A BIOLOGICALLY PLAUSIBLE ACTOR/CRITIC MODEL

Robbie Veldkamp
Cognitive Artificial Intelligence
August 2007

Advisors:
M. Wiering
M. Meeter
H. Kunst

Contents

List of Figures

Chapter 1

Introduction

Humans and animals deal with highly dynamic environments. Food sources and predators change location and appearance all the time and areas that were safe once can all of a sudden grow extremely hazardous. Without the ability to adapt to such changes, most creatures will not survive very long.

Many adaptations are made on a basis of trial and error. Animals will experiment with problems by exploring different kinds of behavior to see which one works best. This trial and error learning is largely dependent on environmental feedback and this leads us to an interesting question. How does environmental feedback make its way into our brain?

1.1 Past Research

In the early eighties researchers found neurological evidence for a *prediction error signal* in the brain. Fluctuations in concentrations of *dopamine*, a neurotransmitter known to have both positive and negative effects on neuron excitability, appeared to reflect errors made by animals when predicting incoming rewards[57].

After this discovery, it did not take long before suggestions were made about the possibility of a prediction error based learning mechanism[57][60][6]. The proposed learning mechanisms showed many similarities to a set of methods known from the field of *reinforcement learning*. Therefore, a number of computational models were built for further investigation. Some of these models focused on the learning itself and did not make any attempt to give more insight into the structure and workings of the brain[54]. Others however, did try to give such insight, but usually restricted themselves to part of the learning problem. Some tried to learn a dopamine-like error signal[11][6] while others used a handcrafted error signal to learn action associations[30][9][34].

1.2 Goal of the Current Study

The current study honors neurobiological knowledge about the interactions of different brain areas and makes no attempt to improve learning algorithms. Instead, the study attempts to create a computational model that learns to produce an informative dopamine-like error signal and is able to use this signal to learn action associations. In accordance with a reinforcement learning technique called *actor/critic learning*, the model assigns these two goals to two different structures. The first structure, the *critic*, consists of brain areas contributing to the production and release of dopamine. The second structure, the *actor*, consists of brain areas responsible for action selection.

The study tries to answer the following central question: can a computer model be constructed, honoring neurobiological knowledge about relevant brain areas and their interconnections, that learns to produce a dopamine-like error signal and is able to use this signal for solving learning problems?

1.3 Relevance to Cognitive Artificial Intelligence (CAI)

CAI aims to be an interdisciplinary science in that it tries to connect the fields of linguistics, cognitive psychology and computer science. Even though the current study is not directly relevant to the field of linguistics, it does contribute to the other two sciences in varying degrees.

It could be argued that the present study is more relevant to cognitive science than it is to the field of artificial intelligence. The study makes no attempt to improve computer learning methods but rather tries to clarify the functional role and interdependencies of some neurobiological systems. Therefore, all the used computational procedures must be consistent with the physical capacities of biological nervous systems. This restriction is typical for cognitive studies and usually of no importance to the field of artificial intelligence.

This unbalance in relevance should be of no real consequence to the CAI relevance of the study. In an interdisciplinary science, study relevance can only be expected to show some unbalances between the different disciplines involved. In this case, the study has some explanatory, question asking and maybe even predictive power for cognitive scientists. For researchers in the field of artificial intelligence, the study might at best reveal some inspiring ideas.

1.3.1 Cognitive Psychology

In general, the creation of computer models, like the one treated here, forces cognitive psychologists to be very specific about their ideas. Computers simply do not understand vague notions and easily overlooked theoretical flaws will undoubtedly cause problems when trying to get a computational model to work. Such flaws will often lead to more detailed scientific questions and possibly even to the destruction of entire theories.

The current study might well produce such results. Even though there are neurobiological theories about the way the brain produces dopamine error signals and other theories about the way the brain learns action associations using such error signals, there does not seem to be a model connecting the two. Trying to construct such a model may lead to inconsistencies on a number of levels. For instance, different theories might prove to assign inconsistent functionalities to relevant brain areas or interactions between error signal learning and action learning might prove to be unworkable. Apart from revealing problems, the model might give some insight into the interplay between *learning to predict* and *learning to act*.

1.3.2 Computer Science

Without a doubt the brain is the best learning device known to man. In the past, many ideas were taken from the brain and used for improving computer learning methods. In the future, neurological findings might well continue to inspire computer scientists to come up with improvements in artificial learning.

In the current study, the connection between cognitive and computer science is interesting. In previous studies, ideas from computer science were used to construct learning theories about the brain. Now, ideas from cognitive psychology are used to construct a computer model of these theories. With some luck, the model inspires computer scientists to explore some new computational ideas while giving cognitive scientists some insight into their own theories. Clearly, the two disciplines are influencing each other.

1.4 Thesis Outline

The next two chapters of the thesis are introductory. Chapter two provides a treatment of actor/critic models from the perspective of reinforcement learning while chapter three treats biological knowledge about dopamine behavior and action selection from the perspective of cognitive psychology. Chapter four shows how this knowledge was integrated into a computational learning model and chapter five describes the results and behaviors of this model. Finally, chapter six answers the central question and holds some concluding remarks.

Chapter 2

Actor-Critic Models

Actor-critic methods[5] belong to the field of *reinforcement learning*[65]. More specifically, they are members of the set of *temporal-difference* methods (TD)[59][48][65] that were designed to solve prediction problems. Prediction problems are about finding the values of situations. They are of the following form: Given the current situation, what benefits can I expect to earn in the future?

The ability to solve prediction problems is very important in everyday life. Knowledge about the value of situations allows us to make informed decisions about what action to take next. The acquisition of this knowledge is based on exploiting past experiences. As will be described shortly, TD learning methods have some advantages over more traditional learning methods when it comes to learning from experience.

In the next section a brief introduction to TD methods will be provided. It is followed by an overview of actor-critic methods as special cases. Other TD methods will not be treated and few words will be spent on reinforcement learning in general as these topics are well out of the scope of this text. For further information the reader is directed to the excellent introductory book by Sutton and Barto[65].

2.1 Temporal Difference Methods

The main idea behind TD methods has been around for some time[1]. However, this idea was not generally understood and it took well into the 1980's for TD methods to become popular. Although by that time many other learning methods had been developed, TD methods proved to have some important advantages. This section explains the main idea behind TD methods and compares it to that of more traditional methods. An example application is used to provide better understanding and will be described first.

[1] In 1959 Samuel used a TD method to create a checkers player [59]

2.1.1 The Cart-Pole Balancing Task

A famous early reinforcement learning problem has been the *cart-pole balancing task*. The objective is to apply forces to a cart such that a pole is kept from falling over. The cart is stuck on a track and can only be moved left or right. If the pole falls past a certain angle, a failure occurs. Also, to keep the system from learning to push the cart left or right indefinitely, track borders are set up. If touched by the cart, these borders cause a failure. A prediction learning system can be taught to perform the task by giving it a reward signal r of $+1$ for every time step t it can function without failure. If the system is able to balance the pole for a longer time, the accumulation of rewards will be larger at the end of the episode[2]. Such an accumulation of reward is called a *return* and is defined as follows:

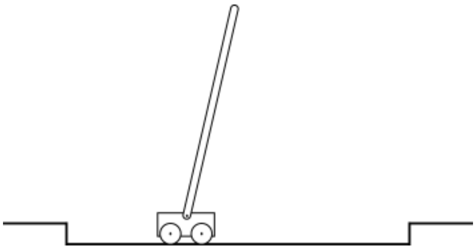

Figure 2.1: The pole-balancing task

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_T, \tag{2.1}$$

where R_t is the return at time step t and T is a final time step. This definition of return works well in tasks that have clear end states. For tasks of a more continues nature, the return definition is usually a bit different:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.2}$$

The γ parameter, $0 \leq \gamma \leq 1$, is called the *discount rate*. The discount rate determines the present value of future rewards since rewards earned further into the future have an increasingly smaller impact on $R(t)$. For $\gamma = 0$, the system is only interested in the reward at time $t + 1$. For $\gamma < 1$, the infinite sum has a finite value as long as the reward sequence $\{r_k\}$ is bounded. For γ approaching 1, the system will take rewards into account that are increasingly further into the future and therefore becomes more farsighted.

Based on return values, the system can make updates to the values of the visited states[3].

[2] When tasks have a logical end point, like *failure* in this example, one usually talks about an *episode* to denote the sequence of states between the start and end of a task

[3] In our example, a state could be defined in terms of pole angle, cart position, cart velocity, etc

But what does the value of a state express? A state value expresses the reward the system can expect to accumulate in the future starting from that particular state. More formally, the state value $V(s_t)$ is the *expected return* following state s. The optimal value function $V(s)$ will map states to value estimates and is defined as follows.

$$V(s) = E\left\{R_t \mid s_t = s\right\}, \tag{2.3}$$

After finding good value estimates for each possible state, the cart can balance the pole by choosing a path through state space that maximizes the return. And so, learning comes down to finding good approximations of the expected return values. The major difference between most prediction learning methods and TD methods is in the way return values are used to estimate state values.

2.1.2 Typical Use of Return Value

Finding good state value estimates can be done in several ways. In our example, most prediction learning methods will update the values of visited states *after* an episode has finished. At this point the return value of the episode is known and can be used to update state values. For instance, the famous *Monte Carlo* method will use the following update rule:

$$V(s_t) \leftarrow V(s_t) + \alpha(R_t - V(s_t), \tag{2.4}$$

where $V(s_t)$ is the value of state s at time t and α is a step size parameter. Clearly, this rule can only be applied when the value of R_t has been established. From the definition of R_t we know that the value can be calculated as soon as the episode has ended. And so, updating the state values has to wait until an episode has finished.

This use of return value is very intuitive but also somewhat unnatural. If we would have to learn to perform the tasks ourselves, learning would certainly not be postponed until the end of an episode. In fact, we would probably feel most of the knowledge is acquired while we are still playing! Unnatural or not, the algorithm was proven to converge to the expected return of any given state[63][24][62][67].

2.1.3 TD Methods

TD methods update state values a bit differently. Such methods change the value of a state as soon as the next state has occurred. The state value will be altered based on the difference between itself and the value of the next state. The update rule is shown in equation 2.5.

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right] \tag{2.5}$$

Here, the γ value is a discount rate and α is again a step size parameter. Comparing equation 2.5 with equation 2.4 it can be said that the Monte Carlo target R_t has been changed into the TD target $r_{t+1} + \gamma V(S_{t+1})$. This new target will still lead us to approximations of expected return because, starting from equation 2.3 and using equation 2.2:

$$\begin{aligned} V(s) &= E\left\{R_t \mid s_t = s\right\} \\ &= E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right\} \\ &= E\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s\right\} \\ &= E\left\{r_{t+1} + \gamma V(s_{t+1}) \mid s_t = s\right\} \end{aligned} \tag{2.6}$$

Eventually, both methods will come up with the same state values. The advantages are in the way these values are discovered.

2.1.4 Advantages of TD Learning

As stated earlier, in many learning problems TD methods have some advantages over other methods. These advantages were clearly pointed out by Sutton[64]. First, since all updates are made in a highly incremental fashion, TD methods are computationally simpler. Obviously this is an advantage when learning tasks have long episodes. In those cases, waiting for R_t to be known simply takes too long. Second, because TD methods make better use of past experience, they tend to converge faster and make better predictions.

This last point can be understood in the following way. Most real world problems require several actions to be performed before a favorable outcome is reached. Traditional methods tend to ignore the information contained in the sequence of states even though this information is often relevant to the state value estimates. In a way, the traditional prediction learning methods can be viewed as *supervised learning* methods since they learn to associate states with return values. Now suppose a completely new state is visited in a certain episode. Further assume that between this state and the end of the episode other states are visited that have been valued before and are in fact very familiar to the system. TD methods will make better use of this *experience* by basing the value of the new state on the values of familiar states much more quickly.

2.2 Actor-Critic Models

Many of the earliest systems using TD methods were actor-critic models. As early as 1977 an actor-critic model was applied to solve the notorious n-armed bandit problem (Witten

[69]) and in 1983 the scheme was used by Barto, Sutton and Anderson [5] to learn a more difficult version of the cart-pole balancing task. In this section, the base architecture of actor-critic models will be explained. Also, some words will be spent on the interest of other sciences in actor-critic models.

2.2.1 Division of Labor

The basic structure of actor-critic models is quite intuitive and is in need of only a short explanation. Actor-critic models have a separate structure to explicitly represent a function that maps states onto actions. This structure is called the *actor* and the state/action mapping function is usually referred to as the *policy*. A second structure, called the *critic* learns and stores the state value function. Whenever the actor performs an action, the new state will be criticized by the critic through a scalar signal called the *TD error*. Both actor *and* critic learning is based on this error. It is defined as follows:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{2.7}$$

Figure **2.2** illustrates the architecture. Both actor and critic use the system's current state as an input. Based on this state the actor will select an action moving the system into the following state. Next, the reward signal coming from the environment is passed to the critic to come up with the TD error signal. Finally, this signal is routed through both actor and critic to allow the structures to update their functions. If, for example, the value of the state turns out to be higher than was predicted by the critic, the actor will intensify its state/action association while the critic will update its state value estimate in order to be more accurate in the future.

2.2.2 Actor-Critic Models in Other Sciences

For some time now, TD methods have been studied by researchers in the fields of biology, psychology and neuroscience. The ability to make accurate predictions about the environment has been assumed to be of key importance to the survival of animals in the wild. Many biological experiments have suggested a reward signal to be involved in animal prediction learning. Since TD methods were designed to solve prediction problems and use reward signals as a basis for learning, these methods seem to have a strong correlation with systems found in nature.

