Biological Models of Reinforcement Learning

Julien Vitay, Jérémy Fix, Frederik Beuth, Henning Schroll, Fred H Hamker

Biological Models of Reinforcement Learning

Julien Vitay, Jérémy Fix, Frederik Beuth, Henning Schroll, Fred H. Hamker

This review focuses on biological issues of reinforcement learning. Since the influential discovery of W. Schultz of an analogy between the reward prediction error signal of the temporal difference algorithm and the firing pattern of some dopaminergic neurons in the midbrain during classical conditioning, biological models have emerged that use computational reinforcement learning concepts to explain adaptative behavior. In particular, the basal ganglia has been proposed to implement among other things reinforcement learning for action selection, motor control or working memory. We discuss to which extent the analogy between the temporal difference algorithm and the firing of dopamine cells can be considered as valid. Our review then focuses on the basal ganglia, their anatomy and key computational properties as demonstrated by three recent, influential models.

1 Introduction

Some recent progress in cognitive neuroscience relies on the understanding that adaptive behavior is not an isolated process in the brain, but rather an emergent property of the interaction between brain, body and the environment [6]. Embodiment, epigenetic development and neuroethology, among others, are neuroscientific research areas that can take huge benefits from the interaction with artificial intelligence, especially involving robotics. The subfield of machine learning called *reinforcement learning* (RL), with its classical agent/environment distinction [28], can be an useful framework to understand not only how the actions of a subject can be learned to maximize future rewards in a given cognitive task, but also how this dependency on reward expectation and motivation can guide the formation and updating of brain representations.

Classical conditioning (like in the famous Pavlov experiment [22]) has been an example of such an interplay between behavioral sciences and computational modeling, particularly with respect to the Rescorla-Wagner model [25]. Increased attention to RL in neuroscience has been given since the seminal studies of Wolfram Schultz [26] who observed that dopaminergic (DA) neurons in the animal's midbrain show similar patterns as the error signal in the temporal-difference (TD) algorithm. The integration of these DA neurons into a functional pathway lead to various computational models of basal ganglia (BG) that mimic the classical actor/critic architecture in order to explain BG functioning in various reward-dependent tasks such as action-selection, motor control or working memory. The idea of this short review is to present what these biological models of BG have captured from the RL paradigm, particularly the TD algorithm, and how the additional biological constraints can potentially influence the development of new RL algorithms able to solve real-world cognitive tasks. An excellent complementary review can be found in [9].

We will first present in section 2 the link between the firing patterns of DA neurons during Pavlovian conditioning and the error signal of the TD algorithm. In section 3, we will present the BG and how its known functional anatomy can be compared with the classical actor/critic architecture. In section 4, we will describe three different recent computational models of BG that rely on different assumptions about basal ganglia functioning. This selection is however highly non-exhaustive, a review about relatively older BG models can be found in [14].

2 Temporal-difference and dopamine

2.1 Classical conditioning

Reinforcement learning closely relates to classical conditioning, which is a simple form of associative learning observed in animals and humans: Given an inborn association between an unconditioned stimulus (US) and an animal's unconditioned response to that stimulus (UR), the animal can be trained to show a response to previously neutral stimuli. The procedure works as follows: A neutral stimulus, called conditioned stimulus (CS), is repeatedly presented before a particular US, thus being a temporal predictor of US presentation for the animal. After a couple of pairings, presentation of the CS leads to a conditioned response (CR) which is usually very similar (but not necessarily identical) to the UR. A classical example of this type of conditioning is the Pavlov experiment: A dog naturally salivates (UR) at the sight of food (US). Each time an experimenter rings a bell (CS), he delivers food after a certain delay. After a couple of associative pairings, the bell will produce salivation (CR) by itself.

2.2 Temporal-difference model

A core idea for RL introduced by Sutton and Barto [28] was the temporal-difference (TD) algorithm for evaluating the value function V^π associated to a policy π. This bootstrapping method uses directly transitions within a Markov decision process to modify its evaluations based on the following error signal:

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

This signal computes the difference between the expected reward after a transition (the reward r_{t+1} actually received after the transition plus the value V of the next state s_{t+1}) and the value of the current state s_t. It can be used to update the values of the states (when δ_t is positive, the real value of the state is higher than its current estimation; when negative it is lower) and also to find the optimal policy (when δ_t is positive, the action that lead to this transition is considered to lead to a "good surprise" and

