2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Serotonin and Dopamine Systems: Internal Areas and Sequential Tasks

Dongshu Wang, Yihai Duan, and Juyang Weng

Abstract— **Serotonin and dopamine transmitters are synthesized in the lower brain but are transmitted widely to many areas of the brain. Emergent representations are critical in understanding their effects. In our prior work [26], their effects on internal, non-motor neurons are studied for pattern recognition tasks only. In this paper, we study their effects on sequential tasks — robot navigation with different settings. They are sequential tasks because the outcome of behavior depends on not only the current behavior as in pattern recognition but also the previous behaviors (e.g., previous navigational trajectories). Analytically, we show that the serotonin and dopamine systems affect the performance of sequential tasks in a compounded way. Experimentally, we show that the effect on the learning rate of feature neurons (in the Y area) allows the agent to approach the friend and avoid the enemy faster as compounding effects of sequential states. Further, we tested the effect of punishment and reward schedule with the same initial locations. We also experimented the effect of punishment and reward schedule with random initial locations. These experiments all indicated that the reinforcement learning via the serotonin and the dopamine systems is beneficial for developing desirable behaviors in this set of sequential tasks — staying close to its friend and away from its enemy. As far as we know, this is the first work that investigates the effects of reinforcer (via serotonin and dopamine) on internal neurons for sequential tasks.**

I. Introduction

THE modulatory system of the brain is often called motivational system or value system in neuroscience and psychology. It is about how neurons in the brain use a few particular types of neural transmitter to indicate certain properties of signals. In the subject of neural modulation, a small group of neurons synthesize and release a particular type of neural modulatory transmitters which diffuse through large areas of the nervous system, producing an effect on many neurons. Functionally, neural modulation is needed for non-associative learning (e.g., sensitization and habituation), classical conditioning, instrumental conditioning (also called reinforcement learning), motivation, emotion and homeostasis.

A modulatory system goes beyond information processing and sensorimotor behaviors. It provides mechanisms to developmental system so that it develops dislikes and likes [1]. Without a modulatory system, the brain does not have a sense of what is important to learn.

A. Neurotransmitter

A main function of the brain is to develop circuits for processing sensory information and generating corresponding motor actions. Such processing is characterized by direct synaptic transmission—the pre-synaptic neuron directly influences the post-synaptic neuron through different types of neurotransmitters. For example, glutamate is a kind of neurotransmitter. Nerve impulses trigger release of glutamate from the pre-synaptic cell [2]. Once released, the glutamate neurotransmitter binds a glutamate receptor, such as the NMDA receptor, in the post-synaptic neuron. A sufficient number of bindings by neurotransmitters results in the firing of the post-synaptic neuron [3]. Each cell typically has many receptors, of many different kinds.

While glutamate and GABA are neurotransmitters whose values are largely neutral in terms of reference at the time of birth, some other neurotransmitters appear to have been used by the brain to represent certain signals with intrinsic values [1], [3], [4]. For example, serotonin (5-HT) seems to be involved with pain, punishment, stress and threats; while dopamine (DA) appears to be related to pleasure, wanting, anticipation and reward [4].

Therefore, 5-HT and DA, along with many other neurotransmitters that have inherent values, seem to be useful for modeling the intrinsic value system of the central nervous system and artificial neural networks.

B. Psychological and biological studies

Psychological studies have provided rich behavioral evidence about the existence of the motivational system [1, 5-13]. It is known that the motivational system is important to the autonomous learning of the brain. However, although there is a very rich literature about models of neuromodulatory systems, such models are limited in terms of computational functions due to a few well known limitations in prior neural network models. Weng argued that they have overcome such limitations [14].

The motivational systems are often referred to as diffuse systems in the sense that each modulatory neuron in such a system uses a particular neurotransmitter (e.g., serotonin

Dongshu Wang is with School of Electrical Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China. Now he is a visiting scholar in Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48823, USA. Corresponding author. (email: wangdongshu@zzu.edu.cn.)

Yihai Duan is with School of Electrical Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China. (e-mail: duanyihai88@163.com)

Juyang Weng is with the Computer Science and Engineering, Michigan State University, East Lansing , MI 48823, USA. He is also with the MSU Cognitive Science Program and the MSU Neuroscience Program. (e-mail: weng@cse.msu.edu).

This work was supported by the Natural Sciences Foundation of China (Grant No. 61174085) and the Specialized Research Fund for the Doctoral Program of Higher Education of the Education Department of China (Grant No. 20114101110005).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

or dopamine) and makes diffuse connections, via such neurotransmitters, with many other neurons [1]. Instead of carrying out detailed sensorimotor information processing, these cells often perform regulatory functions, modulating many postsynaptic neurons (e.g., the cerebral cortex, and the thalamus) so that they become more or less excitable, or make fewer or more connections. The neurotransmitters of such modulatory systems arise from the central core of the brain, typically from the brain stem. The core of each system has a small number of neurons (e.g., a few thousand), but each such neuron contacts more than 100,000 postsynaptic neurons that are widely spread across the brain. The synapses of such modulatory neurons release a particular type of transmitter molecule (e.g., serotonin or dopamine) into the extracellular fluid. Thus, such molecules diffuse to many neurons, instead of being confined to a local neighborhood of a synaptic cleft.

The effects of dopamine and serotonin are not completely understood, according to the literature. Both serotonin and dopamine come in many different forms, since the brain may conveniently use a particular neurotransmitter for different purposes in different parts of the body. For example, some forms of dopamine are related to reward and some serotonin is related to aversion and punishment. There are other forms of serotonin and dopamine that have different effects that are not widely understood [7], [15]. We will focus on the forms of dopamine and serotonin that affect the brain as reward and punishment respectively.