Of all different TD methods, actor-critic models are most popular amongst researchers in the mentioned fields. Most TD learning methods learn action values rather than state values. Policies are then created by searching through the list of action values. Actor-critic methods are different in the sense that they hold an explicit representation of the policy. Because of this explicit association between states and actions, actor-critic models

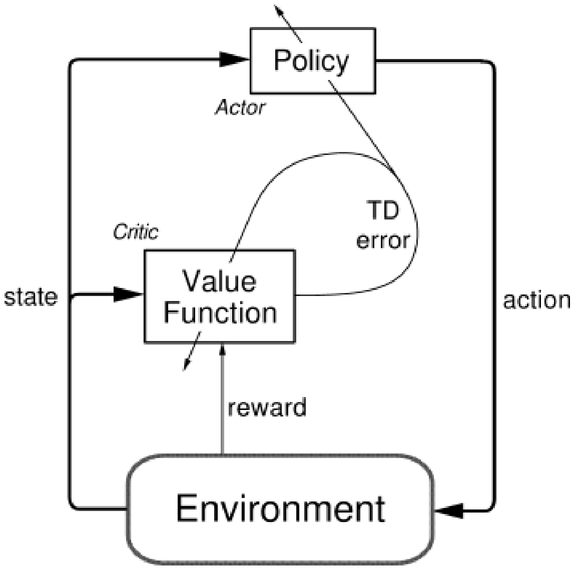

Figure 2.2: The actor-critic architecture

are usually a lot more efficient when it comes to action selection. Since it is hard to see how biological systems would generate policies from action values, biological theories often assume an explicit association to be present in the system. This makes actor-critic models more in line with common biological theories.

Chapter 3

Biological Foundations

This chapter treats the biological foundations of the computational model that will occupy the remainder of this document. It starts out with a global overview of the brain structure most important to the model, i.e. the basal ganglia (BG). After, *dopamine* will be introduced as the neurotransmitter that provides the connection between some neurobiological theories about learning and TD methods in general. The characteristics of dopamine behavior will be treated in detail as this behavior plays a crucial role in the workings of the model. Finally, some of the topological characteristics of dopamine inside the brain are discussed, as well as some of its functional roles.

3.1 The Basal Ganglia

Before diving into too much detail, a few words will be spent on the structure, function and connectivity of the basal ganglia (BG). This will clarify the terminology used in the remainder of this chapter and introduce some assumptions used in designing the model. By no means is the description meant to be exhaustive.

3.1.1 Structure

The BG are a collection of subcortical neuronal groups in the forebrain. The three main subdivisions are the globus pallidus, caudate nucleus, and putamen (figure **3.1**). Taken together, the caudate and putamen are referred to as the *neostriatum*, because they are the most recent BG structures to appear and are developmentally related. The neostriatum together with the globus pallidus form the *corpus striatum*. Some anatomists additionally view one or several other structures as being part of the BG, e.g. the subthalamic nucleus and the substantia nigra. In this document the governing convention will be upheld and the substantia nigra and subthalamic nucleus will be considered parts of the midbrain.

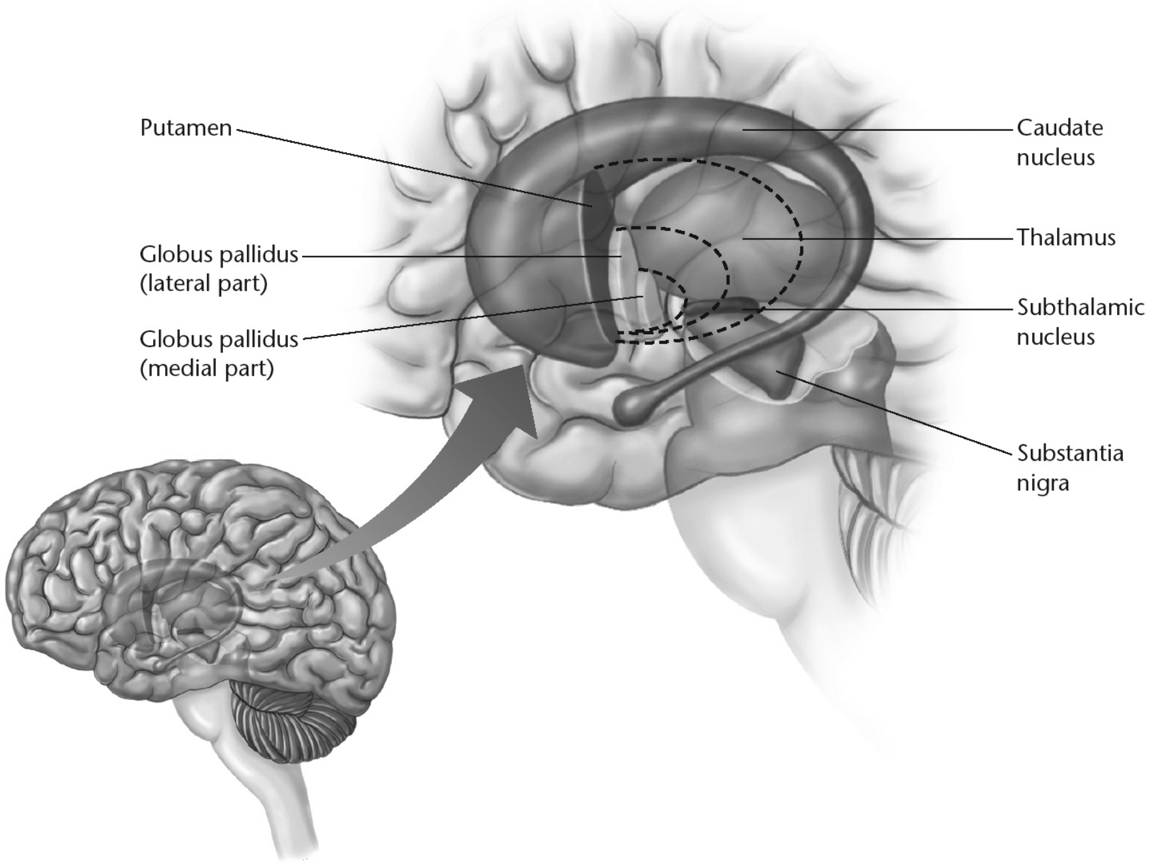

Figure 3.1: Structure of the basal ganglia (taken from [46])

3.1.2 Function

Probably more important than the proper subdivision of structures is their actual function. The BG, subthalamic nucleus, and substantia nigra participate in circuits with the cortex and thalamus to mediate aspects of motor control. These structures are also involved in the mediation of certain cognitive functions. These functions usually require the use of working memory or conscious processes.

Because the BG are recruited in learning stimulus-response (SR) tasks, many researchers assume them to store explicit SR mappings (e.g.[51]). Others believe the BG to facilitate or suppress SR-like associations that are stored in the cortex[53][43]. The latter hypothesis assumes the basal ganglia to have a more modulatory role in performing SR tasks and this assumption is used in this study.

3.1.3 Connectivity

The BG seem to be part of a functional loop that starts and ends in cortical structures. Direct projections exist from all major cortical areas onto the neurons of caudate and putamen (neostriatum), which collectively form the input system of the BG. An important feature of many cortical neurons seems to be their ability to sustain activation patterns over several seconds. This fact came to light when Funahashi et al[58] measured the responses of neurons in the pre-frontal cortex (PFC) in monkeys using a visual delayed response task. They clearly found evidence of CS activated neurons staying active or even increasing their activation until the monkeys made a response. Their findings suggests a cellular contribution to working memory by neurons in the PFC. As we will see in the next chapter, the model exploits this suggestion by assuming such neurons to provide input to the system.

The major outputs of the BG project through globus pallidus and thalamus to primarily motor cortex, premotor cortex and frontal cortex (figure **3.2**). Even though the BG plays a major role in motor control tasks, none of its structures are in projection pathways running from motor cortical areas to the spinal cord. Instead, they are part of a cortical-subcortical motor loop that is thought to monitor aspects of how motor activity as well as nonmotoric functions are progressing.

3.2 The Dopamine Response Signal

Since the early sixties, researchers have suspected dopamine neurons to influence the performance in tasks requiring motoric action[27][10]. However, the exact nature of this role was unknown for many years. Dopamine neurons were shown to fire phasically just before the initiation of motor actions and depletion of tonic dopamine levels were shown to cause problems in motoric control.

This section provides a review of biological knowledge about the behavior of dopamine neurons using an influential study as a guide. First, a relationship between dopamine levels and reward information is explored. Second, the ability of dopamine neurons to predict incoming rewards is discussed. Finally, a comparison is made with the TD error introduced in the previous chapter.

3.2.1 Dopamine and Incoming Reward

In an attempt to gain more insight into the relationship between dopamine levels and arm and eye coordination, a study was done by Romo and Schultz[57]. Monkeys were seated in an enclosed primate chair facing a box containing a small morsel of apple or cookie stuck to the end of a wire. The interior of the box was shielded from sight, but could be accessed by the animal with its hands. At a self-chosen moment, a monkey would release a touch sensitive key, reach inside the box, take the food item, move its arm toward its mouth and

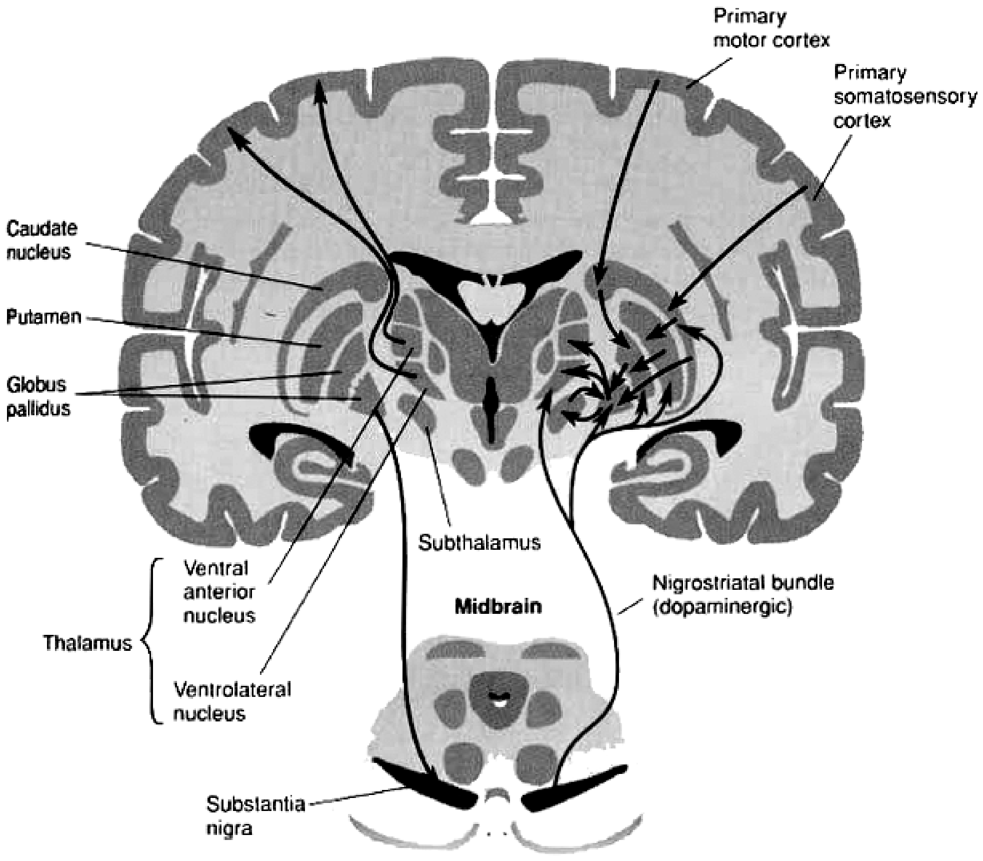

Figure 3.2: Projections of the basal ganglia (taken from [14])

eat the food. Dopamine responses were measured from the moment the key was released until the moment the monkey ate the food.

On touching a morsel of food, the monkeys clearly showed an outburst of dopamine neuron activity. This can be seen in figure **3.3** where the upper picture displays the measured dopamine level and the lower picture displays the activity of all neurons that were individually recorded during the task. Clearly, an outburst occurred as soon as the food was touched and subsided before the animal's hand left the box. Of course, this outburst could have been provoked by the sensation of touch itself in which case it holds no specific relation to the rewarding properties of food. However, something interesting happened when food was placed on the bottom of the box or when no food was present at all. Monkeys would now feel their way around the interior of the box causing touch sensations with no relation to food items. In these cases, monkeys did not show any outbursts of dopamine. Clearly, the outburst had something to do with the presence of food.

By now, many biological studies have shown dopamine neurons to respond phasically to

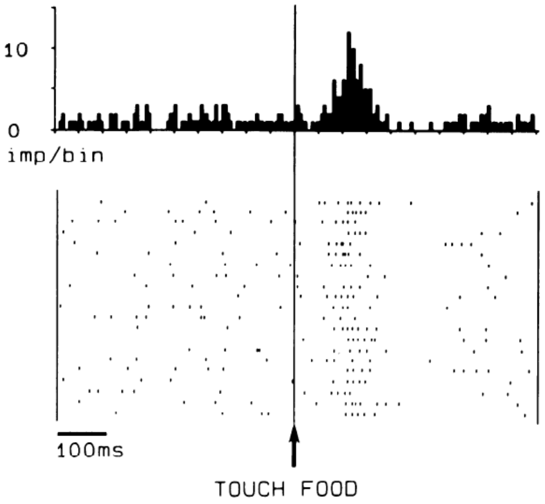

Figure 3.3: On the touch of food, dopamine neuron activity increases (taken from [60])

many different kinds of incoming appetitive stimuli. Recorded responses have been triggered by appetitive stimulation of the somatosensory, visual, and auditory systems suggesting dopamine responses to positively correlate with incoming reward information in general.

The study by Romo and Schultz showed no conclusive evidence supporting the idea of dopamine signals as initializers for spontaneous movement. It did however show some other, rather surprising results that will be treated next.