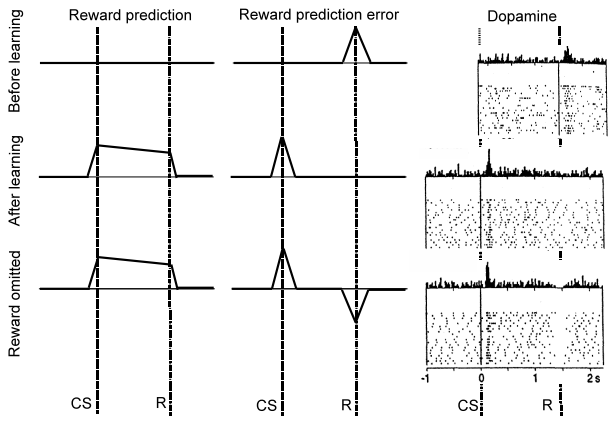

Figure 1: Reward prediction error of the extended TD model and dopamine firing. From [26]. Reprinted with permission from AAAS.

should be preferentially chosen the next time; when negative it should be avoided). This led to the development of the "actor-critic" architecture, where the critic estimates the values of the states and computes δ_t, and the actor learns to perform the optimal action in each state.

Suri and Schultz [27] extended this TD model within a more biologically inspired actor/critic architecture to simulate the response of DA neurons in the context of a spatial delayed response task. The agent receives one of two temporal stimuli (e_1, e_2) as inputs (CS) and has to select one of the two possible actions (a_1, a_2). If it selects the correct one (e.g. $e_1 \rightarrow a_1$, $e_2 \rightarrow a_2$), reward is given after a short delay. The actor is implemented by a direct mapping between the stimuli and the actions, through connection weights which are learned according to a Hebbian rule modulated by the δ_t signal. The authors used a temporal representation of each event ("complete serial compound stimulus") to bridge the delay between the stimulus, actions and rewards. Thus, each stimulus is represented by a series of neurons, each neuron progressively responding through time after the onset of the stimulus: The first neuron is activated at stimulus onset, the second one 100 ms later, and so on. This representation allows to keep track of the appearance of a stimulus and gives information about the time elapsed since its onset. Each of these neurons is considered as a state of the system, even if the transition between them is fixed.

During learning, the delivery of reward progressively increases the value of the neurons that have been active before, accompanied by a positive δ_t, at the condition that the delay between stimulus onset and reward stays constant. Step by step, all the states between reward and stimulus onset get a high value (they all predict oncoming reward). After learning, δ_t will get positive at stimulus onset but not when reward is given, whereas initially, at the beginning of learning, it was only positive at reward delivery. If reward delivery is omitted after learning, δ_t will become negative, because r_{t+1} was expected at this particular time.

The first two columns of Figure 1 show respectively the reward prediction and reward prediction error δ_t of this extended TD model before and after conditioning, as well as in the case where reward is omitted. Reward prediction represents the values of the states (time here is the state space). In agreement with its most prominent predecessor for conditioning (Rescorla-Wagner model [25]) this extended TD model also explains the observed phenomena of extinction, blocking and conditioned inhibition:

– *Extinction:* A CS which was previously associated to a US is repeatedly presented alone (without a subsequent US) and loses its predictive value [22]. The associative strength of this CS then declines asymptotically to zero.

– *Blocking:* One CS (CS_1) has already been learned to predict the upcoming of a US. When a second CS is then simultaneously presented with CS_1, this second CS does not acquire associative strength [15]. The extended TD model can account for this phenomenon by assuming that the total amount of associability is limited.

– *Conditioned inhibition:* Two kinds of trials alternate. Either a particular CS (called CS_1) is constantly followed by a US or a combination of CS_1 and a second CS (named CS_2) is never followed by reward [22]. In this situation CS_1 is associated with the upcoming of the US, whereas CS_2 is associated with the US not being presented, even if CS_1 is present. The extended TD model displays exactly this behavior: CS_1 gains positive associative strength while CS_2 gains negative strength.

2.3 Dopamine firing in the midbrain

Dopamine (DA) is a neurotransmitter mainly produced by two small groups of neurons in the midbrain: ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). They send diffuse although segregated connections to different areas of the brain, such as basal ganglia, most of cerebral cortex, amygdala, hippocampus, thalamus or the superior colliculi. Dopamine has been involved in many aspects of brain functioning (such as motor control, attention, memory, reward anticipation, pleasure, addiction) and dysfunctioning of the DA system leads to severe deficits such as the Parkinson disease, schizophrenia and autism [18].

