in the real robot *Cogbot I*. *Cogbot I* is a humanoid robot with an infrared sensor and a camera, shown in Figure 3.

The infrared sensor and the camera compose the sensory module of the system, in which the infrared sensor represents the dog's ears while the camera represents the dog's eyes. Therefore, an infrared signal, for example, shaking hands near the sensor, represents the sound of bell and works as the conditioned stimulus (CS), while a yellow ball represents food and works as the unconditioned stimulus (US).

Moreover, the registers or buffers in the sensors correspond to the sensory memory. They store the sensory information transiently and provide it to the working memory.

Working memory and the action module compose the core network, whose structure is shown in Figure 4. There are 2 neurons in the working memory: one stores the information related to the sound stimulus, denoted as $\mathrm{wm}_{\mathrm{bell}}$, while the other one stores the information related to the sight stimulus, denoted as $\mathrm{wm}_{\mathrm{see\_food}}$. The outputs of both neurons indicate that the robot has received corresponding stimuli. For example, if the output of $\mathrm{wm}_{\mathrm{bell}}$ is 1, it suggests that the robot *hears* the sound of bell. On the contrary, if the output of $\mathrm{wm}_{\mathrm{bell}}$ is 0, it indicates that the robot *does not hear* any sound. In the action module, there is only 1 neuron corresponding to the action salivation, denoted as a_{salivate}. In order to make the action visible, we use the action of bending back to represent it. Similarly, its output shows whether the dog salivates: 1 means yes and 0 means no.

The long-term memory of the system stores the learning results, mainly the synaptic weights of the core network. Its initial contents are different in acquisition experiment and extinction experiment. For example, in acquisition phase, ω_1 in the long-term memory is initially set to be 0, symbolizing the robot has not associated the sound of the bell with the presentation of food. On the contrary, in extinction phase, the initial value of ω_1 is positive, symbolizing the robot has learned the association. Each time at the beginning of learning, the contents in the long-term memory will be loaded to the working memory.

The reward signal produced by simulated VTA dopaminergic neurons is designed in this way: as food can bring satisfaction to the dog, we think that the dog's statuses will improve if food is presented; that is, the ideal degree of the statuses will increase. Therefore, the reward signal will be positive according to formula (4) and (5). In short, the dog will be rewarded if food is presented. Otherwise, it will not.

In acquisition phase, as the agent gets reward, both of the synaptic weights ω_1 and ω_2 will increase according to formula (13). That is,

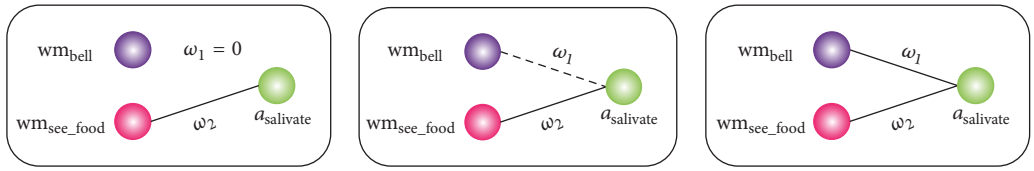
$$\omega(t+1) = \omega(t) + \mu\left(1 + e^{-\omega(t)}\right). \tag{19}$$

On the contrary, in extinction phase, both of the synaptic weights ω_1 and ω_2 will decrease because of no reward according to formula (20):

$$\omega(t+1) = \omega(t) - \mu\left(1 + e^{\omega(t)}\right). \tag{20}$$

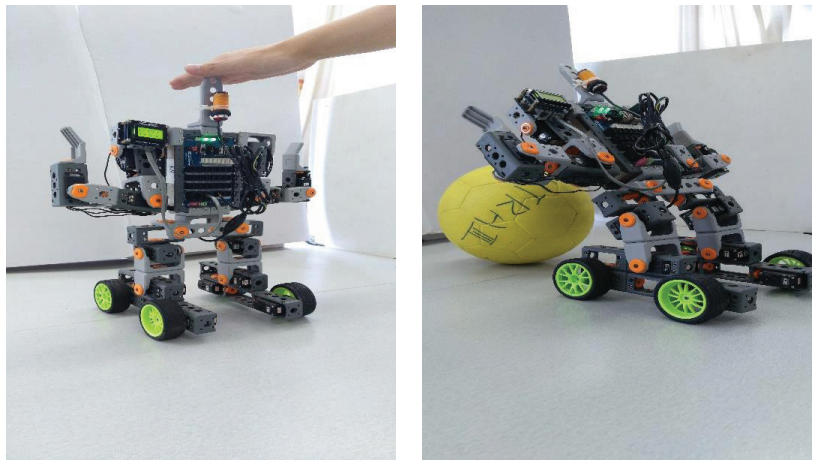
Although both weights change, only the change of ω_1 is observed and recorded in the experiment as it is the key reason for the explanation of the phenomenon.

We use the learning times as the ending condition. In both acquisition and extinction experiments, the robot has to learn 50 times. After that, the experiments come to an end.



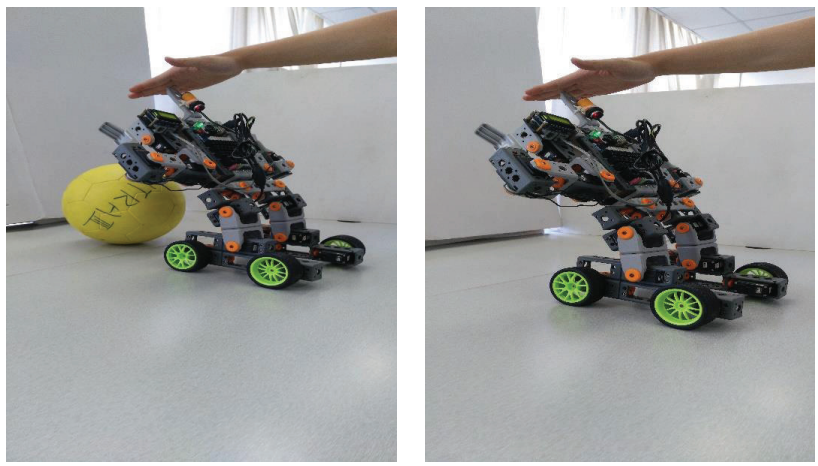
(a) At the beginning (b) During experiment (c) In the end