3.2.2 Reward Prediction

The first interesting finding was discovered by training the animals. Monkeys were conditioned to associate the sound of an opening door with the presence of food in the box. Whenever the sound was heard, the monkeys would reach inside the box and take the food. Occasionally, the door was opened without any food in the box. In these situations, the monkeys showed a dip in tonic dopamine level. The dip occurred at the time of expected reward (the moment the food should have been touched). Figure **3.4** clearly shows this finding. The upper image shows the normal case where food is found on the wire and the unconditioned monkey spontaneously reaches for it. The lower image shows what happens if, instead of a food item, the conditioned monkey finds the bare wire.

A second result was at least as interesting. Conditioned animals did not show any dopamine response on touching food. However, they did show a response on the occurrence of the conditioned stimulus (CS) as can be seen in figure **3.5**. Somehow, the CS had taken the place of the food as a cause for the outburst of dopamine.

In summary, outbursts of dopamine are seen when unpredicted reward information comes in. When the reward becomes predicted after conditioning, the outburst transfers

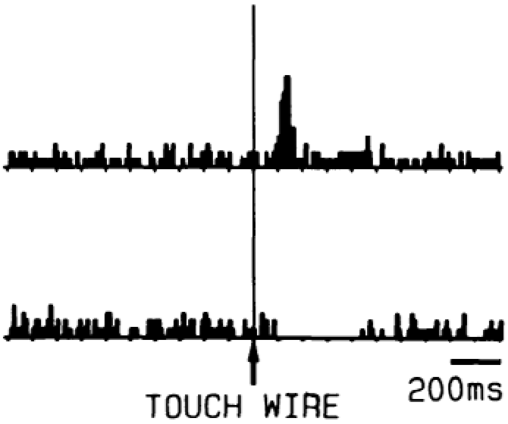

Figure 3.4: A dip in tonic dopamine level occurs when monkeys predict the availability of food but find an empty wire instead (taken from [60])

from the time of reward to the time of reward predicting CS and nothing happens to the tonic dopamine level at the time of actual reward. Finally, when a reward is expected but does not come in, a suppression of tonic dopamine level is seen.

Further research into the behavior of dopamine neurons has brought to light some more characteristics[61]. In the conditioned case, when a reward comes in earlier than expected, it will provoke an outburst of dopamine. However, no suppression occurs at the predicted time of reward. Interestingly, the occurrence of the reward seems to cancel the prediction. Repeating the earlier presentation of reward over several trials will eventually end up in the animal adjusting its reward prediction to fit the new situation. Another interesting effect

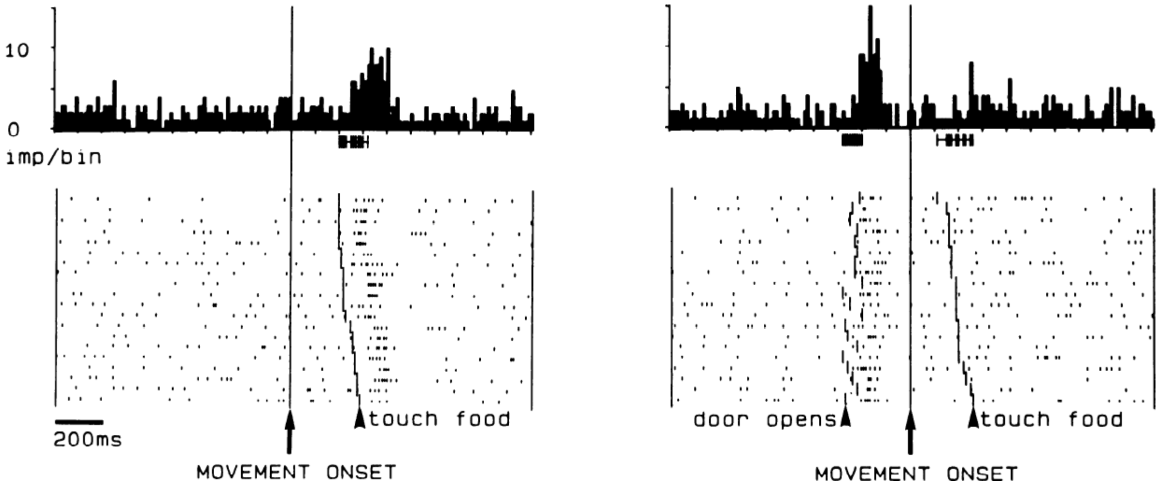

Figure 3.5: Dopamine signal transfers from reward occurrence to CS occurrence (taken from [60])

happens when a stimulus comes in that is very similar, but not equal to the conditioned stimulus. In that case, a moderate outburst occurs at the time of stimulus presentation immediately followed by a slight suppression. Again, no suppression is seen at the time of actual reward (figure **3.6** lower right image).

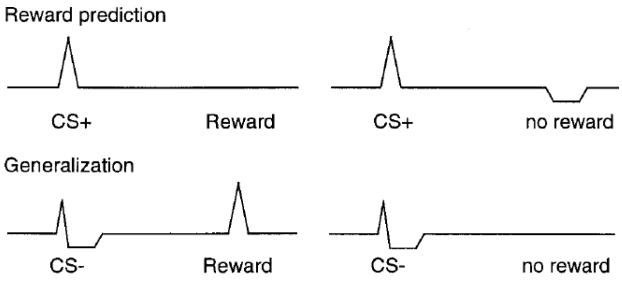

Figure 3.6: Dopamine responses with proper CS predicting reward (upper image) and stimulus similar to CS predicting reward (lower image) (taken from [60])

3.2.3 Dopamine and TD Error

These finds raise the interesting possibility of *CS chains*, where an incoming CS causes a prediction of a following CS. Initially, the dopamine outburst would transfer from the time of reward to the time of incoming CS predicting the reward. Next, another CS could predict the occurrence of the first CS making the dopamine outburst transfer to the earlier CS. This effect has been shown experimentally[60]. Dopamine outbursts will eventually always happen on the earliest reward predicting CS and will not be seen afterward. However, a suppression of tonic dopamine level occurs as soon as a subsequent CS or the reward itself is not seen by the animal.

This behavior of dopamine neurons holds a striking resemblance to the TD error of the previous chapter. In TD learning, the TD error provides the system with a target for learning. The error is based on the difference between a state value and the value of the subsequent state. In biology, a CS can cause an outburst of dopamine predicting future reward. Apparently, the brain can learn to associate the outburst and the CS on basis of a *subsequent CS* which can therefore be related to a *subsequent state* in TD learning. Suppressions of tonic dopamine levels act like the TD error in that the magnitude of a suppression shows the difference between what was expected and what actually came in. When no suppression occurs, the prediction was completely accurate. When tonic dopamine levels are thoroughly suppressed, the prediction was very wrong. This leads to the following formalization:

$$DopamineResponse(Reward) = RewardOccurred - RewardPredicted \tag{3.1}$$

3.2.4 Actor-Critic Learning

Because of the resemblances to the TD error, it is not too surprising that many researchers believe some form of TD learning to take place inside the brain. From biological data it seems perfectly clear that the brain is at least able to make predictions about subsequent outbursts of dopamine. If this was not the case, how could it be that no outburst occurs when a reward is received after a CS has predicted it? For this to work, the outburst needs to be actively prevented from happening. Preventing the outburst can only be done if the brain has some prior knowledge about its magnitude and timing. From TD learning we know that knowledge about future states can be acquired using a combination of the TD error and incoming reward information. Or in our case, dopamine fluctuations and unpredicted outbursts. Therefore, brain structures coding this knowledge could be said to collectively form the *critic* of a creature.

Outbursts of dopamine are now believed to alert animals to incoming reward. Animals can learn about appetitive stimuli because these stimuli will cause a dopamine outburst when they occur without being predicted. Next, animals can learn associations between incoming stimuli and the future occurrence of appetitive stimuli, allowing them to learn behaviors that will increase their chances of actually earning the appetitive stimuli (e.g. approaching behaviors). Sequences of actions can thus be formed, ultimately allowing the animal to retrieve its reward. Learning the values of behaviors in different scenarios can be done using dopamine fluctuations and unpredicted rewards. Explicit associations between incoming stimuli and behaviors can be coded inside the brain. Brain structures containing this code could be said to collectively form the *actor* of a creature.

3.3 Dopamine and the Brain

What are the topological aspects of dopamine in the brain? In this section, the brain structures involved in the activation of dopamine neurons are discussed, as well as the structures that receive dopaminergic projections. Finally, the influence of dopamine on action selection is clarified on the basis of neurobiological studies.

3.3.1 Dopamine Activation

The substantia nigra compacta (SNc) is responsible for most of the dopamine production in the brain, and therefore plays a vital role in reward and addiction. A majority of SNc activation comes from the pedunculopontine tegmental nucleus (PPTN) that is part of the brain stem and caudal to the SNc (see figure **3.7**). Lesions in the PPTN will usually result in hemiparkinsonian-like symptoms[39] which illustrates the importance of the PPTN as activator of neurons in the SNc, since Parkinson disease is caused by a dopamine deficiency and lesions in the PPTN can cause many of the same symptoms. The symptoms and causes of Parkinson's disease will be treated in more detail later in this chapter. PPTN neurons have been found to fire phasically in response to primary reward or reward-predicting conditioned stimuli. This leaves them well situated to provide input to the SNc neurons[25].

From where does the PPTN receive these reward related signals? Brown et al[11] suggest primary reward signals to come from the lateral hypothalamus, whereas reward-predicting signals travel via the ventral striatum - ventral pallidum pathway (see figure **3.7**). This pathway receives its main input from the limbic cortex[50]. A working memory trace of a CS in the limbic cortex could therefore activate the ventral striatum, causing a sustained activation of ventral striatal neurons which will eventually provoke a net excitation of PPTN neurons through double inhibition (ventral striatum to ventral pallidum to PPTN). These suggestions are based on acquired knowledge about the interconnectivity of the structures involved. Neurons in the ventral striatum, called *matrisomes*, have been found to respond to both predicted and primary rewards[71] and project to the ventral pallidum. The ventral pallidum projects about one fourth of its collaterals to the PPTN[52].

3.3.2 Dopamine Inhibition

Providing a mechanism for dopamine responses to primary and predicted rewards does not completely explain the dopamine behavior described earlier. When primary rewards are predicted, no dopamine outburst is seen at the time of actual reward. What mechanism suppresses this response? Furthermore, the suppression in tonic dopamine levels that occurs when a predicted reward is absent or late, is still in need of explanation.

Striosome cells located in the caudate and putamen, provide a strong source of SNc inhibition and, in turn, receive dopaminergic projections from the SNc[19](see figure **3.7**). This could possibly explain both phenomena as was suggested by Brown et al[11]. Based on CS input from the limbic cortex, striosome cells could collectively learn to provide a timed inhibition of SNc neurons. Dopaminergic feedback from the SNc neurons could act as a teaching signal that enables the striosomes to learn CS associations and reward delay times. When a predicted reward comes in, the memory trace of the predicting CS in the limbic cortex will cause a well-timed excitation of striosomes that will annihilate the SNc excitation triggered by the reward information coming in from the PPTN. When a predicted reward fails to occur, the striosomes will still inhibit the SNc activation but, without actual reward excitation from the PPTN, the inhibition now results in a suppression of tonic dopamine levels.

3.3.3 Dopamine Projections

So what brain structures are influenced by dopamine responses? Based on biological data, Groves et al[33] suggest that virtually every striatal neuron will get dopamine input from neurons in the SNc. The cortical dopamine innervation in monkeys is highest in the frontal

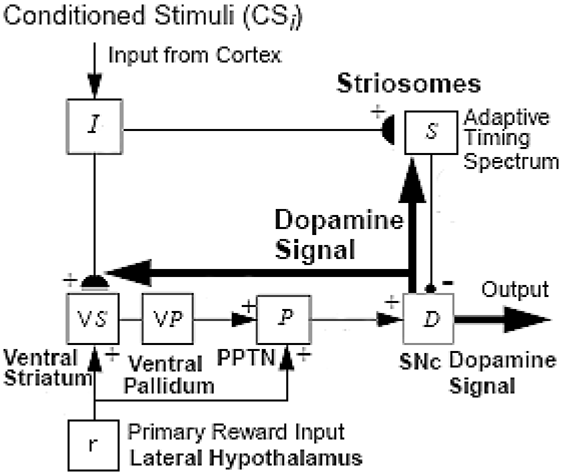

Figure 3.7: Structures activating dopamine release in SNc (based on image in [11])

lobe, is still sizable in parietal and temporal lobes, and is lowest in the occipital lobe[7][68]. Cortical dopamine synapses are predominantly found in layers I and V-VI, contacting a large proportion of cortical neurons there. These data suggest that the dopamine response advances as a wave of activity from the midbrain to striatum and frontal cortex.

There are two distinct dopamine receptor types. About 80% of receptors in the striatum are of type D1 and are located predominantly on neurons projecting to internal globus pallidus (GPi) and substantia nigra pars reticulata. The remaining 20% of dopamine receptors in the striatum are of type D2 and are located mostly on neurons projecting to external globus pallidus (GPe)[8][23][36][1]. When dopamine in released, D1 receptors will increase the excitability of neurons in a depolarized state[13][16][35][47]. Interestingly, D1 receptors will reduce excitability when neurons are in a hyperpolarized state[35]. This could have some interesting consequences since the medium spiny neurons (MSNs) which form ninety to ninety-five percent of all striatal neurons have been found to be bistable in nature[21]. Bistable neurons can switch between two stable but different membrane potentials and so dopamine could affect a neuron in opposing ways depending on which of the two membrane potentials is active.

In contrast to D1 receptors, D2 receptors reduce excitations evoked by other receptors (like NMDA and AMPA receptors) at any membrane potential[15][70].