Dopamine is released in the brain's Ventral Tegmental Area (VTA) and particularly the nucleas accumbens act as a general facilitative agonist for pleasure and is related to the reinforcement of behaviors [16]. Increased dopamine in these areas creates such feelings as euphoric sensations, added energy, and an increase in focus ability. Dopamine is associated with reward prediction [7]. If an agent gets a reward, then dopamine is released in the brain. If an agent is expecting a reward, dopamine is also released.

Serotonin often appears to be dopamines counterpart [17]. Dopamine excites the neurons while serotonin inhibits them. One specific type of serotonin with this effect is 5-HT. Serotonin leads to behavior inhibition and aversion to punishment [17]. For example, if a monkey pushes a lever and receives a shock, then it will avoid pressing that lever [7]. There are two parts of the brain that release serotonin, the dorsal raphe and the median raphe [17]. The dorsal raphe connects serotonin to all of the areas that have dopamine connections [18]. Serotonin from the dorsal raphe interacts with dopamine to cause the agent to avoid behaviors that the dopamine encourages.

In the following, wet discuss the theory for emergent value systems for sequential tasks.

II. Theory for Value System of Sequential Tasks

In terms of context dependence, there are two types of tasks, episodic and sequential.

A. Episodic and sequential tasks

In an episodic task environment, the agent's experience is divided into atomic episodes. Each episode consists of the agent perceiving and then performing a single action. Crucially, the next episode does not depend on the actions taken in previous episodes. In episodic environments, the choice of action in each episode depends only on the episode itself. Many classification tasks are episodic. For example, an agent that has to spot defective parts on an assembly line bases each decision on the current part, regardless of previous decisions; moreover, the current decision does not affect whether the next part is defective. In sequential environments, on the other hand, the current decision could affect all future decisions. Chess, taxi driving and robot navigation are sequential. In these cases, short-term actions can have long-term consequences [19]. Episodic environments are much simpler than sequential environments because the agent does not need to think ahead.

B. Reinforcement in sequential tasks

Reinforcement learning is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment. Without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make [19]. The agent needs to know that something good has happened and that something bad has happened. This kind of feedback is called a reward or punishment. In the natural world, signals from a pain sensor is associated with bad and those from a sweet sensor is associate with good. However, the brain mechanisms of such an association is largely unknown.

Reinforcement learning has been carefully studied by animal psychologists for many years [20-23]. Techniques used in traditional reinforcement learning include: Markov decision process, dynamic programming, Monte Carlo method, temporal difference learning, Q-learning, function approximation, etc. However, those models are *symbolic* in the sense that each mode has specific, handcrafted meanings for a specific task.

Using Brain-like *emergent* representations, each neuron does not have a specific task meaning. Patterns of neuronal firing emerge from interactions of the physical external world. In particular, a motor neuron (or a firing pattern of neurons) does not represent a bad action until it consistently fires with the presence of serotonin. Weng et al. [12] modeled that serotonin and dopamine are associated with pain sensors and sweet sensors, or punishment and reward, respectively, in general. Serotonin inhibits the firing of the current motor neurons and dopamine excites the firing of the current motor neurons. Hopefully, since the diffusions of such neural transmitters are relatively slow, statistically, the level of such neurotransmitters correlates with the responsible motor actions reasonably well.

The above is about motor neurons. In a sequential task, however, each reinforcer (punishment or reward) is a consequence of a sequence of past state trajectories indicated by

sensory-state pairs in terms of $(\mathbf{x}(t), \mathbf{z}(t))$:

$$(\mathbf{x}(t-m), \mathbf{z}(t-m)), ..., (\mathbf{x}(t-1), \mathbf{z}(t-1)), (\mathbf{x}(t), \mathbf{z}(t)) \quad (1)$$

where $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are the sensory and state vectors, respectively. In our Developmental Network (DN) framework, state and action are the same since reporting state is an action. E.g., say bad words is an action punishable.

The Y area, represented by a large number of Y neurons, represents a multi-exchange *bridge* that form feature clusters in the two islands X area and Z area (as discussed in the next section):

$$(\mathbf{x}(t-1), \mathbf{z}(t-1)) \rightarrow \mathbf{y} \rightarrow (\mathbf{x}(t), \mathbf{z}(t))$$

where $\rightarrow$ means "predicts". If the following $(\mathbf{x}(t), \mathbf{z}(t))$ is independent with y, the task is episodic. In a sequential task, the following $(\mathbf{x}(t), \mathbf{z}(t))$ depends on y.

In our prior work [26], we modeled that serotonin and dopamine both increase the learning rate of the firing Y neurons to memorize the important event in episodic tasks. In this paper, we study how serotonin and dopamine increase the learning rate of firing Y neurons for sequential tasks. As we can see from the above analysis, changing the rate using serotonin and dopamine transmitter should improve the performance of learning sequential tasks.

The rest of the paper is organized as follows. In section III, we review the theory behind our model. In section IV, we design the experiment, show and analyze the experiment results. In section V, we present the conclusion and the future study.

III. Network Architecture

A. Developmental network

Development network is the basis of a series of Where-What networks, whose 7th version, namely, the latest version, appeared in [24]. The simplest version of a Developmental Network (DN) has three areas, the sensory area X, the internal area Y, and the motor area Z, with an example in Fig. 1. The internal area Y as a "bridge" to connect its two "banks"— the sensory area X and the motor area Z.