FIGURE 4: The structure of core network in different phases of acquisition experiment. ω_1 and ω_2, respectively, represent the synaptic weights between the neurons in working memory and action module. At the beginning of the experiment, there is no connection between wm_{bell} and $a_{salivate}$, suggesting that the dog will not salivate when it hears the bell sound alone. Thus, ω_1 is 0. Then, it gradually increases, indicating a synaptic connection between wm_{bell} and $a_{salivate}$ appears. As it is not big enough to trigger off salivation, the synaptic connection is represented by dash line instead of solid one in (b). At the end of the experiment, ω_1 has increased a lot, so the connective line between two neurons becomes full in (c). During the whole process, ω_2 increases too according to the learning mechanism. However, since the change of ω_2 is not the highlight of the experiment, it is ignored in figure.



(a) The initial stage: only presenting the infrared signal

(b) The initial stage: only presenting the visual stimulus (yellow ball)



(c) The intermediate stage: the infrared signal and yellow ball are presented together

(d) The final stage: the robot is bending back with only the infrared signal

FIGURE 5: The whole process of the acquisition experiment in *Cogbot I*.

5.1.2. Results of Acquisition Experiment. The process of acquisition is the one that the connection between the sound stimulus and the salivation response is being established. The dog did not salivate at the beginning of the experiment when it heard the bell sound alone. Then, it was fed every time along with the sound. After a few times of such trials, the dog would salivate even if it only heard the sound of the bell.

Such process is reproduced in *Cogbot I*, shown in Figure 5. At first, only the infrared signal presents, and the robot has no reaction at all. Then, when the yellow ball and the infrared

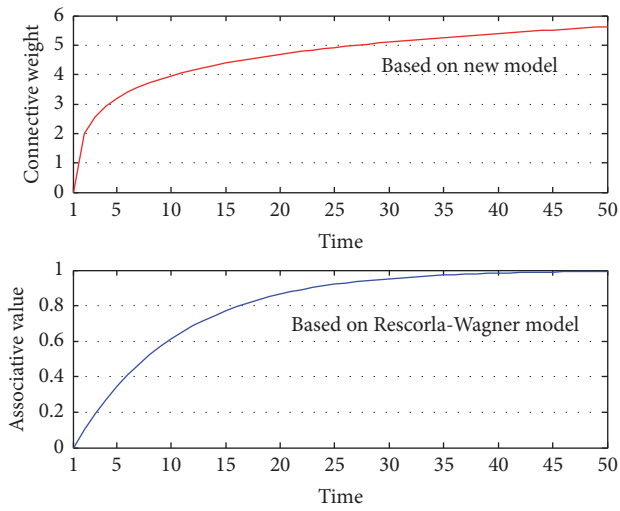


FIGURE 6: The change of ω_1 in the acquisition experiment and the comparison with Rescorla-Wagner model.

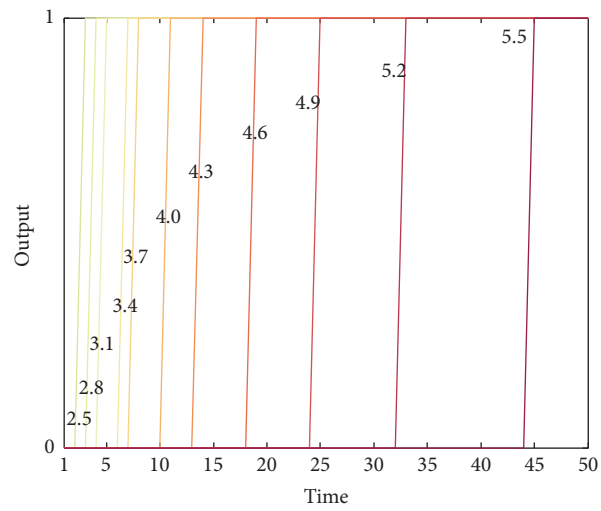


FIGURE 7: The output of a_{salivate} under different threshold values in acquisition experiment. We choose 11 different values in the interval $[2.5, 5.5]$. All the values form an arithmetic progression with common difference 0.3. All b values are marked beside the corresponding lines. The color of the lines indicates the values: the darker the color is, the bigger the b value is.

signal are presented together, the robot bends back since the US is presented. Finally, after a few times of trials, the weight is big enough to excite the corresponding action. Therefore the robot begins to bend back even without the yellow ball.

The reason behind the phenomenon lies in the change of the synaptic weight ω_1. As analyzed Section 5.1.1, the weight ω_1 will continuously increase in acquisition experiment until the end of the experiment. The change of ω_1 is recorded and compared with the data of Rescorla-Wagner model [2] (the learning rate is 0.1), shown in Figure 6.

Despite different meanings of these data (our model records connective weight while Rescorla-Wagner model records associative value), the tendencies reflected by the two models are the same. During the process, both connective weight and associative value continuously increase, which indicates the agent has gradually learned to associate CS with the reward and classical conditioning is being acquired.

According to formula (2), only when ω_1 is bigger than the threshold value of the neuron a_{salivate}, denoted as b, can the action neuron be excited. Therefore, the factor which decides whether the action will be executed without the presentation of food is the value of b. At the beginning of the experiment, ω_1 is quite small and does not exceed b so that the neuron a_{salivate} will not be excited and the action will not be executed, either. However, as ω_1 continuously increases, it will exceed b after a few times of learning. Then, the action can be executed even without food.

We discuss the influence of different threshold values on the output of a_{salivate}, shown in Figure 7. The figure shows that the less b is, the more easily a_{salivate} is excited; that is, its output becomes 1. Since the model is proved to converge, the increment of ω_1 is less and less during the experiment. Thus, the gaps between neighboring lines in Figure 7 become wider and wider.