These dopaminergic neurons exhibit stereotyped phasic excitatory responses of high amplitude, short duration (< 200ms) and short latency (70-100 ms) after several types of events: Delivery of primary rewards; sudden appearance of novel, intense or salient stimuli; and arbitrary stimuli classically conditioned by association with primary rewards [26]. The third column of Figure 1 shows the typical activation of a dopaminergic cell in SNc after classical conditioning: The cell only responds to the CS and not anymore to the reward delivery. This response pattern can be compared to the prediction error signal δ_t of the TD algorithm previously presented. Moreover, when reward is omitted, these DA cells also show a pause in firing (below baseline) that can be considered as a negative value of δ_t.

This analogy between DA firing and the reward prediction error of TD raises the question of the functional role of these DA neurons. If the TD analogy is correct, the phasic bursts observed in DA neurons should act as a "critic" signal for other other brain areas (which could then be considered as the "actor"). We will see in the section 3 that there are some elements supporting this point of view.

2.4 Limits of the TD analogy

Despite this striking similarity between DA firing and the TD error signal in classical conditioning, DA neurons can exhibit a different behavior from what would be expected if their only functional role was to be the "critic". Some of these criticisms can be found in [4, 23].

– *Backward shift in time:* In the extended TD model applied to classical conditioning, the positive δ_t gradually shifts back in time during learning from the time of reward to stimulus onset. However, what is observed in DA neurons is more a gradual decrease of the amplitude of the reward-related DA burst concurrently to a gradual increase of the CS-related activity. Nevertheless, a TD model using certain parameter settings (long-lasting eligibility traces) can mimic these observations [21].

– *Temporal representations:* No strong evidence has been found yet about a neural mechanism allowing to represent the time elapsed since stimulus onset by means of a successive firing of chained neurons ('eligibility traces'). What is rather found is "ramping" activities in the thalamus that increase from baseline activity at different speeds, until they reach their maximum level when an action has to be performed [29]. How these signals can be used to provide a useful temporal representation for a TD-like model is still unclear.

– *Varying stimulus-reward intervals:* The stimulus-reward interval can be uniformly varied (1 to 3 seconds) during the learning phase. Dopaminergic neurons are in this case responding both at stimulus onset and reward delivery, whatever the delay [10]. This can not be directly taken into account by TD-inspired models, which rely on a fixed stimulus-reward interval during learning. This is mainly due to the use of temporal representations for the learning of the critic. In [4], the authors address this problem by building a "mixture of experts" system, where each expert specializes on a specific stimulus-reward interval, but the questions of the number of experts needed, their temporal resolution and their biological plausibility remain unsolved.

– *Uncertainty of rewards:* More recent experiments have shown that the information carried by phasic DA bursts is not so stereotyped and represents quite finely reward amplitude and reward uncertainty [30]. This aspect is not taken into account in most TD-inspired models.

– *Novelty detection:* Contrary to the TD reward prediction error, DA cells also respond to the sudden appearance of novel, intense or salient stimuli, even if they are not associated to reward [12]. However, this response decreases when the subject becomes habituated to such an unrewarding stimulus.

In conditioning tasks with fixed stimulus-reward intervals, the firing of DA cells can be efficiently compared to the error signal of the TD algorithm, and some biological models have taken benefits from this approach (see [9] for a review). However, the above-mentionned limits show that DA cells can have a much more complex behavior and are involved in other tasks than pure conditioning. From the biological perspective, the TD analogy recently served as a guide to understand the functional role of DA and most experiments aimed at confirming this concept, but it may also be that the experiments that critically test the TD analogy have not been done so far.

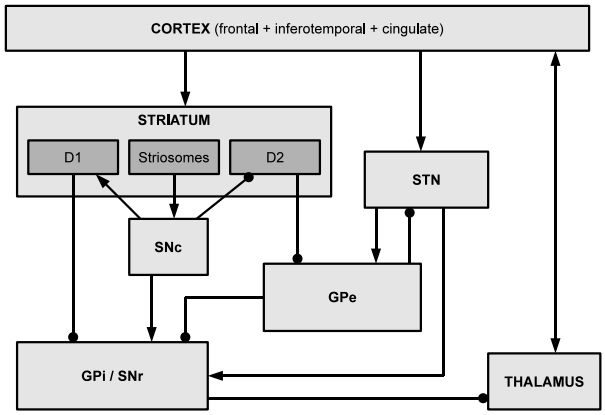

Figure 2: Schematic architecture of the BG. Pointed endings represent excitatory connections, round endings denote inhibitory connections. See the text for details about the structures. Modified from [1].