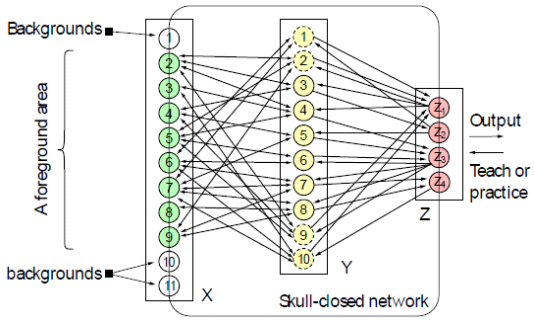


Fig. 1. The architecture of DA. It contains top-down connections from Z to Y for context represented by the motor area. It contains top-down connections from Y to X for sensory prediction (but this part is not used in the work here). Pink areas are human designed or human taught. Yellow areas are autonomously generated (emergent and developed).

The most basic function of an area Y seems to be prediction—predict the signals in its two vast banks X and Z through space and time. The prediction applies when part of a bank is not supervised. The prediction also makes its bank less noisy if the bank can generate its own signals (e.g., X)[14].

A secondary function of Y is to develop bias (like or dislike) to the signals in the two banks, through what is known in neuroscience as neuromodulatory systems.

The DN algorithm is as follows. Input areas: X and Z, Output areas: X and Z. The dimension and representation of X and Y areas are hand designed based on the sensors and effectors of the robotic agent or biologically regulated by the genome. Y is skull-closed inside the brain, not directly accessible by the external world after the birth.

1) At time $t = 0$, for each area A in $\{X, Y, Z\}$, initialize its adaptive part $N = (V, G)$ and the response vector $\mathbf{r}$, where V contains all the synaptic weight vectors and G stores all the neuronal ages.

2) At time $t = 1, 2, ...$, for each area A in $\{X, Y, Z\}$, do the following two steps repeatedly forever:

a) Every area A computes using area function f.

$$(\mathbf{r}', \ N') = f(\mathbf{b}, \mathbf{t}, N) \quad (2)$$

where f is the unified area function described in the following equation (3), $\mathbf{b}$ and $\mathbf{t}$ are areas bottom-up and top-down inputs from current network response $\mathbf{r}$, respectively; and $\mathbf{r}'$ is its new response vector.

b) For each area A in $\{X, Y, Z\}$, A replaces: $N \leftarrow N'$ and $\mathbf{r} \leftarrow \mathbf{r}'$.

If X is a sensory area, $x \in X$ is always supervised and then it does not need any synaptic vector. The $z \in Z$ is supervised only when the teacher chooses to. Otherwise, z gives (predicts) motor output. Next, we describe the area function f.

Each neuron in area A has a weight vector $\mathbf{v} = (\mathbf{v}_b, \mathbf{v}_t)$, corresponding to the area input $(\mathbf{b}, \mathbf{t})$, if both bottom-up part and top-down part are applicable to the area. Otherwise, the missing part of the two should be dropped from the notation. Its pre-action energy is the sum of two normalized inner product:

$$r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \frac{\mathbf{v}_b}{||\mathbf{v}_b||} \cdot \frac{\mathbf{b}}{||\mathbf{b}||} + \frac{\mathbf{v}_t}{||\mathbf{v}_t||} \cdot \frac{\mathbf{t}}{||\mathbf{t}||} = \dot{\mathbf{v}} \cdot \dot{\mathbf{p}} \quad (3)$$

where $\dot{\mathbf{v}}$ is the unit vector of the normalized synaptic vector $\mathbf{v} = (\dot{\mathbf{v}}_b, \dot{\mathbf{v}}_t)$, and $\dot{\mathbf{p}}$ is the unit vector of the normalized synaptic vector $\mathbf{p} = (\dot{\mathbf{b}}, \dot{\mathbf{t}})$. The inner product measures the degree of match between these two directions $\dot{\mathbf{v}}$ and $\dot{\mathbf{p}}$, because $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t}) = \cos(\theta)$ where θ is the angle between two unit vectors $\dot{\mathbf{v}}$ and $\dot{\mathbf{p}}$. This enables a match between two vectors of different magnitudes. The pre-action energy value ranges in [-1, 1].

To simulate lateral inhibition (winner takes all) within each area A, only top-k winners fire and update. Considering $k = 1$, the winner neuron j is identified by:

$$j = \arg \max_{1 \le i \le c} r(\mathbf{v}_{bi}, \mathbf{b}, \mathbf{v}_{ti}, \mathbf{t}) \quad (4)$$

where c is the neuron number in the area A.

The area dynamically scale top-k winners so that the top-k responses with values in [0,1]. For $k = 1$, only the single winner fires with responses value $y_j = 1$ and all other neurons in A do not fire. The response value y_j approximates the probability for $\dot{\mathbf{p}}$ to fall into the Voronoi region of its $\dot{\mathbf{v}}_j$ where the "nearness" is $r(\mathbf{v}_b, \mathbf{b}, \mathbf{v}_t, \mathbf{t})$.