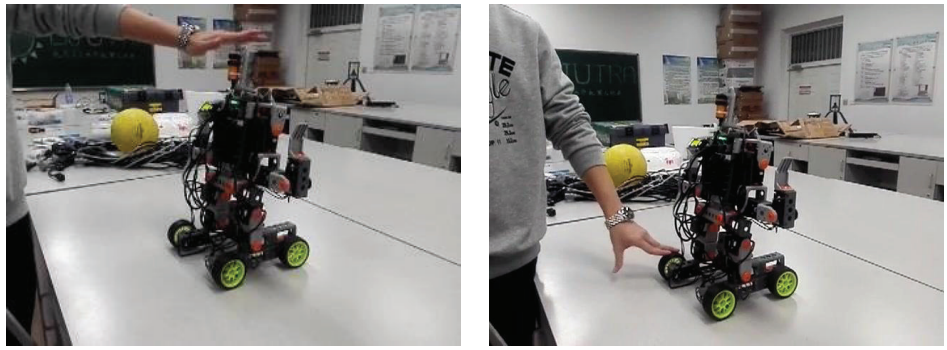
5.1.3. Results of Extinction Experiment. Contrary to acquisition phase, extinction phase is the process during which the association between CS and the conditioned response gradually disappears. In Pavlov's dog experiment, if CS, that is, the bell sound, is presented alone without food repeatedly, the conditioned response, that is, salivation, will be no longer watched.

We reproduce it in the robot *Cogbot I*. The extinction experiment is done right after the acquisition experiment. Therefore, the robot at the beginning of the extinction experiment acquires the classical conditioning. Then, we present CS (the infrared signal) alone without US (the visual signal) every time we make experiments. At the end of the experiment, the robot does nothing when CS is presented alone, shown in Figure 8.

The phenomenon in the extinction experiment can also be explained from the point of the comparison between ω_1 and b. In the extinction experiment, ω_1 at first is so great that it exceeds b and the neuron a_{salivate} is excited to execute the conditioned response. However, since there is no food presented during the experiment, the synaptic weight ω_1 decreases gradually according to formula (20), suggesting that the association between CS and the conditioned response is being weakened. At certain moment, ω_1 decreases to be less than b. Then, the conditioned response cannot be executed any longer.

We record the change of ω_1 during the experiment and compare it with that of Rescorla-Wagner model, shown in Figure 9. Both the models show a similar process during which the association between CS and the conditioned response is diminishing.

We also discuss the influence of different b in the experiment. Figure 10 shows the result. Obviously, the bigger the b value is, the faster the extinction procedure is. Figure 10 still shows a similar phenomenon: the gaps between lines are



(a) Presenting the infrared signal alone (b) The robot does not bow back in the end

FIGURE 8: The results of the extinction experiment in real robot *Cogbot I*.

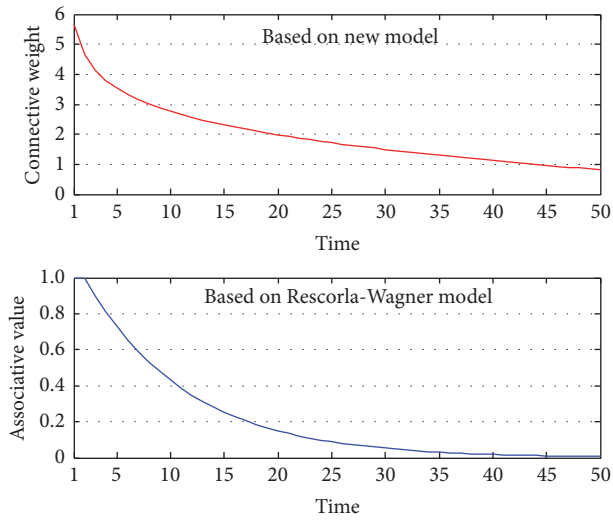


FIGURE 9: The change of ω_1 in the extinction experiment and the comparison with Rescorla-Wagner model.

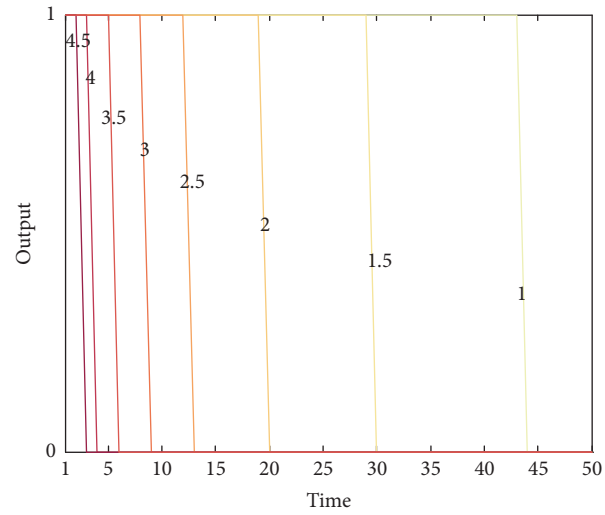


FIGURE 10: The output of a_{salivate} under different threshold values in extinction experiment. All the threshold values are in the interval $[1, 4.5]$ with common difference 0.3. We mark all b values on the corresponding lines. The color of the lines indicates the values: the darker the color is, the bigger the b value is.

getting wider and wider, suggesting the decreasing speed of weight ω_1 is slowing down as the experiment continues.

5.2. Operant Conditioning Experiment: Thorndike's Cat Experiment. Another type of associative learning is operant conditioning, or instrumental conditioning. Whereas classical conditioning focuses on the association between conditioned stimulus (CS) and conditioned response, operant conditioning involves learning from the consequences of the behavior. Operant conditioning principles presented by Skinner [41] suggest that the behaviors which result in reward tend to be repeated by animals while the behaviors without reward tend to be avoided.

However, Skinner was not the first psychologist to study operant conditioning. Indeed, Skinner's theory on operant conditioning is developed on the ideas of Thorndike. Thorndike formally studied operant conditioning and reward learning back in the late 1800s. He designed and carried out a lot of animal learning experiments, among which the escape experiment of a cat in a puzzle box is the most famous one.

In the experiment, a cat was put in a puzzle box designed by Thorndike. The cat was encouraged to escape to reach a piece of fish placed outside. To go outside the box, it had to firstly press a pedal and then lift the latch of the box. Only when it finished the series of actions in right order could it escape successfully.