3 Basal Ganglia

As reviewed above, experimental evidence suggests that the functional role of DA neurons is related to reward-prediction error signals, but the analogy with the TD algorithm is not straightforward. To better understand the functional role of DA neurons, it may be useful to look at the targets of dopaminergic innervation (what would play the role of the "actor"), especially the basal ganglia (BG). The basal ganglia are a set of interconnected subcortical structures in the basal forebrain which are in interaction with the cerebral cortex, the thalamus and the limbic structures. They are composed of two main input structures - the striatum (STR) and the subthalamic nucleus (STN) -, two main output structures - the globus pallidus pars interna (GPi) and the substantia nigra pars reticulata (SNr) -, as well as one inner structure, the globus pallidus pars externa (GPe). The DA neurons of SNc are also considered as being part of the BG.

All these structures are densely interconnected and receive various connections from most parts of the cerebral cortex, the thalamus and limbic structures such as the amygdala and the hippocampus. However, since the influential work of [1], a simplified connectivity for BG has received much attention from modelers and experimentalists. On Figure 2, one can distinguish three main pathways through the BG. The direct pathway takes STR as an input structure (it receives dense topographical connections from the cerebral cortex) which projects in an inhibitory manner to the output structures of BG (GPi and SNr). These output structures are tonically active - which means they have a very high baseline activity - and strongly inhibit neurons in the thalamus (as well as in various motor subcortical structures such as the superior colliculi). When neurons in STR show sufficient activity, they inhibit some GPi/SNr neurons which in turn cease to inhibit thalamic neurons. These thalamic neurons can then be involved in a thalamocortical recurrent loop. As a whole, this double inhibition mechanism in the direct pathway of BG allows to selectively control the opening of thalamocortical loops depending on striatal processing. This direct pathway has been used in isolation in some BG models, for example [3] for

sequence learning or [27] for classical conditioning.

The second pathway in BG involves an intermediary step from STR to GPi/SNr through GPe. This relay is also inhibitory, which means this pathway globally increases the tonic inhibition of the thalamus by BG. This pathway is antagonistic to the direct one, but the projections from GPe to GPi/SNr are much more diffuse, so the compound effect of both direct and indirect pathways resembles more a center-surround effect. Moreover, the neurons in STR that directly project to GPi/SNr and the ones that project to GPe are mainly segregated, with different types of DA receptors. SNc can then act as a "controller" that can selectively favorize one pathway or the other, therefore opening or closing the corresponding thalamocortical loops. When one considers that DA bursts corresponding to reward predicting events favorize the opening of a loop and that DA depletions corresponding to omission of reward close these loops, one can see that this TD-like activation may be a central mechanism for selecting rewarding actions or the content of working memory [19].

The third pathway to consider uses STN as an input, which projects on both GPe and GPi/SNr. This hyperdirect pathway is still not well known, but STN receives direct cortical inputs and has faster conduction times than STR. STN excitatory projections to GPi/SNr are more diffuse than the direct or indirect pathways: it can provide a "global No-Go" signal, that avoids premature responses of the BG to incoming stimuli. The loop between GPe and STN can also provide a timing mechanism for action execution, or allow sequence learning.

This functional sketch is even more complicated by the fact that BG is organized in parallel segregated loops, each in different modalities (motor, limbic, associative...) [2]; or by the fact that DA also modulates learning of the connections between the cerebral cortex and the striatum. What needs to be pinpointed is the fact that BG can act as a controller for activities in the rest of the brain, and that DA has a central role in its functioning. The part of the striatum called striosomes on Figure 2 and the dopaminergic neurons can be functionally considered as a "critic" for this system, whereas the rest of this complex architecture would act as the "actor".

4 Computational models of BG

A good review of former BG models relying on the actor/critic architecture (e.g. [7,13]) can be found in [14]. We will now focus on three different BG models that, even if they do not all directly deal with RL, give further insight into the possible roles of BG in adaptive behavior. The interested reader could also have a look to other models of BG, for example [8] that incorporates an internal model of the environment to perform a tree-search predictive algorithm, or [16] which is a very biologically detailed models of dynamical oscillations in the BG.