All the connections in a DN are learned incrementally based on Hebbian learning—co-firing of the pre-synaptic activity $\dot{\mathbf{p}}$ and the post-synaptic activity y of the firing neuron. Consider area Y, as other area learn in a similar way. If the pre-synaptic end and the post-synaptic end fire together, the synaptic vector of the neuron has a synapse gain $y\dot{\mathbf{p}}$. Other non-firing neurons do not modify their memory. When a neuron j fires, its weight is updated by a Hebbian-like mechanism:

$$\mathbf{v}_j \leftarrow \omega_1(n_j)\mathbf{v}_j + \omega_2(n_j)y_j\dot{\mathbf{p}} \tag{5}$$

where $\omega_2(n_j)$ is the learning rate depending on the firing age n_j of the neuron j and $\omega_1(n_j)$ is the retention rate with $\omega_1(n_j) + \omega_2(n_j) \equiv 1$. The simplest version of $\omega_2(n_j)$ is $1/n_j$, which gives the recursive computation of the sample mean of input $\dot{\mathbf{p}}$:

$$\mathbf{v}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\dot{\mathbf{p}}(t_i) \tag{6}$$

where t_i is the firing time of the neuron. The age of the winner neuron j is incremented $n_j \leftarrow n_j+1$. A component in the gain vector $y_j\dot{\mathbf{p}}$ is zero if the corresponding component in $\dot{\mathbf{p}}$ is zero. Each component in $\mathbf{v}_j$ so incrementally computed is the estimated probability for the pre-synaptic neuron to fire under the condition that the post-synaptic neuron fires. A more complicated version of $\omega_2(n_j)$ is presented in the next section when we discuss the architecture of our motivated system.

B. Motivated Developmental Network (MDN)

According to literature[17], serotonin and dopamine receptors are also found in brain neurons except the motor neurons. It means that the release of serotonin and dopamine, which occurs in RN and VTA areas, should also have effect on neurons in Y_u neurons. Previous researches [1], [25], namely, the original MDN, modeled the effect of serotonin and dopamine on motor areas, but they did not consider the effect of these neurotransmitters on Y_u neurons.

Fig. 2 presents the architecture of the motivated DN. It links all pain receptors with raphe nuclei (RN) located in the brain stem—represented as an area, which has the same number of neurons as the number of pain sensors. Every neuron in RN releases serotonin. Similarly, it also links all sweet receptor with VTA—represented as an area, which has the same number of neurons as the number of sweet sensors. Every neuron in VTA releases dopamine.

Serotonin and dopamine are synthesized by several brain areas. For simplicity, we use only RN to denote the area that synthesizes serotonin and only the VTA to denote the area

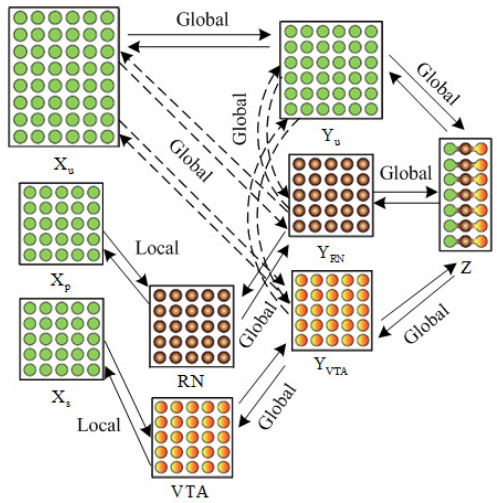


Fig. 2. A motivated DN with serotonin and dopamine modulatory subsystems. It has 9 areas. RN has serotonergic neurons. Neurons in Y_{RN} and Z have serotoninergic synapses. VTA has dopaminergic neurons. Neurons in Y_{VTA} and Z have dopaminergic synapses. The areas Y_u, Y_{RN} and Y_{VTA} should reside in the same cortical areas, each represented by a different type of neurons, with different neuronal densities. Within each Y sub-area and the Z area, within area connections are simulated by top-k competition.

that synthesizes dopamine, although other areas in the brain also involved in the synthesis of these neurotransmitters.

Therefore the sensory area $X = (X_u, X_p, X_s)$ consisting of an unbiased array X_u, a pain array X_p and a sweet array X_s. $Y = (Y_u, Y_{\text{RN}}, Y_{\text{VTA}})$ connects with $X = (X_u, X_p, X_s)$, RN and VTA as bottom-up inputs and Z as top-down input.

Within such a motivated developmental network, the motor area is denoted as a sequence of neurons $Z = (z_1, z_2, \cdots, z_m)$, where m is the number of motor neurons whose axons innervate muscles or glands. Each z_i has three neurons $z_i = (z_{iu}, z_{ip}, z_{is})$, where z_{iu}, z_{ip} and z_{is} $(i = 1, 2, \cdots, m)$ are unbiased, pain and sweet, respectively. And these indicate the effects of glutamatergic synapses, serotoninergic synapses and dopaminergic synapses, respectively.

Whether the action i is released depends on not only the response of z_{iu} but also on those of z_{ip} and z_{is}. z_{ip} and z_{is} report how much negative value and positive value are associated with the i-th action, according to past experience. They form a triplet for the pre-action energy value of each motor neuron, glutamate, serotonin and dopamine.

Modeling the cell's internal interactions of the three different types of neurotransmitter, the composite pre-action value of a motor neuron is determined by

$$z_i = z_{iu}\gamma(1 - \alpha z_{ip} + \beta z_{is}) \tag{7}$$

with positive constants α, β and γ. In other words, z_{ip} inhibits the action but z_{is} excites it. α is a relatively larger constant than β since punishment typically produces a change in behavior much more significantly and rapidly than other forms of reinforcers.