The D1 and D2 receptor characteristics provide a straightforward prediction about the effect of dopamine on synaptic plasticity[30]. Since more active cells undergo long term potentiation (LTP) whereas less active cells undergo long term suppression (LTD) and since dopamine has an effect on the activation level of cells, dopamine signals must have an effect on activity dependent learning. When an outburst of dopamine makes cells more excitable through D1 receptors, these cells are more likely to undergo LTP. When dopamine release makes a cell less excitable, either through D1 receptors on cells in a hyperpolarized state or through D2 receptors on cells in whatever state, this will make the cell likely to undergo LTD. Suppression of tonic dopamine levels will have the opposing effect. In the next section, more words will be spend on this topic.

3.4 Dopamine and Action Selection

It is a widely accepted idea that a functional dopamine system is of crucial importance to the fine tuning of motor actions. In this section, the role of dopamine as an indirect modulator of motor actions is explored, as well as a theoretical mechanism for the interactions between dopamine system and motor cortex. This mechanism was first proposed by Frank[30] who based it on knowledge about the nature of impairments suffered by Parkinson's patients. To provide context for the treatment of the mechanism, this section starts out with a brief overview of these impairments.

3.4.1 Parkinson's Disease

Parkinson's disease (PD) is believed to be caused by an insufficient release of dopamine at receptor sides. Patients suffer tremors, muscle rigidness, postural instability, speech problems and slowness of movement. Furthermore, patients show impairments in higher cognitive tasks. These impairments can be broadly divided into two categories. The first category contains impairments that are *frontal-like* in nature. Patients show decreased ability in performing tasks involving attentional processes or working memory[66][56]. The second category holds impairments in implicit learning. Tasks requiring implicit learning do not require working memory or conscious knowledge of tasks demands and frontal patients do not show these deficits[49].

Compensating the lack of dopamine through dopaminergic medication often leads to other impairments[18][55][4]. These impairments occur whenever medicated patients are asked to perform tasks that rely heavily on the use of negative feedback. A possible explanation for this effect is that dopaminergic medication provokes an overdose of tonic dopamine in relatively undamaged areas of the brain. Elevated levels of tonic dopamine now restrict patients from learning on basis of negative feedback[18][4]. These areas will not be able to suppress dopamine levels below baseline and as a result, trial and error learning becomes impaired.

3.4.2 The Modulating Basal Ganglia

The seemingly unrelated cognitive deficits of Parkinson's disease can be tied together by assuming a modulatory role for the basal ganglia[30]. The basal ganglia are now in charge of suppressing and facilitating SR associations stored in the cortex and do not store any

SR mappings explicitly. The dopamine system can be functionally considered part of the basal ganglia. In doing so, it can be assumed to be a modulator inside a modulator. These assumptions allow us to tie together some of the seemingly unrelated cognitive deficits stemming from dopamine dysfunction in the basal ganglia. Tasks that seem to involve the frontal cortex could be impaired because the BG is interconnected in a functional circuit with the PFC[29][3], while the implicit learning tasks could be impaired because of damage to a neostriatal learning mechanism[41][49].

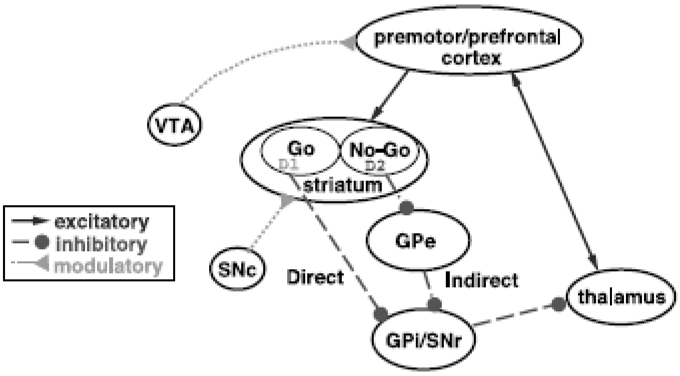

Figure 3.8: Structure of the basal ganglia (taken from [30])

In the context of motor control, various authors have suggested that the BG selectively facilitates the execution of a single motor command[43][31]. Amongst competing actions the BG would modulate motor execution by signaling *Go* for the most appropriate action while signaling *No-Go* for all others[37]. Besides allowing for clear decision making, this functionality helps to string together simple motor actions to form a complex motor sequence[30].

3.4.3 The Go/NoGo Mechanism

As was seen earlier, the major input segment of the BG is the striatum. The striatum receives input from multiple cortical areas and creates a functional loop by projecting through the globus pallidus and substantia nigra to the thalamus, ending up in cortical areas like the premotor cortex (PMC)[2]. Toward the thalamus there are two pathways of information[32]. The first, called the *direct* pathway is formed by inhibitory projections from the striatum to the *internal* segment of the globus pallidus (GPi). The second, called the *indirect* pathway is formed by inhibitory projections from striatum to the *external* segment of the globus pallidus (GPe). The GPe then converges the two pathways by inhibiting the GPi. In the absence of striatal firing, the GPi tonically inhibits the thalamus and so when striatal neurons manage to inhibit the GPi through the direct pathway, this allows the thalamus to get excited from other excitatory projections[38][31]. When this happens, thalamocortical projections enhance the activity of the motor command that is currently represented in the PMC. Viewed in such a way, the direct pathway seems to facilitate Go responses to actions by disinhibiting the thalamus allowing it to help select an action represented in the PMC. Conversely, the indirect pathway facilitates No-Go responses to actions by inhibiting the thalamus which will now suppress an action represented in the PMC.

3.4.4 D1 and D2 Receptors

The biological structure just described raises some interesting possibilities when we consider the distribution of D1 and D2 receptors over the two pathways. D1 receptors predominate in the direct pathway and D2 receptors predominate in the indirect pathway[26][45][22][19]. Therefore, outbursts of dopamine have an excitatory effect on direct/Go pathway activation and an inhibitory effect on indirect/No-Go pathway activation[12][44]. Depletion of dopamine has the opposite effect, biasing the indirect pathway to be overactive[20].

Another interesting possibility comes from the bistable nature of most striatal cells. On an outburst of dopamine, the D1 receptors in the direct/Go pathway will provide some *contrast enhancement*[40][17][28]. Biological noise in striatal Go neurons will have less of an impact on action selection, since neurons in a hyperpolarized state will be suppressed by a dopamine outburst, while neurons in a depolarized state will become more activated. This reduces the signal to noise ratio in the direct pathway and may help to determine which among several responses is most appropriate to select.

Chapter 4

The Model

How can all this neurobiological knowledge be used to construct a computer model of biologically inspired actor/critic learning? Such a model needs to honor the wiring of the brain as well as reproduce the causes and effects of dopamine outbursts. Furthermore, the model needs to be able to actually learn tasks.

This chapter describes the details of a model that was constructed to do all of these things. For clarity, it will only describe the workings of the model without going into too much detail about the learning results or the similarity to brain structures or behavior. Those topics will be thoroughly treated in the next chapter.

The chapter is divided into three parts treating one major subdivision of the model each, i.e. the input module, the actor module and the critic module. Since the actor module is heavily dependent on dopamine for learning the chapter covers the critic first. However, in a loop of the model, the input system first calculates an input pattern after which the actor calculates a response to that input. Next, the critic calculates the new dopamine signal based on prediction and reward and finally both actor and critic use this signal to update their weights. This order of calculations was taken from the actor/critic models as described in chapter 2.

4.1 Model Input

Of great importance to both actor and critic is the input mechanism of the model. When a CS comes in, the system will need to respond by selecting an action, waiting for a possible reward and updating the learning parameters of both actor and critic. In order to be able to make comparisons to biological data, rewards are delayed for a number of time steps with each time step roughly corresponding to a hundred milliseconds. However, the delaying of rewards raises the first problems.

Since the model needs to be able to learn SR associations while all learning is driven by rewards, the gap between stimulus appearance and incoming reward needs to be bridged. Learning targets can not be adequately constructed if information contained in the stimulus is lost before a reward comes in. Somehow this information needs to be stored. Therefore, the model has a separate structure managing the inputs of the system.

4.1.1 State Input

Environment states are communicated to the model in the form of value vectors. A set of subsequent states I is stored as a list of these input vectors. Every time step a vector of state values I_t is passed on to the input neurons. In the model, these input neurons are called *limbic* neurons as the prefrontal limbic system in the brain is known to provide major input to the striatum[19][42].

CS information will linger inside the limbic neurons, degrading over time. The temporal gap between CS and reward is assumed to be bridged by these sustained CS inputs[58] providing a mechanism for the needed storage of information. New incoming values will only become represented by the limbic neurons if they are stronger than the values currently held. Finally, all activation will be canceled completely when a reward r_t is received.

The mathematical notation L_t^i is used to denote the output of limbic neuron i at time t. Similar notation is used for the inputs to each individual neuron, i.e. I_t^i. For a single input neuron L^i, the output at time t now becomes:

$$L_{t+1}^i = \begin{cases} I_{t+1}^i & \text{if } I_{t+1}^i > L_t^i \\ 0 & \text{if } r_t > 0 \\ 0 & \text{if } L_t^{max} - \alpha > 2L_t^i \\ 2L_t^i - L_t^{max} - \alpha & \text{otherwise} \end{cases}, \tag{4.1}$$

where α influences the maximal length of time before the signal degrades to zero and L_t^{max} denotes the maximal value with which the stimulus can come in. After decreasing to zero the limbic output will be kept at zero unless new inputs start the degeneration process again.

At first sight these sustained outputs of limbic cortical neurons might seem a little artificial. The shape of the function was chosen because it sustains neuron activation at a high level for a number of time steps after which it will drop quickly. As was described in the previous chapter, some cortical neurons have the ability to stay active until a response is made by the animal (see [58]). These neurons do not show an exponential decrease of activation, but rather keep their activation at a high level for some time.

4.2 The Critic

The structure of the critic is based on a model by Brown et al[11]. This model was created to explain the dopamine response characteristics described in chapter 3. Based on existing biological knowledge Brown et al suggested a setup that, with some learning, would reproduce these responses. The use of biological knowledge and the explanatory success of the model makes it an excellent candidate for filling the role of critic in the present study. Still, the model was not completely duplicated. Learning rules and update rules were customized and some changes were made to parts of the mechanics. Let us see how it works.

4.2.1 SNc Output

In the previous chapter the SNc was described as a structure providing a large portion of total dopamine production in the brain. In the model, the SNc has been simplified to a single neuron and dopamine output is the linear combination of its activation and suppression. Activation comes from the PPTN neurons that were previously said to have a major excitatory influence on the SNc in the brain. Timed suppressions are provided by striosomal cells located in the striatum. These two structures, PPTN and striosomes will be examined in detail later in this section.

The level of dopamine at any given time t is modeled as the output of the SNc at that time. For clarity, the term D_t is used to express both the SNc output and the dopamine level at time t:

$$D_t = -\Gamma + \sum_p P_t^p W^{pd} - \sum_s S_t^s W^{sd}, \tag{4.2}$$

where W^{pd} is the weight of the connection between PPTN cell p and the unique dopamine cell d and W^{sd} is the weight of the connection between striosomal cell s and the dopamine cell d. Output of a single PPTN cell p at time t is written as P_t^p and similar notation is used for output of striosomal cell s. The activation is corrected by a term Γ which is the threshold value of the combined PPTN neurons and has value 0.1.

In the model all weights of connections with the SNc are usually set to 1 which simplifies the equation to:

$$D_t = -\Gamma + \sum_p P_t^p - \sum_s S_t^s, \tag{4.3}$$

The dopamine level has a zero baseline and is allowed to fluctuate between values -1 and 1. Every time step its value is used for updating learning parameters of different structures which will be explained later on.

4.2.2 Dopamine Neuron Activation

In what way is the SNc connected to the input of the model? In the previous chapter there was mention of a ventral striatum - ventral pallidum pathway that would excite the PPTN through double inhibition. For simplicity the model lumps together these two structures and refers to it simply as the *ventral striatum* (see figure **4.1**). This ventral striatum gets excitatory input from the limbic cortex. Also, it receives primary reward r_t from a single cell structure representing the *lateral hypothalamus.* Activation at time t of a ventral striatal neuron v is expressed as follows:

$$V_t^v = r_t W^{rv} + \sum_l L_t^l W^{lv}, \tag{4.4}$$

where W^{rv} is a non-adaptive weight between lateral hypothalamus and the ventral striatum cell v and W^{lv} is an adaptive weight between the limbic cortex neuron l and ventral striatum cell v. All weights are randomly initialized to have values between .0 and .05.

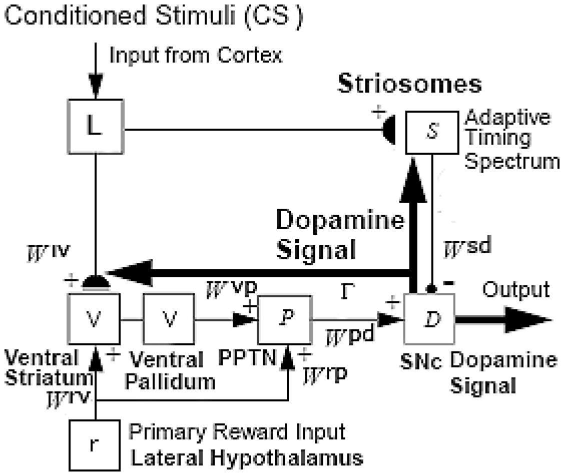

Figure 4.1: Structures activating dopamine release in SNc (based on image in [11])

Through training, the ventral striatum neurons will learn to associate CS signals, coming from the limbic cortex, with future rewards. When a CS is recognized, the ventral striatum cells will excite the PPTN causing a rise in dopamine level. This then provides an explanation of how the brain is able to learn to elicit a dopamine response at the time of incoming CS. However, left this way, the mechanism would cause a *tonically* elevated level of dopamine between the occurrence of a CS and the time of reward, because the limbic input neurons will sustain the CS induced activation for this period of time. From the previous chapter we know the dopamine response to be phasic instead of tonic. Following an assumption made by Brown et al[11], the PPTN will change the signal from tonic to phasic through accommodation. The model uses the following formula to calculate an accommodation value A_t^p for PPTN neuron p at time t:

$$A_t^p = \beta A_{t-1}^p + \gamma P_{t-1}^p, \tag{4.5}$$

where β and γ are constant scalars and P_{t-1}^p is the activation of PPTN neuron p at time $t-1$. This activation is now altered by the accommodation term and is of the following form:

$$P_t^p = -A_t^p + r_t W^{rp} + \sum_v V_t^v W^{vp}, \tag{4.6}$$

where W^{rp} is the weight between the lateral hypothalamus neuron and PPTN neuron p and W^{vp} is the weight between ventral striatal neuron v and PPTN neuron p (see figure **4.1**). So far, all simulations have assigned these PPTN weights values of 1.0. This simplifies the above equation into:

$$P_t^p = -A_t^p + r_t + \sum_v V_t^v, \tag{4.7}$$

Recall that PPTN neurons have a threshold value Γ. The effect of this value is a simple subtraction of the added PPTN signals and was already made part of the SNc activation function.