Thorndike executed the experiment many times and summarized the results in his learning theory. One is *Law of Effect*, which states that the connections between situations and responses followed by satisfaction are strengthened while the connections with discomfort are weakened. For example, the cat in the experiment would tend to repeat the right series of actions once it found executing such series could bring reward. In fact, the idea of *Law of Effect* is totally in accord with Skinner's operant conditioning theory. Another one is *Law of Exercise*, which states that connections between stimuli and response become strengthened with practice and weakened if practice is not continued. For example, in the

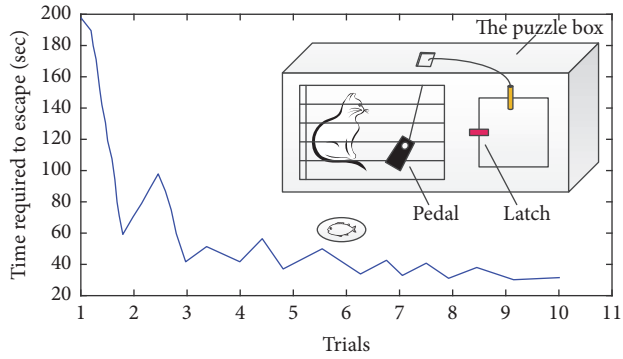


Figure 11: Thorndike's puzzle box and the learning curve in the cat experiment.

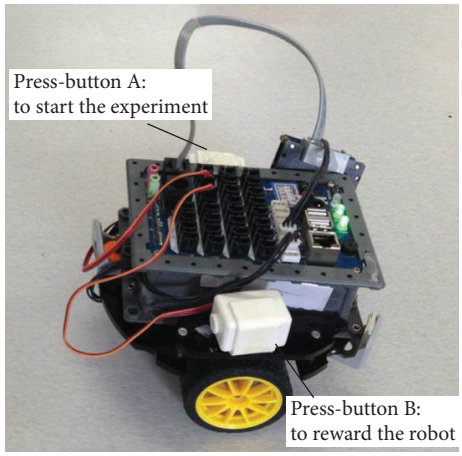


Figure 12: The real robot in the operant conditioning experiment: Cogbot II. Cogbot II is a wheeled robot with 2 press-buttons, among which press-button A is used to start the experiment while press-button B is used to reward the robot.

cat experiment, Thorndike found that the cat more and more adroitly escaped from the box after it grasped the right method. The time it spent in escaping each time tended to decrease. Thorndike recorded the time and drew a picture, called *learning curve*. Figure 11 is the learning curve and the puzzle box of the experiment.

Thorndike's cat experiment is not only a good example for both Law of Effect and Law of Exercise, but also a good example for operant conditioning of animals. Therefore, we reproduce the experiment in real robot *Cogbot II* based on our model.

5.2.1. Experiment Design. The operant conditioning experiment is done in the real robot *Cogbot II*, shown in Figure 12.

We simplify the original settings of the cat experiment in the following way: An apple, corresponding to the fish in Thorndike's cat experiment, is put in the north-east of the robot, shown in Figure 13. After we push down press-button A, the experiment starts. The robot can move northward or move eastward. Firstly moving northward then eastward is considered as the only right way to get reward. Each time

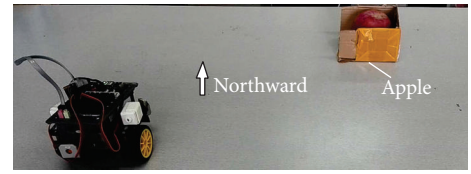


Figure 13: The scene settings of operant conditioning experiment. The aim of the robot is to go northward then eastward to reach the apple.

Table 1: The correspondence between the operant conditioning experiment and the original cat experiment.

The operant conditioning experiment	The original cat experiment
Robot	Cat
Apple	Fish
Move northward	Press the pedal
Move eastward	Lift the latch

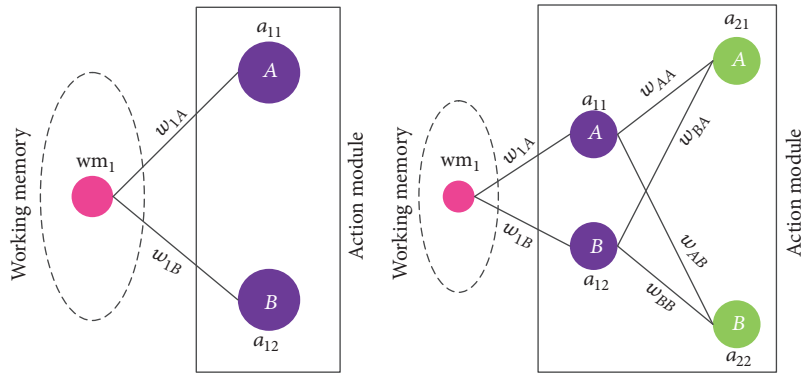
when the robot chooses a direction, the action will last for 2.4 seconds or 3.5 seconds (if the action includes turning). The speed of the robot and the position of the apple are set just fine to allow the robot to reach the apple in a complete northward-eastward action sequence. If the robot happens to go in the right way, press-button B will be pushed down, representing the robot has got the reward.

In short, the correspondence between the experiment and the original Thorndike's cat experiment may be listed in Table 1.

In the experiment, press-button B composes the sensory module of the system. It is press-button B that the robot depends on to sense the reward. If it is pushed down, it represents that the robot feels the reward. Otherwise, it does not. Similarly, the registers or buffers in the press-button serve as the sensory memory.

The working memory receives information from the sensory memory and codes it in a suitable form for the following operation. In this experiment, there is 1 neuron set in the working memory, denoted as wm_1. Its output values are 1, 2, and 3, respectively, symbolizing the 3 statuses of the cat, that is, hungry, half-hungry, and full up.