Then the j-th motor neuron fires and action is released

where

$$j = \arg \max_{1 \le i \le m} \{z_i\} \tag{8}$$

That is, the primitive action released at this time frame is the one that has the highest value after inhibitory modulation through serotonin and excitatory modulation through dopamine, respectively, by its bottom-up synaptic weights. Other Z neurons do not fire.

C. Novelty

Our improvement on the motivated development network lies in the following three aspects:

Firstly, though Zheng [26] also considered the effect of serotonin and dopamine on the neurons in Y_u, their experiment object is face recognition which is a pattern recognition problem. The recognition result is determined only by the current decision. But the robot navigation in unknown environment is a typical sequential problem because the current decision can affect the following one, and the final result is determined by all the former decisions instead of the last one. It is very important and essential to study the effect of serotonin and dopamine on the sequential behavior.

Secondly, previous work [2] studied the effect of serotonin and dopamine on the learning rate of Z qualitatively. Our research studied their effect on the learning rate of Z and Y_u quantitatively. In other words, we calculate the linear punishment (or reward) value according to the distance between the agent and the friend (or enemy) and the threshold, instead of the qualitative way. Quantitative method can approach the real case in physical world better and explain the effect of serotonin and dopamine on the navigation performance more directly.

Finally, we studied the effects of different environments (such as different teachers) on the final navigation performance. To the strict "teacher", even though you do very well, he still consider you have not done enough. On the contrary, to the tolerant "teacher", though you do not do well, he may think you have done quite well. Different environments will produce different effects on the same sequential task. So it is very necessary to study the effects of different environments on sequential tasks.

IV. EXPERIMENT

This section, we will describe the experiment procedure designed to test the above theory and algorithm.

A. Experiment design

In our experiment, we use three robots to test our algorithm. One of the robots is the agent which can think and act, the other two are its friend and enemy, respectively. If the agent approaches the friend robot, it is rewarded with dopamine. If it approaches the enemy, it is punished with serotonin. In this way, the agent will learn to close its friend and avoid its enemy. But it must learn this behavior through its own trial and error experience.

The agent's brain is the motivated DN with three areas, X, Y and Z, as depicted in Fig. 2. Where X is the sensor area which has three sub-areas, X_u is the unbiased area, X_p is the pain area and X_s is the sweet area. At each time step, each area produces a response vector based on the physical state of the world. The X_u vector is created directly from the sensors' input. The X_p vector identifies in which ways the robot is punished and represents the release of the serotonin in RN. The X_s vector identifies in which ways the robot is rewarded and represents the release of the dopamine in VTA.

All of the Y_{RN}, Y_{VTA} and Z sub-areas compute their response vectors in the same way. At the end of each time step, the neurons in Y_{RN}, Y_{VTA} and Z areas that fired update themselves. Their weights are updated according to the former equations (5) and (6). Their ages are updated as follows: $a_i \leftarrow a_i + 1$. The Z area recombines the results from its collaterals to compute a single response vector.

According to the work of [24], serotonin and dopamine levels are released at different levels rather than binary values. The release gives specific neurons in the Y_{RN} and Y_{VTA} areas a non-zero response, which, in our version of modulated developmental network, will have effect on the learning rate of neurons in Y_u. Moreover, the roles of a motor neuron and an inter neuron are very different. The former roughly corresponds to the action that is responsible for the corresponding punishment and reward; the latter corresponds to the memory of the corresponding event. Therefore, serotonin and dopamine should increase the efficiency of learning in Y_u, instead of directly discouraging and encouraging the firing. One way to reach such an effect is to increase the learning rate depicted as follows:

$$\omega_2(n_j) = \min((1 + \alpha_{\text{RN}} + \alpha_{\text{VTA}})\frac{1}{n_j}, 1) \tag{9}$$

where α_{RN} and α_{VTA} are the constants related with RN and VTA respectively. This expression shows that reward and punishment change the learning rate in Y_u neurons. If neurons in X_p and X_s do not fire, responses in RN and VTA are zero. Thus the learning rate (9) will turn into its original form $1/n_j$.

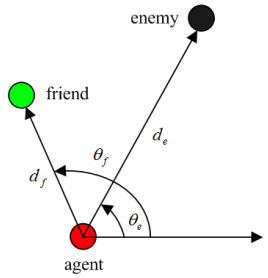


Fig. 3. The setting of the wandering plan which includes the agent, the friend robot and enemy robot. The size of the square space used is 500×500.

B. Input and Output

The Y and Z_u areas are initialized to contain small random data in their state vectors. The Z_p and Z_s areas are initialized to zero since the agent has no idea which actions will cause pleasure or pain. The ages of all neurons are initialized to

1. The number of neurons in Y layer and the fire number k can be selected based on the resources available. The size of Z area is equal to the number of actions that the agent can perform. At any time, the agent can perform one of nine possible actions, it can move in each of the cardinal or intercardinal directions or it can maintain its current position. So the neurons in Z areas has 9 rows and 3 columns. 3 columns denote the Z_u, Z_p and Z_s respectively.

The size of each vector in the X area is determined by the transformation function through which the robot can sense the locations of its friend and enemy. If we define the following entities, a (agent), f (friend), e (enemy) , we can draw a sketch of the location relation among the three robots as figure 3, and get the following expressions:

$$\begin{aligned}
\theta_f =& \arctan(a_x - f_x, a_y - f_y),\\
d_f =& \sqrt{(a_x - f_x)^2 + (a_y - f_y)^2},\\
\theta_e =& \arctan(a_x - e_x, a_y - e_y),\\
d_e =& \sqrt{(a_x - e_x)^2 + (a_y - e_y)^2},\\
x_u =& \{\cos\theta_f, \sin\theta_f, \cos\theta_e, \sin\theta_e, \frac{d_f}{d_f + d_e}, \frac{d_e}{d_f + d_e}\},
\end{aligned}$$

where θ_f and θ_e are the angle between the heading of the agent and the direction of the friend robot and enemy robot, respectively; d_f and d_e are the distance between the agent and the friend and the enemy, respectively.