4.2.3 Dopamine Neuron Suppression

Now that there is a mechanism for dopamine neuron activation we are in need of a mechanism explaining its timed suppression. Brown et al[11] suggested that striosomes, located in the striatum, are the source of this inhibition and set up a detailed biological model explaining their collective ability to time responses. This model was not duplicated here for a couple of reasons. First, the striosome model of Brown et al was set up to explain the ability of timed responses and nothing more, making it essentially a model inside a model. Since the current study does not aim to reproduce the results of this additional model it was decided to accept it as a given. Second, a structure producing very similar behavior could be set up rather easily avoiding the implementation and calculation of a highly complicated mechanism. For detailed information about the original model, the reader is directed to the original document[11]. Only a brief review is given here.

According to Brown, striosome neurons are able to time their activation collectively. Incoming CS information from the limbic cortex would make striosome neurons fire with different delays. Timed activation would be obtained by strengthening the connection between the limbic neurons and the striosomes that just happen to fire at the right time, i.e. the time of subsequent reward. This would cause a timed inhibition of the SNc, suppressing the incoming reward signal.

The current model duplicates the basic idea, but leaves out the biological causes for the delayed firing of striosomes. For a set of subsequent time steps, one striosome at a time will fire a response. This firing starts one time step after the occurrence of a CS on the input layer and lasts for a while (ten time steps usually). By strengthening the connection between the striosome with the right delay and the unit representing the CS in the limbic cortex, a timed suppression of dopamine is obtained. So, for some CS onset time t^{CS} a striosome s will fire at time $t^{CS}+s$ (where s is treated as an index) with strength:

$$S_t^s = \sum_l L_t^l W^{ls}, \tag{4.8}$$

where W^{ls} is an adaptive weight between limbic cortex neuron l and striosomal cell s. Similar to the ventral striatal weigths, these weights are initialized randomly with values between .0 and .05. When trained properly, the striosome dedicated to the actual delay between CS and primary reward will provide the largest output inhibiting the dopamine signal provoked by the primary reward coming in through the PPTN.

4.2.4 Training the Critic

Note that all adaptive weights of the critic exist either between limbic cortex and striosomes or between limbic cortex and ventral striatum (in **4.1** adaptive connections are displayed as half circles). By training these structures properly, the system starts to reproduce many of the behaviors of dopamine concentrations inside the brain. As in TD learning, where the TD error is used to train both critic and actor, the current model uses the dopamine signal to do all the training. In this case, the actor exploits the deviations from tonic dopamine level to train the ventral striatum as well as the striosomes. The difference between tonic dopamine level and the magnitude of the phasic outburst is used as an additional modifier in a Hebbian learning rule. The weight between limbic cortex neuron l and ventral striatum neuron v will be modified using the following update mechanism:

$$\Delta W^{lv} = \eta D_t^+ L_t^l V_t^v \left[1 - W^{lv}\right] - \varphi D_t^- L_t^l V_t^v W^{lv}, \tag{4.9}$$

where η and φ are constants representing LTP and LTD respectively and are both set to .5. D_t^+ and D_t^- denote positive and negative deviation from tonic dopamine level that gets mapped into the [0, 1] range before the calculation is carried out. Only one can be greater than zero and the other will be set to zero. Weights will be set to zero if application of the equation above makes them drop below zero. Time indices are left out of weight symbols for convenience of reading.

A similar rule is used to update the weights of the connections with the striosomes. As can be seen, all learning is *Hebbian-like* in the sense that it is activity dependent in nature. As in Hebbian learning, mutually firing neurons are associated. However, their learning is now influenced by the current dopamine level. When dopamine is at its normal level, no learning is done because the deviation terms are both zero. When dopamine is above baseline, the connections of limbic cortex to striosomes and ventral striatum will be strengthened. When dopamine is below baseline, the connections will be weakened.

As a result, ventral striatal cells will learn to associate patterns displayed by the limbic cortex with incoming reward. Similarly, the reward induced dopamine signal will strengthen the connection between the limbic cortex pattern and the striosomal cell that just happens to be firing at the right time. After a sufficient training period this neuron will inhibit the reward induced dopamine outburst.

4.3 The Actor

The actor is based on a model built by Frank[30]. In an attempt to explain some seemingly unrelated symptoms of Parkinson's patients and known effects and side effects of dopaminergic medication, a model was built that could learn SR associations based on a dopamine error signal. As with the critic, this model aimed to do justice to existing biological knowledge about the connectivity of different neurological systems in the brain, making it an excellent candidate for filling the actor role in the present model.

The model of Frank assumes the basal ganglia to perform a modulating function in action selection. It is based on the idea that different possible responses are competing at any given time and the basal ganglia help determine which response should win the competition. Therefore, the model has a Go and a NoGo response for every action. These responses are generated using two different neural pathways that we will call the Go and the NoGo pathways. Below, the mechanism is treated in detail and figure **4.2** depicts the structure of the actor..

4.3.1 The Go/NoGo Pathways

As was seen earlier, striatal neurons receive input from cortical structures. Therefore, the input of both actor and critic is coming from the limbic input neurons and in both models these connections are adaptive. Additionally, the actor receives excitatory input from the pre-motor cortex (PMC), which is treated as a subsystem of the actor.

The striatal neurons are divided in modules. For every possible action there is a module of striatal neurons providing a Go association and a module of striatal neurons providing a NoGo association. The output of Go neuron g for action a is written as G^{ag} and for NoGo neuron n as N^{an}. All Go and NoGo neurons get input from the limbic system plus the output of the PMC neuron dedicated to their action. The output of PMC neuron a is simply written as A^a treating a as both neuron index and action index for ease of reading.

The output of Go neuron g dedicated to action a now becomes:

$$G_t^{ag} = A_t^a W^{ag} + \sum_l L_t^l W^{lg}, \tag{4.10}$$

where W^{lg} is the weight between limbic neuron l and Go neuron g and W^{ag} is the weight between the PMC neuron a dedicated to action a and Go neuron g. These weights are initialized with values taken from a normal distribution with mean .025 and standard deviation .005. A similar equation is used for calculating the output of NoGo neurons. Clearly, we must first calculate an initial PMC signal for each action in order to get a response from a Go or NoGo module. The PMC bases its initial action choice solely on the input from the limbic neurons:

$$A_t^a = \sum_l L_t^l W^{la}, \tag{4.11}$$

where W^{la} denotes the adaptive weight between limbic neuron l and PMC neuron a and is randomly initialized to be between .0 and .05. The PMC uses a simple Hebbian learning rule to learn to associate actions with input patterns from the limbic cortex. After many co-occurrences of input and action the PMC will be able to decide on an action all by itself. However, a threshold secures that only many of these co-occurrences can make the PMC decide to take action on its own and the length of training time will not be enough to make this happen. Therefore during training, the PMC needs bottom up support from the thalamus.

The magnitude of this support is decided in a battle between Go and NoGo responses. During training, the model will learn to associate high Go responses with beneficial input patterns and high NoGo responses with input leading to depressions of dopamine. Through a neural mechanism capable of calculating the difference between Go and NoGo responses for every action, the actor will be able to support or suppress an action selected by the PMC. How do Go and NoGo neurons learn opposing responses to input patterns? This will be described next.

Striatal Associations

The difference between Go and NoGo neurons is in there dopamine receptors. The Go association neurons have receptors of type D1 that were described earlier to have a stimulating effect on learning in the presence of dopamine. The NoGo associations have receptors of type D2 and will respond to outbursts of dopamine by making striatal neurons less excitable and therefore less likely to associate a limbic input pattern. As a result, the striatal Go neurons will associate strongly when dopamine levels are high, while the NoGo associations will associate better when dopamine levels are low. For ease of reading the inputs from the limbic neurons and the input from PMC neuron a dedicated to action a are merged into one input vector L^a and so L_t^{al} is meant to denote the output of input l from the input pattern dedicated to action a at time t.

$$\Delta W^{lg} = \eta D_t^+ L_t^{al} G_t^{ag} \left[1 - W^{lg}\right] - \varphi D_t^- L_t^{al} G_t^{ag} W^{lg}, \tag{4.12}$$

for weights between input and Go layer neurons. For NoGo neurons this becomes:

$$\Delta W^{lg} = \eta(1 - D_t)^+ L_t^{al} N_t^{an} \left[1 - W^{lg}\right] - \varphi(1 - D_t)^- L_t^{al} N_t^{an} W^{lg} \tag{4.13}$$

It is important to note that the dopamine signal gets scaled between zero and one *before* these calculations are carried out. Furthermore, LTP and LTD parameters η and φ both have values of .075.

By reversing the dopamine signal for NoGo neurons the effect of D2 receptors is simulated. And so the model does not just learn what actions to prefer based on some input pattern, but also learns explicitly what actions to reject. Now all that is needed is some neural mechanism to combine these signals into Go/NoGo responses for actions. Let us see how this is done.

From Association to Action

The Go and NoGo neurons in the striatum have inhibitory projections to neurons in the globus pallidus (GP). There are two pathways, the NoGo neurons with the D2 receptors project to the *external GP* (GPe) while neurons of the Go pathway with the D1 receptors project to the *internal GP* (GPi)[3]. All these connections are inhibitory and so are the projections from GPe to GPi. Both GPi and GPe are modeled to have one neuron for each action. The Go neurons for action a only project to the GPi neuron i for action a and the output of this neuron is labeled GPi^{ai}. Similarly, the NoGo neurons for action a only project to the GPe neuron e for action a and the output of this neuron is labeled GPe^{ae}. The output GPe^{ae} at time t now becomes:

$$GPe_t^{ae} = 1 - \sum_g N_t^{an} \tag{4.14}$$

The sum of Go outputs is subtracted from 1.0 to show the inhibitory nature of the signal. Specifying the level of tonic inhibition to be exactly one is of no real importance since the output of GPe neurons are subtracted from the outputs of GPi neurons. These are now calculated using:

$$GPi_t^{ai} = 1 - \sum_g G_t^{ag} - GPe_t^{ae}, \tag{4.15}$$

The GPi neurons now have one inhibitory value for each action that was calculated by taking the difference between the collective Go and NoGo signals for each action. Inhibitory signals below zero are not allowed and will be set to zero when they occur.

As a result, strong associations of Go neurons with some input pattern will inhibit the GPi through the *direct pathway* while strong associations of NoGo neurons will lift inhibition from the GPi via the *indirect pathway.* Since the GPi tonically inhibits the thalamus, the end result of high NoGo activation in the striatum is an inhibition of the thalamus. Conversely, a high activation of Go neurons will lift inhibition from the thalamus. The thalamus can now be activated by other brain systems supporting the PMC activation. The model uses the PMC itself to excite the thalamus and so there is a PMC-thalamus-PMC modulating loop. The output of action specific thalamus neuron t is labeled as T^{at} and calculated using:

$$T_t^{at} = A_t^a - GPi_t^{ai}, \tag{4.16}$$

where A_t^a was calculated using equation 4.11. This signal is added to the initial value of the corresponding PMC neuron and a *winner takes all* rule decides which action is going to be carried out provided that at least one of the PMC action neurons has an activation above the threshold.

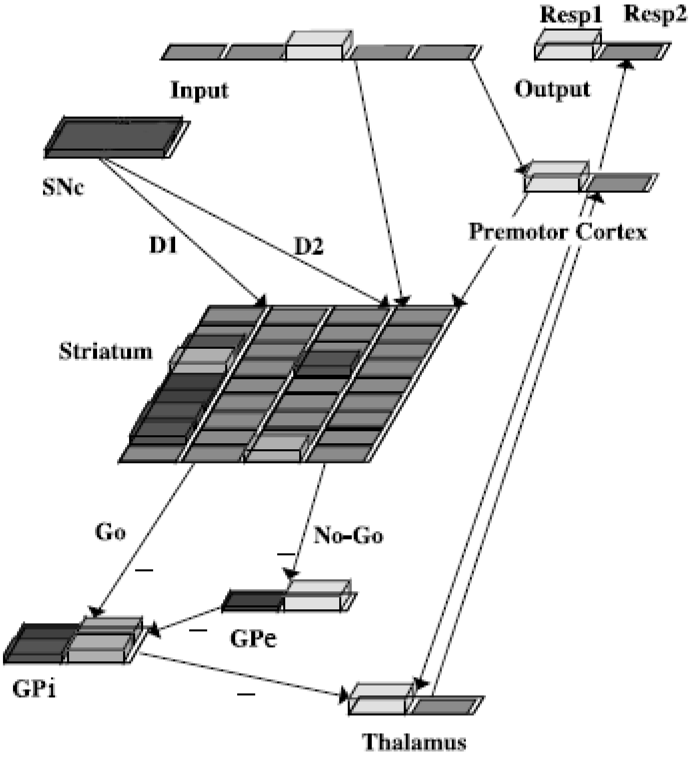

Figure 4.2: Structures involved in action selection (adapted from [30])

4.3.2 Exploration

To find out what actions should go with what input pattern, the model needs to try out different actions in different situations. This can not be done if the system only chooses *winning* actions. What we need is an exploration mechanism. By adding random noise to the output of the PMC the system is made to explore. Of course this form of exploration is not suspected to be part of an actor/critic mechanism in the brain. Unfortunately, as of yet there does not seem to be any biological theory of exploration detailed enough to be used in the present study. The next chapter will describe the details of the exploration mechanism that was used for experimentation.