4.1 Brown, Bullock and Grossberg (1999)

The idea that a single unitary mechanism is responsible for the responses of DA cells to both CS and rewards has been contested by [5]. Previous BG models derived from the TD analogy considered that the DA bursts for CS and reward were provoked by excitation from either primary rewards (limbic structures) or state evaluation in the striosomes of STR. The pause in DA firing was caused by a temporal mismatch between these two sources. In [5], the authors explored more precisely the possible biological sources of stimulation of DA neurons. They propose that the pedunculopontine tegmental nucleus (PPTN) is the unique source of excitation of DA neurons. This nucleus receives itself information about primary rewards (from lateral hypothalamus) as well as information about the appearance of a CS (from the output of BG, the "actor" part). The relevance of the appearance of a CS for reward has to be learned in the BG through association between a working memory of this stimulus and the delivery of reward.

The pause in DA firing when reward is omitted is solely due, according to the authors, to inhibition of SNc coming from the striosomes of STR. The question arose to know how to compute the timing of this inhibition, which should occur only at the time reward is expected. They proposed to use a mechanism for striosomal neurons called "spectral timing", which is an intracellular calcium-dependent timing mechanism. This mechanism generates something similar to the temporal representation, but only the neurons active at the time of reward delivery are selected for the learning process: there is no backward chaining in time.

At the time reward occurs (or is expected), there is an interaction in SNc between the reward-related excitation coming from PPTN and the inhibition coming from the striosomes. At the beginning of learning, the inhibition is weak, and the DA cells respond to the delivery of reward. After learning, excitation and inhibition compensate for each other, and the SNc neurons do not respond anymore to rewards, even if PPTN neurons still excite them. If reward is omitted, only the learned inhibition is transmitted to SNc, leading to a pause in DA firing.

This model is an elegant attempt to provide a biologically plausible explanation to the firing pattern of DA neurons during classical conditioning. Dopaminergic firing does not rely anymore on the temporal derivative of the predicted reward, but on separate mechanisms for reward-related bursts, CS onset-related bursts and pause in DA firing. However, it only uses the direct pathway of BG to compute the excitatory drive to PPTN leading to CS onset-related DA burst and spectral timing is not yet fully confirmed in striosomal cells.

4.2 O'Reilly and Frank (2006)

O'Reilly and colleagues recently proposed in [20] a biologically-inspired algorithm for conditioning called PVLV (Primary Value / Learned Value), which is principally derived from the previously presented model of Brown, Bullock and Grossberg [5], although with significant differences.

The PVLV algorithm relies on two separate modules depicted on Figure 3. The PV (Primary Value) module learns to predict the appearance of a primary reward, by means of an excitatory component (PVe, supposed to be located in the lateral hypothalamus LHA) and an inhibitory component (PVi, supposed to be located in the striosomes of STR). These two components have opposite effects on the firing of the DA neurons. The PVe component simply reflects the delivery of primary rewards, similarly to PPTN in the model of Brown, Bullock and Grossberg. The PVi component learns to predict the time interval between CS onset and reward delivery. Contrary to [5], this timing mechanism is not supposed to lie in the striosomal cells, but rather

to come from an external signal computed in the the cerebellum, where precise timing-related activities have been found [17]. The PV is only active at the time when reward is delivered or expected to be delivered. It explains the inital reward-related DA bursts before learning, its slow decrease in amplitude during learning, as well as the depletion of the dopaminergic signal when a predicted reward is omitted.

The LV (Learned Value) module learns to respond to the appearance of conditioned stimuli that are reliably associated to reward. It is also decomposed into an excitatory component (LVe, which is thought to be located in the central nucleus of the amygdala CNA) and an inhibitory component (LVi, also located in the striosomes of STR). Similarly to PVi, the LVe component learns to associate a CS with reward, but only at the time reward is actually given or expected (as computed by the PV system), through a learning rule derived from the Rescorla-Wagner rule [25]. This feature is important for understanding the model, because a permanent representation of the CS has to be present at the time reward is delivered or expected. This works perfectly for delay conditioning (the CS is still present when reward is given), but require an additional memory mechanism for CS in the case of trace conditioning (the CS is removed before reward is given). The LVi subsystem only slowly learns to cancel the LVe burst, to denote habituation to a CS.