The pain sensor and the sweet sensor has just one value to denote the fear and desire. The fear threshold is set 125, namely, if $d_e > 125$, there is no punishment. If $30 < d_e \leq 125$, punishment value is set 4. Otherwise, the punishment value is calculated through the fear threshold divided by the actual distance d_e.

Similarly, the desire threshold is set 50, namely, if $d_f < 50$, there is no reward. If $50 < d_f \leq 150$, the reward value is calculated through the actual distance d_f divided by the desire threshold. Otherwise, the reward value is set 3.

C. Experiment setup

We designed a simulation environment to illustrate how such a motivated agent would response in the presence of the friend robot and enemy robot. The motivated robot (agent) is controlled by the motivated "brain" which is actually a motivated development network. The "brain" releases serotonin and dopamine for the enemy and friend based on the specific circumstances. Through the simulation, the agent will learn by reinforcement, deciding which one to avoid and which one to go after, based on the release of the serotonin or dopamine.

At each time step, the horizontal and vertical coordinates are collected for each entity. With these data, we can calculate the distance between the agent and its friend and enemy. Through observing the distances of the agent to its friend and enemy, we can measure the learning procedure of the agent.

The agent starts with a behavior pattern determined by its initial neural network configuration. The unbiased regions are initialized with small random data while the biased regions are initialized to zero. This gives the initial appearance of a random behavior pattern. Eventually, it performs an action that causes it to be rewarded or punished, causing it to be either favor or avoid that action when placed in similar situations in the future.

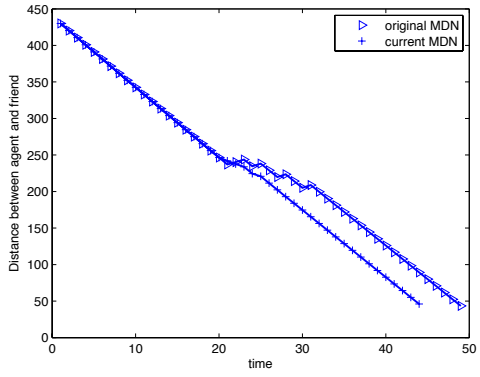
D. Results and analysis

In order to test the effect of serotonin and dopamine on the algorithm performance of current MDN, we compare the distance between the agent and its friend (or enemy) under the original MDN (based on Daly's work, reference 2, it only considered the effects of serotonin and dopamine on the motor area qualitatively) and the current MDN (we designed, it not only considered the effects of serotonin and dopamine on the motor area, but also considered their effects on the Y_u area quantitatively) shown in Fig. 4 (a), (b) respectively. From the figure 4, we can see that at the initial time steps, the agent does not know that the friend can cause pleasure and the enemy can cause pain. So it moves randomly and does not feel pain and pleasure. But when the agent moves towards the friend randomly, it realizes that it can receive pleasure as seen between time step 20 and 25. Similarly, when the agent moves towards the enemy, it realizes that it receives pain. So from then on, the agent will stay close to its friend while avoiding the enemy. This indicates that the reinforcement learning is useful in teaching the agent to stay close its friend and away from its enemy. The agent is able to figure out how to react in a given situation rather than having to be explicitly taught where to go.

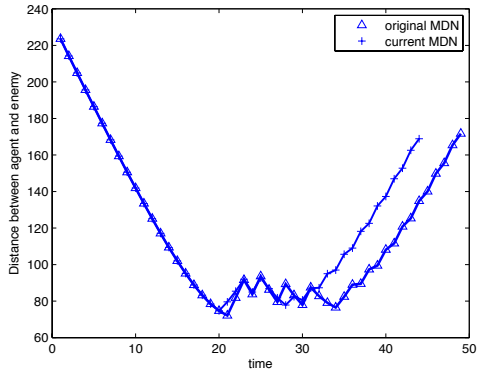
Except this, from Fig. 4, we can also see that under the same condition, the current MDN can get a smaller d_f and bigger d_e than the original MDN which demonstrates the effects of the serotonin and dopamine on the learning of Y_u. Learning rate of Y_u in current MDN introduces the effect of serotonin and dopamine, but the learning rate of Y_u in original MDN is its initial expression $\omega_2(n_j)$ =1/n_j. So the learning rate of current MDN is bigger than the original MDN, its learning speed is faster than that of the original MDN. So the current MDN can approach the friend or avoid the enemy faster than the original MDN.