Chapter 5

Results

A two-armed bandit task was used to investigate the workings of the model. All results in this chapter were obtained trying to solve either this problem or a variant of it and no special modifications were made to enforce particular behaviors of the system. The first section introduces the problem, the second section shows the results of learning the task, the third section treats the behavior of the critic during training and the fourth section does the same for the actor. Finally, some preliminary results suggest a direction for future research.

5.1 The Two Armed Bandit Problem

In order to test the learning capabilities of the system a version of the so-called *two armed bandit problem* was implemented. Given a slot machine with two levers, the system needs to find out what lever to prefer. Or stated differently, which lever returns the most profit in the long run? To differentiate the rewarding effect of the two levers, each lever has a normal distribution associated to it with a mean different from the other lever. When the lever is pulled, a reward value will be drawn from its associated distribution.

5.1.1 Task Difficulty

Even though this problem is famous, it is by no means the most difficult of problems commonly used as benchmarks in the field of reinforcement learning. However, for our purposes it is enough to show the system is capable of learning. Moreover, this study does not aim to compete with established reinforcement learning methods.

The two-armed bandit problem not being the hardest of tasks does not mean it is no challenge. The system will have to learn to prefer one out of two different actions each producing rewarding values. This is harder than the problem used by Frank who simply *assigned* all dopamine responses in his experiments. In contrast, the current system will have to *learn* to produce useful dopamine signals all by itself.

5.1.2 Walking Through the Problem

One run of the system consists of forty thousand time steps. Each time step, the limbic system receives a string of four zeros and once every forty time steps the third input entry will be set to .5. Whenever this trigger value shows up, its magnitude alone is enough to have the PMC suggest an action. The Go and NoGo neurons will collectively come up with an inhibition value for each action and the GPi and GPe mechanism will subtract these values. If the bottom-up thalamus support is strong enough to force PMC activations above a threshold of .1, the winner-takes-all rule settles the competition.

As explained above, performing a certain action means playing a corresponding lever. This lever will now draw a number out of its associated normal distribution and that number becomes the reward. The normal distribution of the first lever has a mean of .3 with standard deviation .1 and the normal distribution of the second lever has a mean of 0.7 with also a standard deviation of .1.

The critic will now take the reward and compute a dopamine signal. It too sees the trigger on the input layer and is learning to associate it with incoming reward. Through exploration, the system will try both actions repeatedly and over time the striosomes will come to expect a reward value somewhere between the scaled mean values of the two levers. The most beneficial lever now causes slight dopamine increases while the least beneficial lever causes slight suppressions. These signals make the actor learn to prefer the better of the two levers and since the impact of exploration decreases over time, the better lever starts to be chosen more often than not. Finally, when all exploration has stopped, the prediction of the striosomes will converge toward the scaled mean value of the best lever.

Exploration is simulated by adding random values to the outputs of the PMC. The system forces the random numbers to be added a couple of time steps after the occurrence of a trigger value. Such a mechanism is of course very artificial, but remember that the study does not aim to come up with a biologically plausible exploration system. Over 39500 time steps, the magnitude of the random values will decrease to zero. However, before that time, the PMC will have strengthened its associations with the trigger value and the Go neurons will lift inhibition from the thalamus whenever the PMC is suggesting the correct action. Together this results in a correct response.

5.2 Learning the Task

The model is capable of learning the two-armed bandit task. Out of ten runs the model learned to prefer the correct action nine times. The only run resulting in a preference for the incorrect action was clearly on its way to change its mind. This could be determined from the increases and decreases of Go and NoGo signals. In this case, the model started out with random weights strongly biasing the incorrect action for selection by the PMC. The growing difference between incorrect action Go and NoGo signals should have overcome this preference in order to inhibit the biased PMC signal enough to have the right action come out on top. Since this difference was actually growing, it seems that, with more learning time, the system would have solved the problem.

In figure **5.1** the left picture shows the times the incorrect action I was chosen in a run of forty thousand time steps. The right picture shows the same for the correct action II. The winner takes all rule has assigned values to both actions of either one or zero, resulting in the pictures below. Because of strong exploration values, both actions are chosen more or less random for a long time. Only in the end of a run the difference becomes very obvious. Here, exploration values are very low and eventually do not occur anymore.

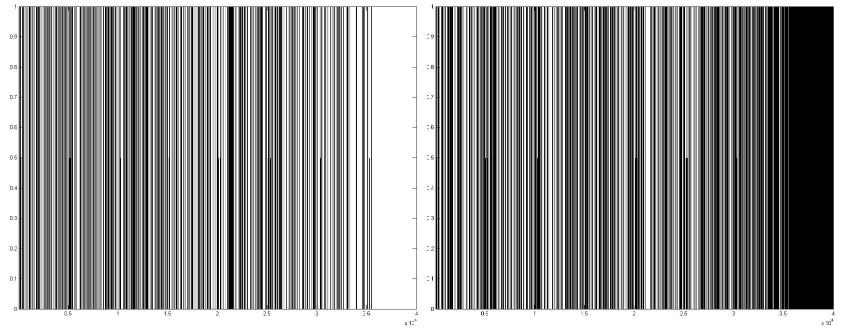

Figure 5.1: Left: incorrect action responses during a run of forty thousand time steps. Right: correct action responses in the same run

Figure **5.2** shows the same result in proportions taken over five hundred time steps. In the beginning both actions are chosen about half of time. Eventually, the algorithm only chooses the correct action II.

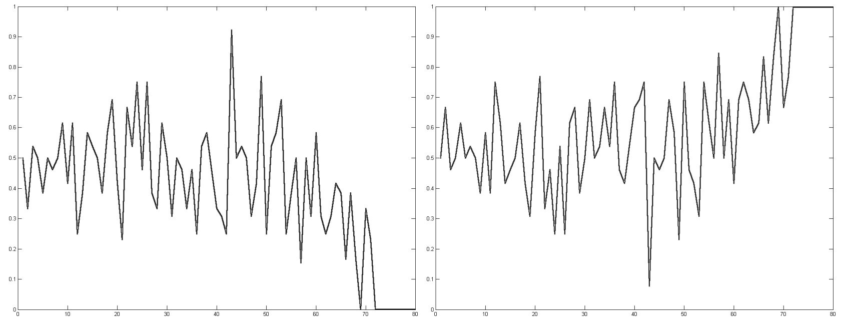

Figure 5.2: Left: proportions of choosing the incorrect action grouped in sequences of five hundred time steps. Right: same for correct action choices

5.3 The Critic

What is the behavior of the critic in this task? Can the biological behaviors of dopamine be reproduced? This section will treat these behaviors one by one, comparing the model to the brain.

5.3.1 Transfer of Dopamine Outburst

When striosomes have not learned to give off high inhibition values as a response to a CS, primary rewards coming in through the PPTN will cause the SNc to fire phasically, resulting in a dopamine outburst (figure **5.3**). Such an outburst will make the ventral striatum strengthen its connections to the current pattern in the limbic cortex and after sufficient co-occurrences of the pattern and primary reward the ventral striatum will start to excite the SNc all by itself as soon as the pattern is seen. It will do this by exciting the PPTN which will change the signal through accommodation, creating the phasic character of the dopamine response. Together this results in another dopamine outburst at the onset of a CS (see figure **5.4**).

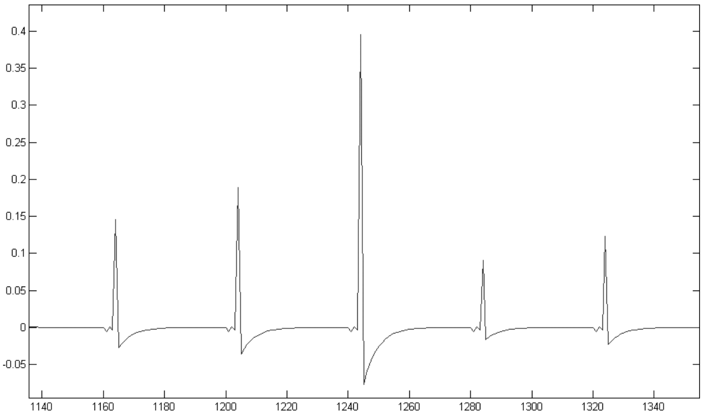

Figure 5.3: Unpredicted rewards cause dopamine outbursts. Rewards are drawn from an action normal distribution resulting in outbursts of different magnitudes. As can be seen, rewards come in cycles of forty time steps. The little dopamine peak just before the reward peak is a start of the system learning a dopamine response on the occurrence of a CS (trigger).

5.3.2 Inhibition of Expected Reward

Unpredicted outbursts of dopamine will cause the striosomes to learn a timed association with the input pattern on the limbic cortex and over time this association becomes strong enough to counter the dopamine outburst all together.

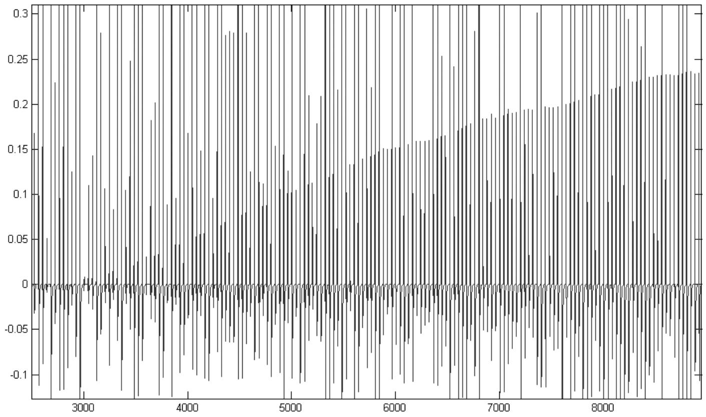

Figure 5.4: Ventral striatal neurons increase their signal on CS appearance

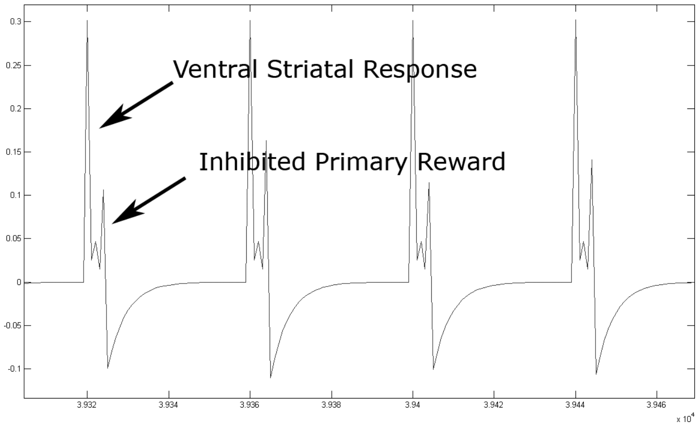

Figure 5.5: Striosomes are learning to cancel dopamine outburst caused by incoming primary reward

This mechanism also explains the suppressions of tonic dopamine level when predicted rewards fail to come in. In that case, the PPTN fails to fire at time of predicted reward and no SNc excitation occurs. However, the striosomes will still fire their timed inhibition signal, causing the dopamine level to drop below baseline.

5.3.3 Altered Timing of Predicted Reward

From chapter 3 we know that primary rewards coming in sooner than predicted will provoke a dopamine outburst without a subsequent suppression. In the model, striosomes and the input mechanism discussed in the previous section take care of this together. Since the striosomes have learned a timed response, they can not inhibit the reward when it comes in early. The reward is therefore free to cause a dopamine outburst. The occurrence of

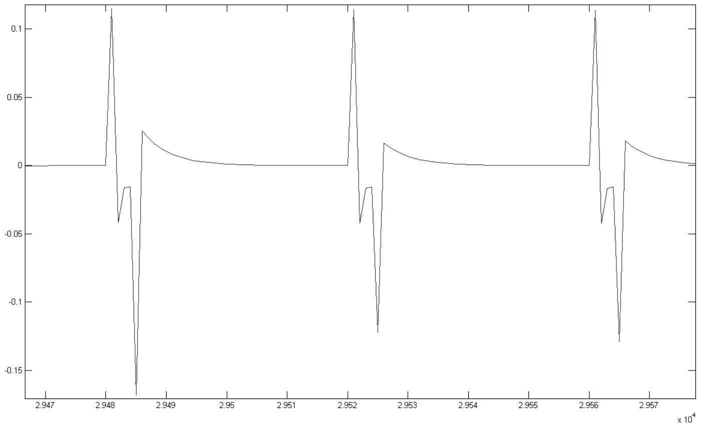

Figure 5.6: Incoming CS makes the ventral striatal cells invoke a dopamine response after which a failure of incoming predicted reward causes striosomes to suppress dopamine below baseline

the reward will then cancel the activation of limbic neurons, removing the association the striosomes are basing their firing on. And so, no further inhibition is seen. Furthermore, after a little training the striosomes will learn to adapt to the new situation, since the earlier reward still causes a dopamine outburst which will be used for training.

Rewards coming in later than expected cause the model to suppress tonic dopamine signal at the time of expected reward. When the actual reward finally does come in, no inhibition from striosomes occurs and an outburst of dopamine is provoked. When the new reward delay is systematic, the model is able to adapt to the new situation. This is in complete agreement with the biological data seen in chapter 3.