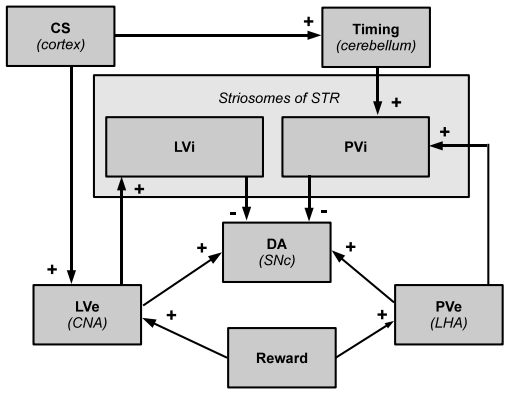

Figure 3: Architecture of the PVLV algorithm. PVLV is made of two subsystems: PV (Primary Value, PVe and PVi) learns to predict the appearance of a primary reward after CS onset. LV (Learned Value, LVe and LVi) learns to respond for a CS that is reliably associated to reward. Modified from [20].

This model can account for several classical conditioning paradigms such as aquisition, extinction, blocking, overshadowing and conditioned inhibition. This algorithm can be also used to perform second-order conditioning (CS-CS associations) by using the computed DA signal to bias cortical associations between the two CS. This is in opposition to the way TD-like algorithms deal with second order conditioning. More interestingly, contrary to TD-like algorithms, PVLV is not sensible to varying stimulus-reward intervals, which is of great interest when applied to real-world tasks. Whereas the other algorithms represent information about the sequence of events (time) in the stimulus representation, the PVLV algorithm uses an additional timing mechanism in cerebellum. According to the authors such external timing mechanism is not crucial for the model but it is essential for particular experimental tasks, e.g. to fit the observed pause in DA firing.

However, this algorithm is not designed to stand on its own, except in the case of delay conditioning. For trace conditioning, the working memory of the briefly flashed CS has to rely on another circuit involving the prefrontal cortex and BG. In [19], the authors have designed a complete BG model (both direct and indirect pathways) which learns in interaction with the PVLV module to keep the CS that are frequently associated to reward in working memory. This complex architecture is able to solve cognitive tasks like the 1-2-A-X task or the Store-Ignore-Recall task.

4.3 Gurney, Prescott and Redgrave (2001)

Dopamine cells respond with a latency between 70 and 100 ms, which is shorter than the latency of the saccades bringing the stimulus onto the fovea for a more detailed analysis (150 to 200 ms). Signals regarding the identity of visual objects are detected in the inferotemporal cortex with a latency of 80 to 100 ms after stimulus onset, therefore at least at the same time as dopamine firing, raising the problem of how this information can reach the dopamine areas almost immediately. In [24], Redgrave and Gurney consequently conclude that the rich and detailed representations in the cerebral cortex are a bad candidate to provide the dopamine system with an accurate reward prediction. They propose that dopamine responses are triggered as a consequence of limited pre-attentive processing that would be computed in the superior colliculus (SC), which have very early visual responses and can quickly provide the dopaminergic neurons with information about the novelty or the reward association of a stimulus at a fixed position, without having to process its visual details.

According to [23], DA firing would principally be used by BG and other areas to indentify which aspect of context or behavior is crucial in causing unpredicted events. Through repetition of interactions between the agent and his environment, DA would help to distinguish between the consequences of the agent's own actions and what is caused by external events. This highlights the role of DA in operant conditioning, contrary to the classical Pavlovian conditioning which requires a fine analysis of the details of the stimulus and would be treated by other cortical structures like orbitofrontal cortex.

These ideas lead the authors to propose a totally different functional model of BG in [11] which is intended to perform action selection instead of RL (which they claim is computed somewhere else in the brain). They propose to rearrange the functional connectivity of the BG by giving a central role to STN. They distinguish two pathways: one is called the selection pathway and comprises a part of STR, STN and the output nuclei GPi/SNr, the other is called the control pathway and contains STR, STN and GPe. The selection pathway performs the selection between different salient events (cortical representations, actions...) through a disinhibition mechanism similar to the direct pathway of classical BG models. The role of the control pathway is to regulate processing in the selection pathway through the connections from GPe to both STN and GPi/SNr. The intensity of this regulation is under the influence of DA, which only signals the behavioral interest of salient events. This BG model has been successfully applied to action selection in robotic tasks.