In the second part, we study the effect of same environment and different punishment and/or reward values on the robot navigation. Initially, the three robots are set to constant locations in different cases. In other words, their relative locations are fixed and the distances between the agent and the friend (or enemy) are invariable. But the punishment and/or reward values are variable. Then we analyze the effects of these different punishment and/or reward values on the robot behavior improvement. The effects are shown in Fig. 5. From Fig. 5, we can see:

(1) Very small punishment. In this situation, punishment is set 1 and reward is set according to section B. From Fig. 5 (a), we can see that the nearest distance between the agent and the enemy is 0. At the former one time step, the agent is far from its friend. The agent receives the punishment and reward at the same time, but it moves through the enemy instead of moving far away from it and moves towards the

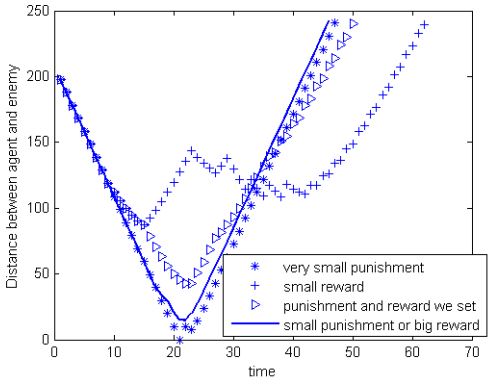


(a)

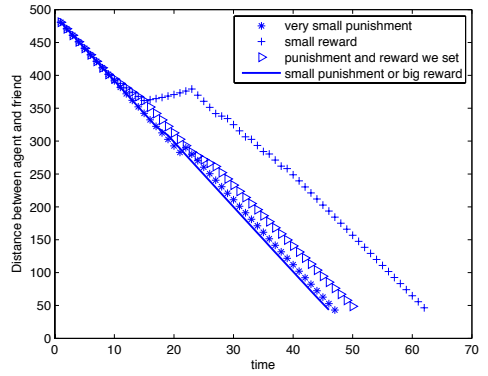


(b)

Fig. 4. Distance comparison between agent and friend (a) and enemy (b) under original MDN and current MDN.



(a)



(b)

Fig. 5. Distance comparisons among different punishments and/or rewards.

friend. So we can conclude that when the agent, friend and enemy are located in the same line, and if the reward value is bigger than the punishment value, the punishment can be omitted. This phenomenon also means that the reward is set too big and the punishment is relatively too small. This suggests that minimum punishment should be bigger than the maximum reward. This is one of the requirements in designing punishment and reward value. Their whole depictions defined in section B are shown in Fig. 6.

(2) Small reward. In this situation, reward is set 0.3. At time step 20 and the following time steps, we can see that the agent does not receive punishment (because the distance between the agent and the enemy is bigger than the fear threshold 125) and only receive the reward. But the distance d_f increases instead of decreasing. This phenomenon shows that when the agent is far away from the enemy, the effect of weights is bigger than that of the reward. It also means that the reward is set relatively small.

(3) Punishment and reward value we set. From Fig. 5 (a) and (b), we can see that in this situation, the agent can not only keep certain distance from the enemy, but also moves towards the friend in short time. These means that the punishment and reward values we set are suitable.

(4) Small punishment or big reward. In this situation, the maximum reward value is set 5 or bigger, punishment value is set according to section B, or the reward keeps invariable and the maximum punishment is set 2. The effects of small punishment on algorithm performance are between that of (1) and (3). The effects of big reward is similar to that of the small punishment. In this situation, the agent can also keep a certain distance from the enemy, but the distance is relatively small, compared with that of the (3). It can also reach the friend in short time.

In the final part, we test the effects of different environmental parameters (different punishment and reward thresholds) on the sequential task quantitatively. The experiment results are illustrated in figure 7. six groups of numbers along the horizontal axis denote the different distances between the agent and the friend (or enemy), and these distances are adopted as the reward (or punishment) thresholds in the corresponding experiments, and the vertical axis denotes the average distances we measured. If the distance between the agent and the friend is bigger than the corresponding reward threshold, the agent will receive a reward from its friend which will attract the agent to move towards the friend. Otherwise, the agent will not receive the reward. Similarly, if the distance between the agent and the enemy is smaller than the corresponding punishment threshold, the agent will receive a punishment from its enemy which will force the agent to move away from the enemy. Otherwise, the agent

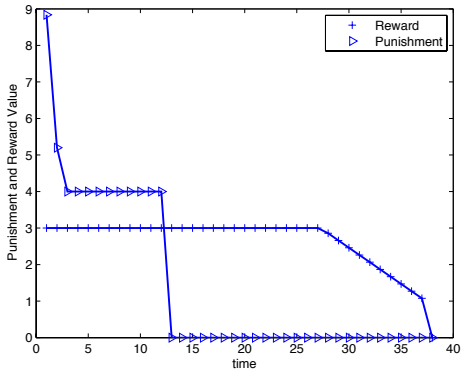


Fig. 6. Punishment and reward change with time.

will not receive the punishment.

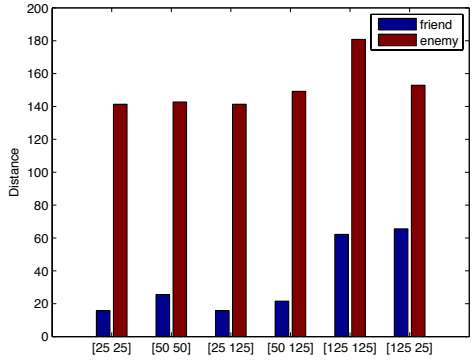


Fig. 7. Average distance between agent and its friend and enemy in different situations.

V. Conclusions

We analyzed the effect of serotonin and dopamine systems on Y neurons for sequential tasks. We conducted three simulation experiments to test the effects of serotonin and dopamine on robot navigation performance. In future work, the agent will be placed in a more complicated situation with multiple friends and enemies to further study the effects of serotonin and dopamine.

References

[1] J. Weng, *Natural and artificial intelligence-introduction to computational brain-mind.* BMI press, Okemos, Michigan, 2012.