5.3.4 Learning CS Chains

The system is able to first associate a CS with a dopamine outburst and then associate the CS with an earlier occurring CS. This capability is quite important because it allows for the possibility of creating long CS chains. A creature could thus learn entire sequences of behavior leading up to a reward.

Figure **5.7** shows what happens when the model is allowed to learn the two-armed bandit task, after which an earlier CS gets introduced to show up five time steps before the original CS. The dopamine outburst starts to quickly shift toward the earlier CS and over time, the original CS induced outburst gets inhibited by the striosomes. As stated in chapter 3, the effect is known from biology, where self-induced dopamine outbursts always shift to the earliest occurring CS[60].

Theoretically it should be possible to have the actor learn to perform some action on basis of the first outburst, then perform some other action on basis of the second outburst earning its reward afterward. The chain of actions could be expanded by introducing an even earlier CS. Learning would we based on dopamine outbursts entirely produced by the system itself.

As of yet, no attempt has been made to implement a task requiring such a chain of actions and experimentation with action sequences is left for future research. However, the pole-balancing task described in chapter 2 would make a suitable benchmark.

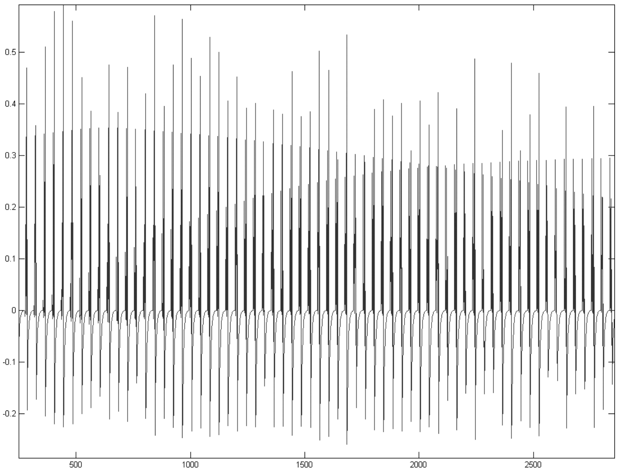

Figure 5.7: The system learns a dopamine response to a new CS introduced five time steps before the original CS. The figure clearly shows the increase of dopamine responses to the earlier CS and the decrease of already learned responses to the original CS. Time variables were reset and so the horizontal axes starts at zero again

5.3.5 Generalization

The only biological characteristics of dopamine the model can not accurately reproduce are caused by stimuli similar but not equal to a learned CS. From chapter 3 we know that when this happens, dopamine levels show a mild suppression right after CS onset and no suppression is seen afterward. It is as if animals immediately spot their mistake and somehow cancel the activation pattern in the limbic cortex. This effect might well be caused by systems outside the scope of this study and there does not seem to be any existing model able to reproduce it. For now, the effect is noted as material for future research.

5.4 The Actor

Unlike the critic, the actor has no real biological measurement data to reproduce. However, illustrating the performance of its elements clarifies the workings of the model. Therefore, the learning behavior of the actors' adaptive parts is treated below.

5.4.1 Go/NoGo Responses

The GPi will only inhibit an action represented in the thalamus when the NoGo signal of that action is higher than its Go signal. Without this inhibiting influence the PMC is free to choose whatever action it likes best. After training we expect to see no inhibition for the correct action and enough inhibition for the incorrect action to guarantee the correct action coming out on top.

Figure **5.8** shows the learning of Go and NoGo responses for the incorrect action. As can be seen, the responses first adapt themselves in the wrong direction only to correct this later on and moving into the right direction thereafter. Given the two-armed bandit task this behavior is not surprising. In the early stages of training, the incorrect action produces a reward that is relatively small but unexpected, causing a moderate outburst of dopamine. After a while, the striosomes have come to expect a larger reward since performing the correct action usually produces such a larger reward. Now the striosomes start causing dopamine *suppressions* when an incorrect action is performed and the Go and NoGo layers change their direction of learning.

The Go and NoGo layers of the correct action do not need to change learning direction. For these layers it is quite easy to learn high Go responses and low NoGo responses (see figure **5.9**).

5.4.2 The PMC

Since the PMC uses a simple Hebbian learning rule to associate its outcome with the state representation on the input layer, at some point its responses become strong enough to take action all by itself. When this happens, a preferred action for a certain input state will be triggered without the need of a Go/NoGo mechanism. All that is needed is enough co-occurrences of stimulus and response. After learning the appropriate Go/NoGo signals, learning such SR mappings becomes trivial. The stimulus and response only occur together and so it is just a matter of time for the PMC to learn how to take action on its own. This reflects the idea that the basal ganglia play a merely modulatory role in action selection.

5.5 Task Reversal

A preliminary result was found when switching the normal distributions after the task was already learned correctly. Since the system has learned to prefer action II it will choose this action every time. However, this action now results in dopamine suppression since the striosomes have come to predict the higher reward, but a lower reward is actually coming in. These suppressions cause Go responses of the previously correct action to come back down while NoGo responses are starting to rise. In this experiment the model is given a constant low value of exploration and at some point the exploratory random numbers added

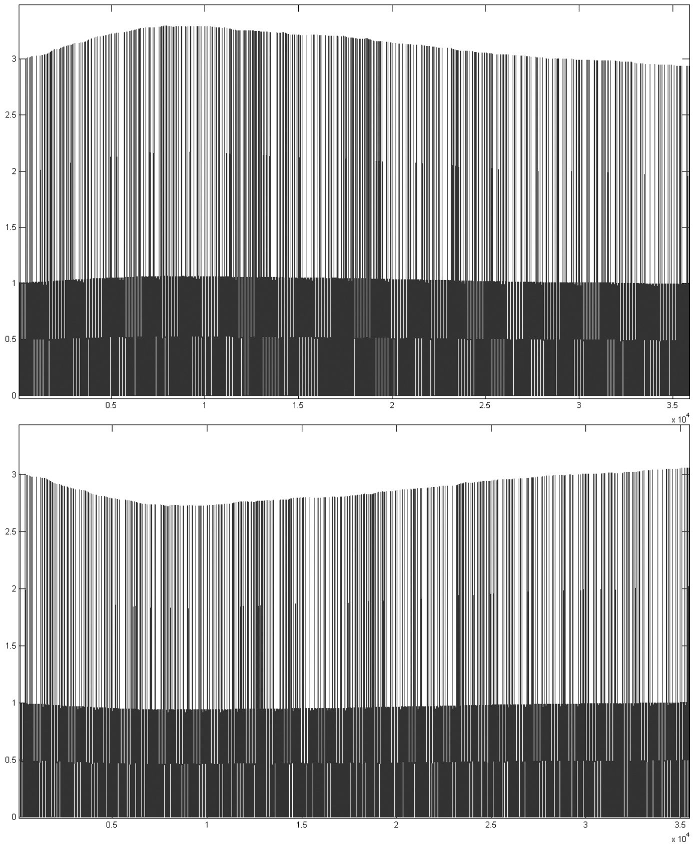

Figure 5.8: Incorrect action Go (upper image) and NoGo (lower image) response learning over 35000 time steps. The large peaks reflect responses to incoming CS with the PMC selecting the incorrect action. The peaks in between are caused by association with the CS on the input layer when the correct action was chosen.

to the PMC become important again. When this happens the system starts choosing both actions now and again and sometimes learns to reverse its preference.

Too few runs have shown complete reversal to accept this theory as fact. However, all runs have shown the Go and NoGo responses changing direction as described above. The runs not showing full reversal are believed to have been suffering from the striosomes adapting too quickly to the new situation. After reversal, the striosomes learn to adjust their inhibition until dopamine level stays at base at time of reward. After they have learned to do this, no suppression of dopamine is seen anymore and the system stops learning all together.

This latter theory would fit nicely with some theories about side-effects of dopaminergic medication in Parkinson patients. If brain areas, relatively unharmed by the disease, are influenced by such medication they will suffer an overdose of dopamine. In our model the

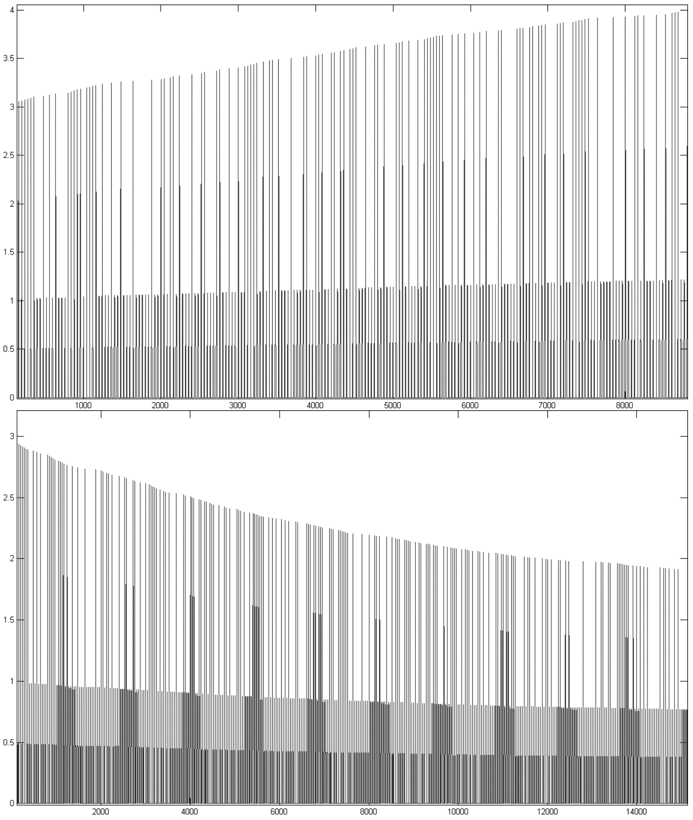

Figure 5.9: Go and NoGo response learning for correct action

striosomes would have a hard time producing dopamine suppressions in overdosed areas. This leads to impairments in tasks requiring negative feedback. As was seen in chapter 3, such side-effects are well known in literature.

Chapter 6

Conclusion

What can be concluded from the results of the previous chapter? How does this relate to earlier models? In this chapter we return to the central question that was posed in chapter 1. Next, the study will be related to traditional actor/critic methods and previous studies concerning biologically plausible dopamine based learning. Finally, some suggestions are made for future work.

6.1 Answering the Central Question

As described in chapter 1, the study attempted to find an answer to the following question: can a computer model be constructed, honoring neurobiological knowledge about relevant brain areas and their interconnections, that learns to produce a dopamine-like error signal and is able to use this signal for solving learning problems?

This question consists of three subquestions. How close does the model copy the workings of relevant brain areas? To what extent is the model produced error signal similar to an actual dopamine error signal? Is the model capable of learning? Below, each question will be treated one at a time.

6.1.1 Model Structures vs. Brain Areas

All structures in the model where built using neurobiological knowledge of brain areas involved. This knowledge was treated in chapter 3 and care was taken not to deviate from it. However, several choices were made without direct support from neurobiology. In particular, equations and constants were often chosen for convenience. Fortunately, almost all equations could be based on well known neurocomputational formulas and care was taken to assign sensible values to constants. Therefore, it should lie within the capabilities of biological nervous systems to produce the same results.

This last statement holds for all but the exploration mechanism. The exploration mechanism is completely artificial. Therefore, all results are biologically plausible assuming the presence of a sufficiently functioning exploration mechanism in the brain. The nature and implementation of such a mechanism is left for future research.

6.1.2 Similarity Between Error Signal and Dopamine

The dopamine response learned by the model is in many ways similar to the dopamine response of natural systems. The system shows dopamine outbursts to novel rewarding input and shifts responses to the first incoming CS. After training, the system suppresses dopamine responses when expected rewards fail to come in and deals with early and delayed incoming rewards in a manner similar to biological dopamine systems.

The only behavior the model does not accurately reproduce occurs when the brain deals with input similar but not equal to the CS information it was trained on. As was treated in chapter 3, a dopamine response is seen when such input comes in, but is followed quickly by a slight suppression after which no responses of any kind are seen anymore. As is the case with the current model, no previous model seems to be able to reproduce it and the effect is left for future research.

6.1.3 Learning Capabilities

As was shown in the previous chapter, the model is capable of learning the two-armed bandit task. For the purpose of this study that result is sufficient. However, no claims can be made about the kinds of problems or the complexity of problems the model is able to learn. For instance, it would be interesting to find out to what extend the model is able to chain CS inputs in order to learn sequences of actions ending up in high reward values. Can the model learn the pole balancing task as described in chapter 2? Once again, this is left for future research.

6.1.4 Conclusion

And so, assuming the presence of a sufficiently functioning exploration mechanism in the brain, the model resembles the relevant brain areas close enough to have some explanatory force. At the very least it shows the possibility of an actor/critic like learning mechanism in the brain that uses dopamine responses as error signals.

6.2 Relation to Actor/Critic Methods

In the end, the model shows many similarities with actor/critic methods known from the field of reinforcement learning. The actual structures collectively labeled *critic* or *actor* were of course categorized by hand and it is completely possible that a better categorization will be made in the future. However, as in a traditional critic, a set of structures collectively learns state values and uses them to produce error signals on which it learns itself. Furthermore, as in a traditional actor, a set of structures uses the error signal to learn SR associations and performs actions resulting in reward information. Finally, learning to produce error signals and learning SR mappings occurs at the same time and the order of computation is exactly like the order used in traditional actor/critic models.

Somewhat different is the models' usage of time. Many neurons need a couple of time steps to reset themselves due to accommodation or lingering inputs. Also, to be biologically plausible we can not have rewards following actions in the exact same time step. Of course, traditional reinforcement learning methods often do use immediate rewards and learn in the same code cycle. However, in the present study this is neither possible nor appropriate. Therefore, the model uses delays of several time steps between trials and special methods had to be constructed to keep information available for learning.

6.3 Relation to Previous Studies

Previous studies have explored computational models able to reproduce the dopamine response in a biologically plausible way. Other studies, biologically plausible or not, have shown the possibility of error signal based learning. Finally, cognitive science has produced a number of theories arguing the dopamine response to play the functional role of an error signal.