As mentioned in the second paragraph of this section, each time the robot can choose 2 actions: move northward or eastward. Therefore, 2 neurons in the action module are set to represent them, denoted as a_{11} and a_{12}. Among them, a_{11} corresponds to moving northward, and a_{12} corresponds to the other. If an action is selected, the corresponding neuron's output is 1; otherwise it is 0. Since the robot learns from single action, there is only 1 layer of neurons in the action module at the beginning of the experiment, shown in Figure 14(a). If the robot is not rewarded all through the single-action-learning phase (set as learning 30 times in this experiment), a new layer consisting of 2 neurons will be added in the action module (shown in Figure 14(b)), which symbolizes that the robot realizes the single action learning does not work and begins to learn more complicated action series. Meanwhile,



(a) In the phase of learning single action (b) In the phase of learning action series

Figure 14: The structure of the core network in the operant conditioning experiment. In the phase of learning single action, there is only 1 layer of neurons in the action module, while there are 2 layers in the module representing the action sequence including 2 actions. The synaptic weights between the working memory and the action module are denoted as w_{1A} and w_{1B}, while the weights between the 2 layers of the action module are denoted as w_{AA}, w_{AB}, w_{BA}, and w_{BB}, respectively, as shown in the picture.

Table 2: State transitions in Thorndike's cat experiment.

	a_1	a_2	a_1a_2	a_2a_1
$wm_1 = 1$	$wm_1 = 1$	$wm_1 = 1$	$wm_1 = 2$	$wm_1 = 1$
$wm_1 = 2$	$wm_1 = 1$	$wm_1 = 1$	$wm_1 = 3$	$wm_1 = 1$
$wm_1 = 3$	$wm_1 = 2$	$wm_1 = 2$	$wm_1 = 3$	$wm_1 = 2$

all neurons in the action module are considered to be excited as long as they receive the outputs of other neurons; that is, the threshold values of the neurons are set to be minus infinity. Every time when the robot executes an action or a sequence of actions, the results for the action or the sequence, that is, the number of times of being rewarded or not, will be saved in the working memory.

Long-term memory stores the results of last time learning. Every time when experiments start, the content in the long-term memory will be loaded to the working memory. When the robot first learns, all of the weights are initialized as 0, indicating that it has no experience to make advantage of.

The reward signal produced by the simulated VTA dopaminergic neurons is computed in the following way.

Firstly, the ideal degree for each state is defined as the value of the corresponding output of wm_1; that is, the ideal degrees for the state of being hungry, half-hungry, and full are, respectively, 1, 2, and 3.

Each time when the robot chooses an action or a sequence of actions, the states will be transformed in accordance with Table 2. The first line in the table represents all possible action combinations including single action and action series, while the first column represents all the statuses of the cat before executing actions. The entries in the table show the states after executing the actions. For example, the entry on line 1 at column 4 indicates that the hungry cat will be still hungry ($wm_1 = 1$) if it chooses the wrong action series a_2a_1.

Let $k = 1$; then we can get the reward signal δ according to formula (4) and (5) based on the state transition and the definition of ideal degree for each state. In general, if the robot chooses the right sequence of actions, the reward signal will be positive; otherwise it will be negative.

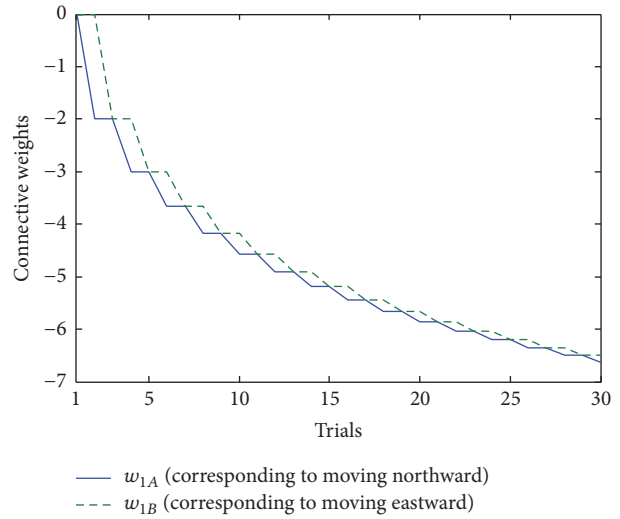


Figure 15: Change of connection weights between neurons in single-action-learning phase.

In single-action-learning phase, the robot is set to try 30 times. Then, it switches to learn action sequence, which includes 100 trials.

5.2.2. Results and Analysis. As mentioned above, the learning process can be divided into 2 phases: single action learning and action sequence learning. Therefore, experiment results will be listed by stages in this section. The following analysis shows that the results can serve as the evidences of *Law of Effect* and *Law of Exercise*.

In single-action-learning phase, no matter which action the robot chooses, it will not get rewarded. Hence, according to the learning mechanism, neither of the two weights related to the actions will increase. Figure 15 shows how the weights change in the phase.