5 Discussion

Adaptive behavior in animals and humans can be analysed through the paradigm of reinforcement learning as the behavior of a rational agent aims at optimizing the future rewards it can obtain through interactions with its environment. It may be of particular interest to investigate more precisely whether some analogy can be found between brain processes and RL algorithms. Such an analogy has been suggested concerning the DA neurons of the midbrain during classical conditioning. They tend to behave similarly to the reward prediction error signal of the classical TD algorithm. A further look at the functional connectivity of these neurons also leads to the idea that these DA neurons and a part of STR could be viewed as critic, whereas the BG and the cerebral cortex as a whole could be viewed as the actor in the classical RL actor/critic architecture. This paradigm has guided the development of several biological models of BG performing reinforcement learning.

However, we have listed experimental evidence that the functioning of these brain areas is not solely devoted to this particular paradigm and that the same neurons show different patterns depending on the type of action and learning the agent is involved in. In particular, the models presented in sections 4.1 and 4.2 interpreted the available data to suggest that the particular behavior of DA during conditioning is not guided by a single unitary mechanism computing the time derivative of reward prediction, but rather by the interaction of several information flows coming from distinct brain areas. On top of their greater biological plausibility, these models are able to make predictions about the role of the BG in processes that are indirectly related to reinforcement learning, such as working memory or more abstract cognitive tasks.

Some authors even take a more radical approach (like in section 4.3) and object to assign the DA bursting patterns such a central role in adaptive behavior, especially in selection of action. They suggest that DA only signals quantitatively the behavioral importance of salient events, without giving precise information about their value.

The exact role of DA in brain processes is therefore not yet fully understood. Its firing pattern in classical conditioning can denote a role in criticizing other brain areas, but it could also be a side effect of other processes elsewhere in the brain. It can act at different levels, by modifying the intracellular properties of various neurons or by modulating the learning of synaptic strength. Its effect also varies depending on which area is targeted: contrary to STR, the cerebral cortex is much more affected by the tonic level of DA (its baseline) than the phasic bursts [31]. As a consequence, RL processes in the brain should be seen as a global process involving various brain areas that cooperate to fulfill a task, without any clear demarcation between parts of the RL algorithms.

The performance of the presented models have been demonstrated on a number of tasks that go further than the pure RL paradigm, such as working memory or decision making. Whether these high-level processes only take benefit of the "built-in" RL properties of the dopaminergic system (in an epigenetic sense), or whether they belong to a more global process that can be reduced by specific observations to something that looks like RL (the magnifying glass effect), is still an open question. What is important for each RL or BG model is to specify its domain of validity until a comprehensive theory of DA can emerge. From the technical point of view, TD models have been very powerful in solving callenging tasks, but we are still far away from fundamental solutions for real-world cognitive agents. From our point of view, future research should also focus on demonstrating that additional biological details also lead to improved performance for cognitive agents behaving in the real world.

Acknowledgements

This review article has been supported by the German research foundation (Deutsche Forschungsgemeinschaft) grant (DFG HA 2630/4-1) "A neurocomputational systems approach to modeling the cognitive guidance of attention and object/category recognition" and by the FP7-ICT program of the European Commission within the grant "Eyeshots: Heterogeneous 3-D Perception across Visual Fragments".

References

[1] R. L. Albin, A. B. Young, and J. B. Penney. The functional anatomy of basal ganglia disorders. *Trends Neurosci*, 12(10):366–375, 1989.

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

[3] G. Berns and T. Sejnowski. A computational model of how the basal ganglia produce sequences. *J Cogn Neurosci*, 10:108–121, 1998.

[4] M. Bertin, N. Schweighofer, and K. Doya. Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Netw*, 20(6):668–675, 2007.

[5] 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, 1999.

[6] H. J. Chiel and R. D. Beer. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci*, 20(12):553–557, 1997.

[7] J. L. Contreras-Vidal and W. Schultz. A predictive reinforcement model of dopamine neurons for learning approach behavior. *J Comput Neurosci*, 6(3):191–214, 1999.

[8] N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711, Dec 2005.

[9] K. Doya. Reinforcement learning: Computational theory and biological mechanisms. *HFSP J.*, 1(1):30–40, 2007.

[10] C. D. Fiorillo and W. Schultz. The reward responses of dopamine neurons persist when prediction of reward is probabilistic with respect to time or occurence. *Society for Neuroscience Abst.*, 27:827–853, 2001.