[2] J. Daly, J. Brown and J. Weng, "Neuromorphic motivated systems," *Proceedings of international joint conference on neural networks*, San Jose, California, USA, July 31-August 5, 2011, pp. 2917-2924.

[3] B. R. Cox, J. L. Krichmar,"Neuromodulation as a Robot Controller, A Brain-Inspired Strategy for Controlling Autonomous Robots," *IEEE Robotics & Automation Magazine* No. 9, pp. 72-80, September, 2009.

[4] M. J. Crockett, A. Apergis-Schoute, B. Herrmann, M. Lieberman, U. Muller, T. W. Robbins, and L. Clark, "Serotonin modulates striatal responses to fairness and retaliation in humans," *The journal of Neuroscience*, vol. 33, No. 8, pp. 3505-3513, 2013.

[5] G. Baldassarre, F. Mannella, V. G. Fiore, P. Redgrave, K. Gurney, M. Mirolli, "Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model," *Neural Networks*, vol. 41, pp. 168-187, 2013.

[6] Z. Ji , M. D. Luciw, J. Weng, "A Biologically-Motivated Developmental System Towards Perceptual Awareness in Vehicle-Based Robots," *Proceedings of the Seventh International Conference on Epigenetic Robotics: Modeling, Cognitive Development in Robotic Systems*, Lund University Cognitive Studies.2007, pp. 1335-1342.

[7] S. Kakade, P. Dayan, "Dopamine: generalization and bonuses," *Neural network*, no. 15, pp. 549-559, 2002.

[8] M. Kubisch, M. Hild, S. Hofer, "Proposal of an Intrinsically Motivated System for Exploration of Sensorimotor State Spaces," *Proceedings of the Tenth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund University Cognitive Studies, 2010, pp. 51-56.

[9] S. Cavallaro, "Genomic analysis of serotonin receptors in learning and memory," *Behavioral Brain Research*, vol. 195, no. 1, pp. 2-6, 2008.

[10] S. Niekum, A. G. Barto, L. Spector, "Genetic Programming for Reward Function Search," *IEEE transactions on autonomous mental development*, vol. 2, no. 2, pp. 83-90, 2010.

[11] P. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic Motivation Systems for Autonomous Mental Development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265-286, 2007.

[12] J. Weng, S. Paslaski, J. Daly, C. VanDam, and J. Brown, "Modulation for emergent networks: Serotonin and dopamine," *Neural Networks*, no. 41, pp. 225-239, 2013.

[13] H. Wersing, S. Kirstein, M. Gotting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick1, J. Steil, H. Ritter, E. Korner, "A biologically motivated system for unconstrained online learning of visual objects," *Artificial Neural networks -ICANN 2006, Lecture Notes in Computer Science*,vol. 4132, pp. 508-517, 2006.

[14] J. Weng, "Why have we passed "neural networks no not abstract well"," *Natural intelligence: the INNS magizine,*, vol. 1, no. 1, pp.13-22, 2011.

[15] K. E. Merrick, "A Comparative Study of Value Systems for Self-Motivated Exploration and Learning by Robots," *IEEE transactions on autonomous mental development*, vol. 2, no. 2, pp.119-131, 2010.

[16] K. Nelson, "Motivation and the Brain: How incentives affect the brain and motivation," *Available: www.google.com*.

[17] N. D. Daw, S. Kakada, and P. Dayan, "Enemy interactions between serotonin and dopamine," *Neural networks*, vol. 15, no.4-6, pp. 603-616, 2002.

[18] J. L. Krichmar, "The neuromodulatory system: a framework for survival and adaptive behavior in a challenging world," *Adaptive behavior*, vol. 16, no. 6, pp. 385-399, 2008.

[19] S. J. Russel, P. Norvig, *Artificial Intelligence: A Modern Approach. Third Edition.* Upper Saddle River, New Jersey, Prentice hall, 2010.

[20] P. Dayan, B. W. Balleine, "Reward, Motivation and reinforcement learning," *Neuron*, vol. 36, no. 10, pp. 285-298, 2002.

[21] X. Huang, J. Weng, "Inherent value systems for autonomous mental development," *International Journal of humanoid Robotics*, vol. 4, no. 2, pp. 407-433, 2007.

[22] X. Huang, J. Weng, "Novelty and reinforcement learning in the value system of developmental robots," *Proceedings of the Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Edinburgh, Scotland, August 10 - 11*, 2002, pp. 1-9.

[23] S. Singh, R. L. Lewis, A. G. Barto, J. Sorg, "Intrinsically motivated Reinforcement Learning: An Evolutionary Perspective," *IEEE transactions on autonomous mental development*, vol. 2, no. 2, pp. 70-82, 2010.

[24] X. Wu, G. Qian and J. Weng,"Skull-closed autonomous development: WWN-7 dealing with scales," *.Proceedings of the international conference on brain-mind*, July 27 - 28, 2013, East Lansing, Michigan, pp. 1-9.

[25] S. Palaski, C. Vandam and J. Weng, "Modeling dopamine and serotonin systems in a visual recognition network," *Proceedings of international joint conference on neural networks*, San Jose, California, USA, July 31-August 5, 2011, pp. 3016-3023.

[26] Z. Zheng, K. Qian, J. Weng, Z. Zhang, "Modeling the effects of neuromodulation on internal brain areas: serotonin and dopamine," *Proceedings of international joint conference on neural networks*, Dallas, Texas, USA, August 4-9, 2013, pp. 1404-1411.