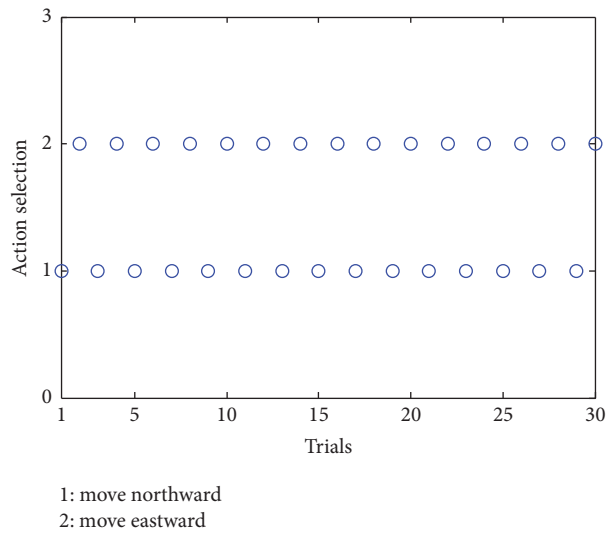
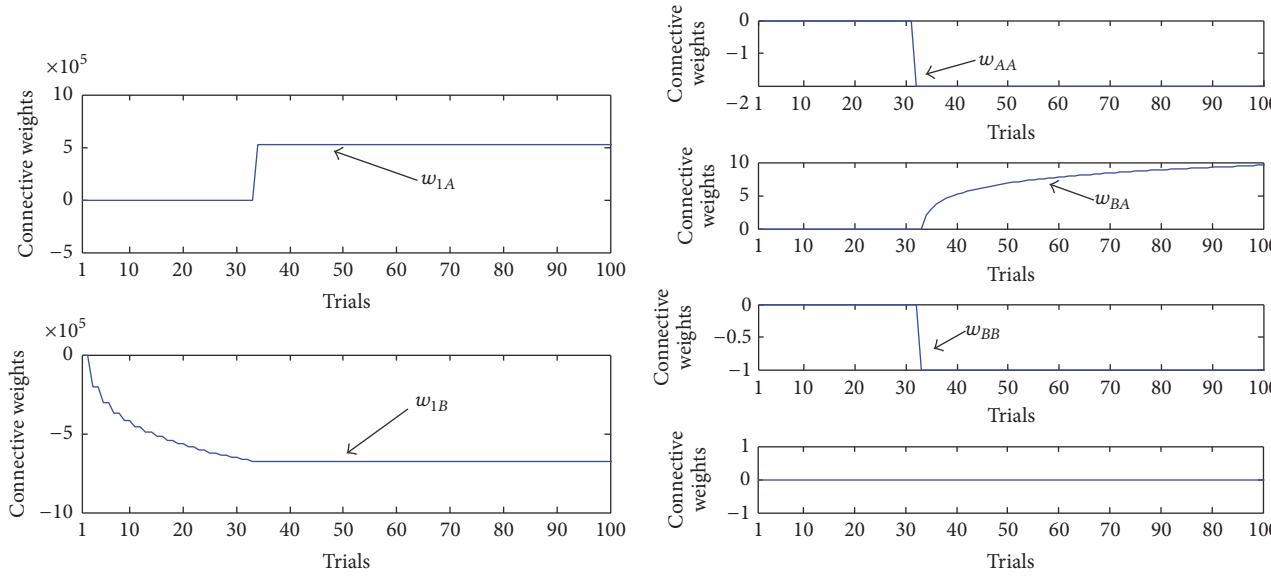


Figure 16: Action selection in single-action-learning phase. 1 represents the fact that the robot chooses to move northward, while 2 represents that it chooses to move eastward.



(a) The change of w_{1A} and w_{1B}

(b) The change of w_{AA}, w_{AB}, w_{BA}, and w_{BB}

Figure 17: The change of the connective weights of the core network in action-sequence-learning phase.

As shown in Figure 15, since no reward follows the two actions, the weights continually decrease from the original value 0 every time when the related actions are executed.

Indeed, the results in Figure 15 testify one aspect of *Law of Effect*. As the theory states, the connections between the stimulus and the response will be weakened if there is no reward. The synaptic weights between working memory and action module actually symbolize the connections between the stimulus and the response; so what is shown in Figure 15 is completely consistent with the theory.

As a result, such changes influence how the robot chooses actions. Figure 16 shows the change of the selections in single-action-learning phase. The alternate decreases of the weights make the robot choose the actions in turn during the learning process, as it selects actions in winner-take-all way.

After 30 times of trials, the robot stops learning single action and begins to learn action sequence. The structure of the action module is becoming more complicated and transformed as shown in Figure 14(b).

The change of the synaptic weights is shown in Figure 17. Obviously, only when the robot chooses the right sequence, that is, moving northward first then eastward, can it get reward. Therefore, only the synaptic weights related to this action sequence will increase while other synaptic weights decrease or keep unchanged for the corresponding actions have not been selected all the time.

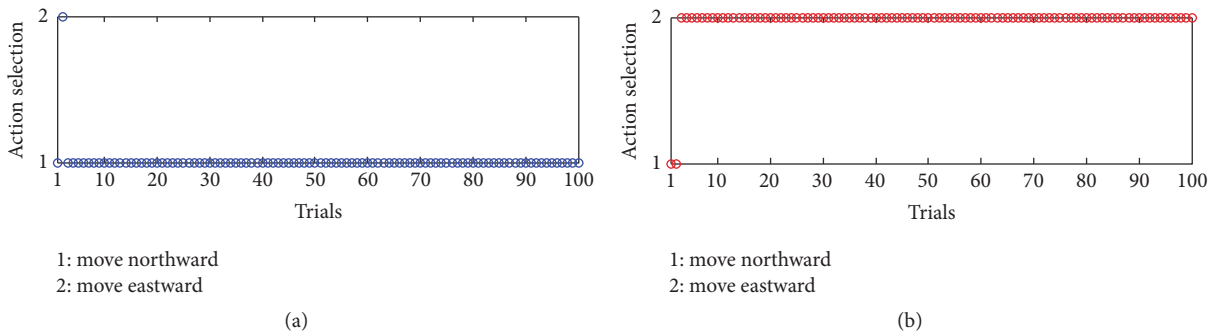


Figure 18: Action selection in serial-action-learning phase. (a) records the change of the first action in the sequence, while (b) is about the change of the second action. The numbers in the figure, 1 or 2, represent the two actions the robot can choose each time.

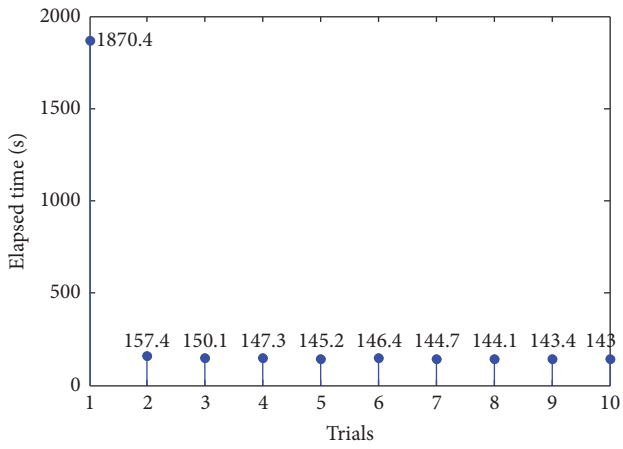


Figure 19: The elapsed time in each experiment.