Interestingly, it seems no study has actually produced a computational model able to learn SR associations using a truly dopamine-like error signal. In fact, the behaviors of error signals used by most action learning studies are generally very simple and really not that similar to the behaviors of dopamine responses at all. This introduces the following question: are dopamine responses suitable error signals? Previous studies have left this topic wide open.

By building a computational model using an error signal that *is* similar to natural dopamine responses (in shape, timing and development) and showing it to be capable of learning, the current study has contributed to the believe that modern theories about dopamine are on the right track. Moreover, by basing its design on two other computational models (Brown et al[11], Frank[30]) a thorough connection to previous work was established and a computer model was produced that honors neurobiological knowledge about the brain areas involved.

6.4 Future Research

The present model is very fertile for future exploration. Aside from the suggestions made earlier in this chapter, many aspects of dopamine and its interplay with brain structures could be investigated. For instance, even though Frank[30] has experimented with the effects of dopamine deficiency and overdose on actor structures, no investigation has been clarified the effects on the critic structures. In the present model, the dopamine level very much influences critic learning and it can only be expected that dopamine overdose, as a result of medication, or deficiency, as a result of e.g. Parkinsons' disease, will have some effect on its performance. Using the present model, experiments can be set up and predictions can be made about the effects of unnatural dopamine levels on the brain structures of the critic.

The bistable nature of striatal neurons has not been incorporated in the present model. Doing so could end up in interesting Go layer behavior since the presence of dopamine will enhance contrast in the outputs of bistable neurons. In his model, Frank[30] has already experimented with bistable neurons, but his method of learning was somewhat different and the effects on the present model are therefore unclear.

Bibliography

[1] Levey AI, Hersch SM, Rye DB, Sunahara RK, Niznik HB, Kitt CA, Price DL, Maggio R, Brann MR, and Ciliax BJ. Localization of d1 and d2 dopamine receptors in brain with subtype-specific antibodies. *Proc Natl Acad Sci U S A.*, 90(19):8861–8865, Oct 1993.

[2] G. E. Alexander, M. R. DeLong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci*, 9:357–381, 1986.

[3] Crutcher MD Alexander GE. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.*, 13(7):266–71, Jul 1990.

[4] Gotham AM, Brown RG, and Marsden CD. 'frontal' cognitive function in patients with parkinson's disease 'on' and 'off' levodopa. *Brain.*, 111:299–321, Apr 1988.

[5] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. pages 81–93, 1990.

[6] G.A. Barto. pages 215–232, 1995.

[7] B. Berger, S. Trottier, C. Verney, P. Gaspar, and C. Alvarez. Regional and laminar distribution of the dopamine and serotonin innervation in the macaque cerebral cortex: A radioautographic study. *The Journal of Comparative Neurology*, 273(1):99–119, 1988.

[8] C Bergson, L Mrzljak, JF Smiley, M Pappy, R Levenson, and PS Goldman-Rakic. Regional, cellular, and subcellular variations in the distribution of d1 and d5 dopamine receptors in primate brain. *J. Neurosci.*, 15:7821–7836, 1995.

[9] Gregory S. Berns and Terrence J. Sejnowski. A computational model of how the basal ganglia produce sequences. *J. Cognitive Neuroscience*, 10(1):108–121, 1998.

[10] W. Birkmayer and O. Hornykiewicz. The effect of 3,4-dihydroxyphenylaline (=dopa) on parkinsonian akinesia. *Wien.Klin.Wochenschr.*, 73:787–8, 1961.

[11] J. Brown, D. Bullock, and S. Grossberg. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J Neurosci*, 19(23):10502–10511, December 1999.

[12] Joshua W. Brown, Daniel Bullock, and Stephen Grossberg. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Netw.*, 17(4):471–510, 2004.

[13] Cepeda C, Buchwald NA, and Levine MS. Neuromodulatory actions of dopamine in the neostriatum are dependent upon the excitatory amino acid receptor subtypes activated. *Proc Natl Acad Sci U S A.*, 90(20):9576–80, Oct 1993.

[14] Neil R. Carlson. *Foundations of physiological psychology*. Pearson/Allyn and Bacon, Boston ; Munich [u.a.], 6. ed., international ed. edition, 2005.

[15] C. Cepeda, S. H. Chandler, L. W. Shumate, and M. S. Levine. Persistent na+ conductance in medium-sized neostriatal neurons: characterization using infrared videomicroscopy and whole cell patch-clamp recordings. *J Neurophysiol*, 74:1343–1348, 1995.

[16] C Cepeda, CS Colwell, JN Itri, SH Chandler, and MS Levine. Dopaminergic modulation of nmda-induced whole cell currents in neostriatal neurons in slices: contribution of calcium conductances. *J Neurophysiol*, 79:82–94, 1998.

[17] Jonathan D. Cohen and David Servan-Schreiber. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1):45–77, Jan 1992.

[18] R. Cools, R.A. Barker, B.J. Sahakian, and T.W. Robbins. Enhanced or impaired cognitive function in parkinsons disease as a function of dopaminergic medication and task demands. *Cerebral Cortex*, 11:1136–1143(8), December 2001.

[19] Gerfen C.R. The neostriatal mosaic - multiple levels of compartmental organization in the basal ganglia. *Annu Rev Neurosci*, 15:285–320, 1992.

[20] Gerfen CR. Molecular effects of dopamine on striatal-projection pathways. *Trends Neurosci.*, 23:S64–S70, Oct 2000.

[21] Gerfen C.R. and Wilson C.J. The basal ganglia. 12:371–468, 1996.

[22] Gerfen C.R., Keefe K.A., Bloch B., le Moine C., Surmeier D.J., Reiner A., Levine M.S., and Ariano M.A. Neostriatal dopamine receptors. *Trends neurosci.*, 17(1):2–5, 1994.

[23] Gerfen CR, Engber TM, Mahan LC, Susel Z, Chase TN, Monsma FJ, and Sibley DR. D1 and d2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, 250(4986):1429–32, Dec 1990.

[24] Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.

[25] J.F. Dormont, H. Condé, and D. Farin. The role of the pedunculopontine tegmental nucleus in relation to conditioned motor performance in the cat. *Experimental Brain Research*, 121(4):401–410, 1998.

[26] Ince E, Ciliax BJ, and Levey AI. Differential expression of d1 and d2 dopamine and m4 muscarinic acetylcholine receptor proteins in identified striatonigral neurons. *Synapse.*, 27(4):357–66, Dec 1997.

[27] H. Ehringer and O. Hornykiewicz. Distribution of noradrenaline and dopamine (3-hydroxytyramine) in human brain: Their behaviour in extrapyramidal system diseases. *Klin. Wochenschr.*, 38:1236–9, 1960.

[28] Rolls ET, Thorpe SJ, Boytim M, Szabo I, and Perrett DI. Responses of striatal neurons in the behaving monkey. 3. effects of iontophoretically applied dopamine on normal responsiveness. *Neuroscience.*, 12(4):1201–12, Aug 1984.

[29] Middleton F.A. and Strick P.L. Basal ganglia output and cognition: Evidence from anatomical, behavioral, and clinical studies. *Brain and Cognition*, 42(2):183–200, March 2000.

[30] Michael J. Frank. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *J. Cogn. Neurosci.*, 17(1):51–72, January 2005.

[31] Chevalier G and Deniau JM. Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, 13:277–280, 1990.

[32] Alexander GE, Crutcher MD, and DeLong MR. Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Prog Brain Res.*, 85:119–46, 1990.

[33] P. M. Groves, M. Garcia-Munoz, J. C. Linder, M. S. Manley, M. E. Martone, and S. J. Young. Elements of the intrinsic organization and information processsing in the neostriatum. pages 51–96, 1995.

[34] K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. ii. analysis and simulation of behaviour. *Biol Cybern*, 84(6):411–423, June 2001.

[35] Salvador Hernndez-Lpez, Jos Bargas, D. James Surmeier, Arturo Reyes, and Elvira Galarraga. D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an l-type ca2+ conductance. *J Neurophysiol*, 17(9):3334–3342, May 1997.

[36] S. M. Hersch, B. J. Ciliax, C. A. Gutekunst, H. D. Rees, C. J. Heilman, K. K. Yung, J. P. Bolam, E. Ince, H. Yi, and A. I. Levey. Electron microscopic analysis of d1 and d2 dopamine receptor proteins in the dorsal striatum and their synaptic relationships with motor corticostriatal afferents. *J Neurosci*, 15(7 Pt 2):5222–5237, July 1995.

[37] Okihide Hikosaka. Role of basal ganglia in initiation of voluntary movements. pages 153–167, 1989.

[38] Frank M. J., Loughry B., and O'Reilly R. C. Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, 1:137–160(24), 1 June 2001.

[39] Kojima J., Yamaji Y., Matsumura M., Nambu A., Inase M., Tokuno H., Takada M., and Imai H. Excitotoxic lesions of the pedunculopontine tegmental nucleus produce contralateral hemiparkinsonism in the monkey. *Neuroscience letters*, 226(2):111–114, 1997.

[40] Cohen J.D., Braver T.S., and Brown J.W. Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, 12:223–229(7), 1 April 2002.

[41] Hay J.F., Moscovitch M., and Levine B. Dissociating habit and recollection: evidence from parkinson's disease, amnesia and focal lesion patients. *Neuropsychologia*, 40:1324–1334(11), 2002.

[42] Brog JS, Salyapongse A, Deutch AY, and Zahm DS. The patterns of afferent innervation of the core and shell in the "accumbens" part of the rat ventral striatum: immunohistochemical detection of retrogradely transported fluoro-gold. *J Comp Neurol.*, 338(2):255–78, Dec 1993.

[43] Mink JW. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog Neurobiol.*, 50(4):381–425, Nov 1996.

[44] Gurney K, Prescott TJ, and Redgrave P. A computational model of action selection in the basal ganglia. i. a new functional

anatomy. *Biol Cybern.*, 84(6):401–10, Jun 2001.

[45] Keefe K.A. and Gerfen C.R. D1-d2 dopamine receptor synergy in striatum: effects of intrastriatal infusions of dopamine agonists and antagonists on immediate early gene expression. *Neuroscience*, 66:903–913(11), 1995.

[46] J. Kalat.

[47] Y. Kawaguchi, C. J. Wilson, and P. C. Emson. Intracellular recording of identified neostriatal patch and matrix spiny cells in a slice preparation preserving cortical inputs. *J Neurophysiol*, 62:1052–1068, 1989.

[48] A. Harry Klopf. Brain function and adaptive systems: A heterostatic theory. *Technical Report AFCRL72 -016*, Mar. 1972.

[49] B. J. Knowlton, J. A. Mangels, and L. R. Squire. A neostriatal habit learning system in humans. *Science*, 273(5280):1399–1402, September 1996.

[50] T. Ljungberg, P. Apicella, and W. Schultz. Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol*, 67(1):145–163, 1992.

[51] Packard MG and Knowlton BJ. Learning and memory functions of the basal ganglia. *Annu Rev Neurosci.*, 25:563–93, March 2002.

[52] G.J. Mogenson and M. Wu. Subpallidal projections to the mesencephalic locomotor region investigated with a combination of behavioral and electrophysiological recording techniques. *Brain Res Bull.*, 16Mar(3):383–90, 1986.

[53] Hikosaka O. Neural systems for control of voluntary action–a hypothesis. *Adv Biophys.*, 35:81–102, 1998.

[54] Andrés Pérez-Uribe. Using a time-delay actor-critic neural architecture with dopamine-like reinforcement signal for learning in autonomous robots. pages 522–533, 2001.

[55] Swainson R., Rogers R.D., Sahakian B.J., Summers B.A., Polkey C.E., and Robbins T.W. Probabilistic learning and reversal deficits in patients with parkinson's disease or frontal or temporal lobe lesions: possible adverse effects of dopaminergic medication. *Neuropsychologia*, 38:596–612(17), 1 May 2000.

[56] R. D. Rogers, B. J. Sahakian, J. R. Hodges, C. E. Polkey, C. Kennard, and T. W. Robbins. Dissociating executive mechanisms of task control following frontal lobe damage and parkinsons disease. *Brain*, 121:815–842, 1998.

[57] Ranulfo Romo and Wolfram Schultz. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, 63(3):592–606, 1990.

[58] Funahashi S, Bruce CJ, and Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol.*, 61(2):331–49, Feb 1989.

[59] A. L. Samuel. Some studies in machine learning using the game of checkers. pages 71–105, 1995.

[60] W. Schultz. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.*, 13(3):900–913, March 1993.

[61] W. Schultz. Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80(1):1–27, July 1998.

[62] Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 361–368. The MIT Press, 1995.

[63] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[64] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44, August 1988.

[65] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, March 1998.

[66] Woodward T.S., Bub D.N., and Hunter M.A. Task switching deficits associated with parkinson's disease reflect depleted attentional resources. *Neuropsychologia*, 40:1948–1955(8), 2002.

[67] John N. Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94, 1996.

[68] S. Mark Williams and Patricia S. Goldman-Rakic. Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody. *Cerebral Cortex*, 3:199–222, 1993.

[69] Ian H. Witten. An adaptive optimal controller for discrete-time markov environments. *Information and Control*, 34(4):286–295, August 1977.

[70] Zhen Yan, Wen-Jie Song, and D. James Surmeier. D2 dopamine receptors reduce n-type ca2+ currents in rat neostriatal cholinergic interneurons through a membrane-delimited, protein-kinase-c-insensitive pathway. *J Neurophysiol*, 77:1003–1015, 1997.

[71] C.R. Yang and G.J. Mogenson. Hippocampal signal transmission to the pedunculopontine nucleus and its regulation by dopamine d2 receptors in the nucleus accumbens: an electrophysiological and behavioural study. *Neuroscience*, 23(3):1041–55, 1987.