[11] K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biol Cybern*, 84(6):401–410, 2001.

[12] J. C. Horvitz. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96(4):651–656, 2000.

[13] J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generate and use neural signal that predict reinforcement. In J. L. Davis J. C. Houk and D. G. Beiser, editors, *Models of information processing in the basal ganglia*. The MIT Press, Cambridge, MA, 1995.

[14] D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw*, 15(4-6):535–547, 2002.

[15] L. J. Kamin. "attention-like" processes in classical conditioning. pages 9–31, 1968.

[16] A. Leblois, T. Boraud, W. Meissner, H. Bergman, and D. Hansel. Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J Neurosci*, 26(13):3567–3583, Mar 2006.

[17] M. D. Mauk and D. V. Buonomano. The neural basis of temporal processing. *Annu Rev Neurosci*, 27:307–340, 2004.

[18] A. Nieoullon. Dopamine and the regulation of cognition and attention. *Prog Neurobiol*, 67(1):53–83, 2002.

[19] R. C. O'Reilly and M. J. Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput*, 18(2):283–328, 2006.

[20] R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz. Pvlv: the primary value and learned value pavlovian learning algorithm. *Behav Neurosci*, 121(1):31–49, 2007.

[21] W. X. Pan, R. Schmidt, J. R. Wickens, and B. I. Hyland. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J Neurosci*, 25(26):6235–6242, Jun 2005.

[22] I. P. Pavlov. *Conditioned reflexes*. Oxford University Press, London, 1927.

[23] P. Redgrave and K. Gurney. The short-latency dopamine signal: a role in discovering novel actions? *Nat Rev Neurosci*, 7(12):967–975, 2006.

[24] P. Redgrave, T. J. Prescott, and K. Gurney. Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci*, 22(4):146–151, 1999.

[25] R. A. Rescorla and A. R. Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory*, pages 64–99. Appleton-Century-Crofts, 1972.

[26] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[27] R. E. Suri and W. Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871–890, 1999.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998. A Bradford Book.

[29] M. Tanaka. Cognitive signals in the primate motor thalamus predict saccade timing. *J Neurosci*, 27(44):12109–12118, 2007.

[30] P. N. Tobler, C. D. Fiorillo, and W. Schultz. Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645, 2005.

[31] J. Vitay and F. H. Hamker. Sustained activities and retrieval in a computational model of the perirhinal cortex. *J Cogn Neurosci*, 20(11):1993–2005, 2008.

Contact

Prof. Dr.-Ing. Fred H Hamker
Fakultät für Informatik, Technische Universität Chemnitz
Straße der Nationen 62, D-09107 Chemnitz
Tel.: +49 (0)371-53137875
Email: fred.hamker@informatik.tu-chemnitz.de
WWW: http://www.tu-chemnitz.de/informatik/KI

Bild

Julien Vitay obtained a diploma in electrical engineering from Supélec (France) in 2002 and a Ph.D. degree in computer science in 2006 from the University Henri Poincaré Nancy-I (France). He is currently a post-doctoral fellow in the research group of Fred Hamker and focuses on computational models of the basal ganglia for the cognitive control of visual perception and action selection.

Bild

Jérémy Fix obtained a diploma in electrical engineering from Supélec (France) in 2005 and a Ph.D. degree in computer science in 2008 from the University Henri Poincaré Nancy-I (France). He is currently a post-doctoral fellow in the research group of Fred Hamker and focuses on computational models of the basal ganglia for the cognitive control of visual perception and action selection.

Bild

Frederik Beuth received a diploma in computer science from the TU Chemnitz in 2008. He is currently a PhD student in the research group of Fred Hamker.

Bild

Henning Schroll is studying Psychology at the University of Münster. He is currently working on his diploma thesis, applying basal ganglia modeling to psychological tests.

Bild

Fred Hamker received his diploma in electrical engineering at the Univ. of Paderborn in 1994 and his PhD in computer science at the TU Ilmenau in 1999. He was a PostDoc at the Univ. of Frankfurt and the California Institute of Technology (Pasadena, USA). In 2003, he established a research group at the Univ. of Münster in the department of psychology. Since 2009, he is professor for Artificial Intelligence at the TU Chemnitz. His research group persues an experimental and model-driven approach to explore visual perception and cognition. It aims at understanding brain processes by neurocomputational models that explain experimental data and that are able to perform real-world tasks.