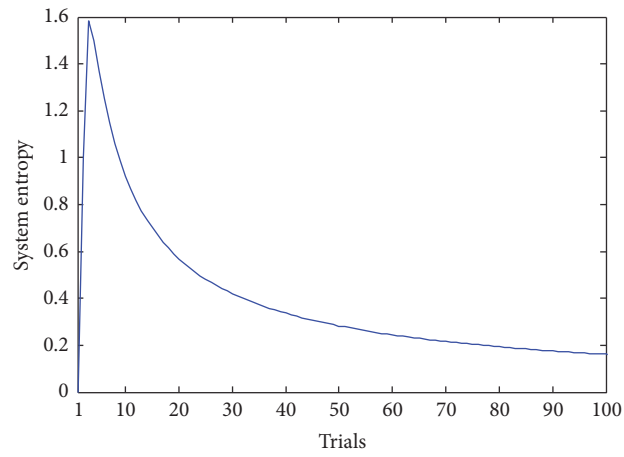


Figure 20: The change of system entropy in action-sequence-learning phase.

Similarly, the changes of the weights influence the robot's decision. Figure 18 illustrates how the robot chooses actions during the phase.

At the beginning of the experiment, since the robot has not learned the association between actions and reward, it makes some wrong decisions. For example, at the first trial, the robot chooses to move northward two times. The second trial is not right, either. However, from the third trial, the robot makes the right choice. As the right action sequence brings about reward and satisfaction, the robot repeatedly chooses the same sequence until the end.

Thus, results in Figures 17 and 18 testify the other aspect of *Law of Effect*, which states that the connections between stimuli and responses will be strengthened if the responses are followed by reward, and those responses or actions will tend to repeat.

We also record the spending time of each experiment to validate *Law of Exercise*. When the experiment begins, the timer starts. System entropy, denoted as SE, is calculated during the whole procedure. If SE < 0.3 is observed, the experiment ends and the timer is stopped. Figure 19 shows the results.

Figure 19 illustrates that the time spent in the experiment is inclined to decrease over time. The robot spent a lot of time in the first trial as it has no experience. However, along with the progress of the experiment, the weights related to the right action sequence were continuously strengthened. Moreover, with the help of the memory mechanism, it had more experience to speed up learning so that it became more and more adept in this task. Such changes are completely consistent with *Law of Exercise*, which states that practice makes perfect.

The whole procedure including single-action-learning phase and action-sequence-learning phase is unsupervised. All the learning behaviors only depend on the interactions between the agent and the environment. Meanwhile, it is also a self-organized procedure in which the structure of the neural network changes autonomously until it converges.

As mentioned above, the degree of self-organization can be measured by system entropy. Figure 20 shows how the system entropy changes during the second phase. The decrease of the system entropy in the phase indicates that the system is changing from disorder to order so that the degree of self-organization increases.

6. Conclusion

We have presented a cognitive model based on neuromodulated synaptic plasticity on the issue surrounding associative learning. We apply it to reconstructing two famous conditioning experiments. The results of the experiments in real robots prove the suitability and validity of the proposed cognitive model in different learning tasks. The results also prove the idea that both classical conditioning and operant conditioning are able to be unified in a general frame. The two types of conditioning share a similar neural mechanism and can be unified at the level how stimulus and response connect and how the connections change in the environment.

Moreover, the statistical feature of our model indicates that associative learning is a kind of statistical learning. Some research reports that certain statistical principles like Bayesian rule possibly work in associative learning, in accordance with this study [42].

Finally, this study shows that associative learning is self-organized. As the learning mechanism is unsupervised, the synaptic connections between neurons, or the associative strengths between stimulus and response, change and develop in a self-organized way until new organization forms, signifying the convergence of the model and the emergence of intelligence.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors acknowledge support from National Natural Science Foundation of China (no. 61375086, no. 61573029); Key Project of S&T Plan of Beijing Municipal Commission of Education (no. KZ201610005010); Young Excellent Talent Program of Beijing Institutions (YETP 1610).

References

[1] M. Domjan, *The Principles of Learning and Behavior*, Nelson Education, Scarborough, Canada, 2014.

[2] R. A. Rescorla and A. R. Wagner, "A theory of Pavlovian conditioning: the effectiveness of reinforcement and non-reinforcement," in *Classical Conditioning II: Current Research and Theory*, A. H. Black and W. F. Prokasy, Eds., pp. 64–69, Appleton-Century-Crofts, New York, NY, USA, 1972.

[3] C. R. Raymundo and C. G. Johnson, "An artificial synaptic plasticity mechanism for classical conditioning with neural networks," in *Proceedings of the International Symposium on Neural Networks (ISNN '14)*, pp. 213–221, Springer, November-December 2014.

[4] M. Ziegler, R. Soni, T. Patelczyk et al., "An electronic version of Pavlov's dog," *Advanced Functional Materials*, vol. 22, no. 13, pp. 2744–2749, 2012.

[5] A. C. Courville, N. D. Daw, and D. S. Touretzky, "Similarity and discrimination in classical conditioning: a latent variable account," *Advances in Neural Information Processing Systems*, vol. 17, pp. 313–320, 2005.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2011.

[7] J. Huang, X. Ruan, L. Li, R. Wei, Q. Fan, and X. Wu, "Operant conditioning learning model based on BP network," in *Proceedings of the 33rd Chinese Control Conference (CCC '14)*, pp. 8386–8390, IEEE, Nanjing, China, July 2014.

[8] T. Taniguchi and T. Sawaragi, "Incremental acquisition of behaviors and signs based on a reinforcement learning schemata model and a spike timing-dependent plasticity network," *Advanced Robotics*, vol. 21, no. 10, pp. 1177–1199, 2007.

[9] E. Y. Cheu, C. Quek, and S. K. Ng, "ARPOP: an appetitive reward-based pseudo-outer-product neural fuzzy inference system inspired from the operant conditioning of feeding behavior in aplysia," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 317–329, 2012.

[10] A. Cyr, M. Boukadoum, and F. Thériault, "Operant conditioning: a minimal components requirement in artificial spiking neurons designed for bio-inspired robot's controller," *Frontiers in Neurorobotics*, vol. 8, article 21, pp. 1–32, 2014.

[11] T. V. Maia, "Reinforcement learning, conditioning, and the brain: successes and challenges," *Cognitive, Affective & Behavioral Neuroscience*, vol. 9, no. 4, pp. 343–364, 2009.

[12] I. P. Pavlov and G. V. Anrep, *Conditioned Reflexes*, Courier Corporation, 2003.

[13] E. L. Thorndike, "Animal intelligence: an experimental study of the associative processes in animals," *The Psychological Review: Monograph Supplements*, vol. 2, no. 4, p. i-109, 1898.

[14] T. J. Carew, E. T. Walters, and E. R. Kandel, "Classical conditioning in a simple withdrawal reflex in *Aplysia californica*," *The Journal of Neuroscience*, vol. 1, no. 12, pp. 1426–1437, 1981.

[15] B. Brembs, F. D. Lorenzetti, F. D. Reyes, D. A. Baxter, and J. H. Byrne, "Operant reward learning in Aplysia: neuronal correlates and mechanisms," *Science*, vol. 296, no. 5573, pp. 1706–1709, 2002.

[16] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*, Psychology Press, 2005.

[17] K. D. Cantley, A. Subramaniam, H. J. Stiegler, R. A. Chapman, and E. M. Vogel, "Hebbian learning in spiking neural networks with nanocrystalline silicon TFTs and memristive synapses," *IEEE Transactions on Nanotechnology*, vol. 10, no. 5, pp. 1066–1073, 2011.

[18] W. Gerstner, W. M. Kistler, R. Naud et al., *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*, Cambridge University Press, Cambridge, UK, 2014.

[19] T. Manninen, K. Hituri, J. H. Kotaleski, K. T. Blackwell, and M.-L. Linne, "Postsynaptic signal transduction models for long-term potentiation and depression," *Frontiers in Computational Neuroscience*, vol. 4, article 152, 2010.

[20] A. Soltoggio and K. O. Stanley, "From modulated Hebbian plasticity to simple behavior learning through noise and weight saturation," *Neural Networks*, vol. 34, pp. 28–41, 2012.

[21] L. M. Giocomo and M. E. Hasselmo, "Neuromodulation by glutamate and acetylcholine can change circuit dynamics by regulating the relative influence of afferent input and excitatory feedback," *Molecular Neurobiology*, vol. 36, no. 2, pp. 184–200, 2007.

[22] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[23] R. Legenstein, D. Pecevski, and W. Maass, "A learning theory for reward-modulated spike-timing-dependent plasticity with

application to biofeedback," *PLoS Computational Biology*, vol. 4, no. 10, Article ID e1000180, 2008.

[24] V. Pawlak and J. N. D. Kerr, "Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity," *The Journal of Neuroscience*, vol. 28, no. 10, pp. 2435–2446, 2008.

[25] B. Porr and F. Wörgötter, "Learning with 'relevance': using a third factor to stabilize Hebbian learning," *Neural Computation*, vol. 19, no. 10, pp. 2694–2719, 2007.

[26] T. Yang and M. N. Shadlen, "Probabilistic reasoning by neurons," *Nature*, vol. 447, no. 7148, pp. 1075–1080, 2007.

[27] J. I. Gold and M. N. Shadlen, "Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward," *Neuron*, vol. 36, no. 2, pp. 299–308, 2002.

[28] M. Pfeiffer, B. Nessler, R. J. Douglas, and W. Maass, "Reward-modulated Hebbian learning of decision making," *Neural Computation*, vol. 22, no. 6, pp. 1399–1444, 2010.

[29] B. Nessler, M. Pfeiffer, and W. Maass, "Hebbian learning of Bayes optimal decisions," in *Advances in Neural Information Processing Systems*, pp. 1169–1176, 2009.

[30] S. M. Jaeggi, M. Buschkuehl, P. Shah, and J. Jonides, "The role of individual differences in cognitive training and transfer," *Memory & Cognition*, vol. 42, no. 3, pp. 464–480, 2014.

[31] F. I. M. Craik and R. S. Lockhart, "Levels of processing: a framework for memory research," *Journal of Verbal Learning and Verbal Behavior*, vol. 11, no. 6, pp. 671–684, 1972.

[32] D. E. Rumelhart and J. L. McClelland, *The PDP Research Group: Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Foundations, 1986.

[33] R. P. Kesner and E. T. Rolls, "A computational theory of hippocampal function, and tests of the theory: new developments," *Neuroscience & Biobehavioral Reviews*, vol. 48, pp. 92–147, 2015.

[34] R. C. Atkinson and R. M. Shiffrin, "Human memory: a proposed system and its control processes," *Psychology of Learning and Motivation*, vol. 2, pp. 89–195, 1968.

[35] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *ERE WESCON Convention Record*, part 4, pp. 96–104, IRE, New York, NY, USA, 1960.

[36] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[37] R. C. Malenka, E. J. Nestler, and S. E. Hyman, "Chapter 6: widely projecting systems: monoamines, acetylcholine, and orexin," in *Molecular Neuropharmacology: A Foundation for Clinical Neuroscience*, A. Sydor and R. Y. Brown, Eds., pp. 147–157, McGraw-Hill Medical, New York, NY, USA, 2nd edition, 2009.

[38] W. Schultz, "Getting formal with dopamine and reward," *Neuron*, vol. 36, no. 2, pp. 241–263, 2002.

[39] A. Lak, W. R. Stauffer, and W. Schultz, "Dopamine prediction error responses integrate subjective value from different reward dimensions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 6, pp. 2343–2348, 2014.

[40] H. M. Bayer and P. W. Glimcher, "Midbrain dopamine neurons encode a quantitative reward prediction error signal," *Neuron*, vol. 47, no. 1, pp. 129–141, 2005.

[41] B. F. Skinner, *The Behavior of Organisms: An Experimental Analysis*, BF Skinner Foundation, 1990.

[42] L. P. Sugrue, G. S. Corrado, and W. T. Newsome, "Choosing the greater of two goods: neural currencies for valuation and decision making," *Nature Reviews Neuroscience*, vol. 6, no. 5, pp. 363–375, 2005.