Cogn Neurodyn (2010) 4:91–105
DOI 10.1007/s11571-010-9109-x

REVIEW

Computational models of reinforcement learning: the role of dopamine as a reward signal

R. D. Samson · M. J. Frank · Jean-Marc Fellous

Received: 26 September 2009 / Revised: 17 February 2010 / Accepted: 24 February 2010 / Published online: 21 March 2010

Abstract Reinforcement learning is ubiquitous. Unlike other forms of learning, it involves the processing of fast yet content-poor feedback information to correct assumptions about the nature of a task or of a set of stimuli. This feedback information is often delivered as generic rewards or punishments, and has little to do with the stimulus features to be learned. How can such low-content feedback lead to such an efficient learning paradigm? Through a review of existing neuro-computational models of reinforcement learning, we suggest that the efficiency of this type of learning resides in the dynamic and synergistic cooperation of brain systems that use different levels of computations. The implementation of reward signals at the synaptic, cellular, network and system levels give the organism the necessary robustness, adaptability and processing speed required for evolutionary and behavioral success.

Keywords Reinforcement learning · Dopamine · Reward · Temporal difference

R. D. Samson
Evelyn F. McKnight Brain Institute and Neural Systems, Memory and Aging, University of Arizona, Tucson, AZ 85724, USA

M. J. Frank
Department of Cognitive and Linguistic Sciences and Department of Psychology, Brown Institute for Brain Science, Brown University, Providence, RI 02912, USA

J.-M. Fellous (✉)
Department of Psychology and Applied Mathematics, University of Arizona, 1501 N. Campbell Av, Life Sciences North, #384, Tucson, AZ 85724, USA
e-mail: fellous@email.arizona.edu

Introduction

In computer sciences, reinforcement learning (RL) combines theories from machine learning, artificial intelligence and dynamic programming. It refers to the trial-and-error learning of the set of actions an agent must take to maximize future rewards or minimize punishments (Sutton and Barto 1998). RL is fundamental to the psychological and neuroscientific studies of reinforcement and conditioning, and to the neuroscience of decision-making in general including neuroeconomics (Glimcher and Rustichini 2004; Camerer 2008). Since the early 1980s, there has been a growing interest in mapping computational theories of RL to their underlying neural mechanisms. On the theoretical side, it has become clear that RL results from the complex interactions between different computational subsystems describing processes internal to the organisms and accounting for its interactions with the environment (Fig. 1; Sutton and Barto 1998; Freeman 2007). On the biological side, it has also become clear that there is no single brain area responsible for the implementation of RL and that learning and reward processing are highly distributed functions involving dozens of dynamically interacting brain structures (Dayan and Balleine 2002). As this review will show, RL is computationally implemented at multiple levels, from chemical to systems.

In RL, an agent interacts with its environment in order to learn the best actions it must perform to maximize the sum of future rewards. RL models are typically composed of 4 main components (Fig. 1; Sutton and Barto 1998). (1) A reward function, which attributes a desirability value to a state. The reward is often a one-dimensional scalar $r(t)$, computed on the basis of information streams from the environment. It is an instantaneous scalar that can carry a positive or a negative value. (2) A value function (also

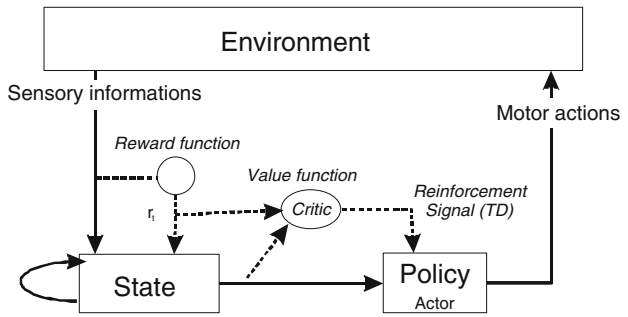

Fig. 1 Schematic representation of the reinforcement process. An agent is involved in a constant and dynamical action-perception loop (*plain arrows*) which involves the world, the agent's state, and the agent's policies. Reinforcement learning makes use of a parallel stream of information (*dashed lines*) providing additional input dimensions to the State and Policy modules. These reinforcing signals are typically low in information content

known as *critic*) which determines the long term desirability of a state, based on how much future rewards can be expected for being in that state. The value function may or may not be contingent upon actions taken by the agent. In most models, the output of this function is computed as the Temporal Difference (TD) error between estimated and actual rewards. (3) A policy function (also known as *actor*) which maps the agent states to possible actions, using the output of the value function (the reinforcement signal) to determine the best action to choose. (4) A model of the environment which includes a representation of the environment dynamics required for maximizing the sum of future rewards. These general RL concepts are further explained and elaborated by various authors (Dayan and Daw 2008; Kaelbling et al. 1996; Montague et al. 2004a; Sutton and Barto 1998; Worgotter and Porr 2005; Dayan and Abbott 2001).

The general goal of RL models is to maximize the sum of future rewards. However, if rewards are delivered after a delay, or with low probabilities, it can be challenging to determine which state or action to associate with the current reward. This issue is known as the 'temporal credit-assignment problem' (Barto et al. 1983; Sutton 1984). One way of addressing this issue with computational models is to incorporate an eligibility trace. An eligibility trace can be defined as a slowly decaying temporal trace of an event, used to determine which state or action is to be reinforced when a reward arrives later in time under a given policy (Sutton and Barto 1998). Eligibility traces have been implemented computationally at various levels and a few examples will be presented in this review.

The computer science notion of RL was originally inspired from the psychological theories of reinforcement, and the underlying process of associative learning (Cardinal et al. 2002; Mackintosh 1983; Wise 2004; Wise and Rompre 1989). At the neural level, reinforcement allows for the strengthening of synaptic associations between pathways carrying conditioned and unconditioned stimulus information. At cognitive and system levels, reinforcement processes may be the basis of the neural substrates of the emotional state (Rolls 2000). The intrinsic nature of rewards in psychology and computer science, however, is different. In computational RL theories, rewards are simple, 'information poor' scalars influencing learning, while in psychology, rewards are signals from the environment that can arouse the organism (Berridge 2007).

Despite the differences between computational RL theories and the psychological and physiological aspects of conditioning, findings from these research fields are complementary and converging toward a better understanding of the brain mechanisms underlying reinforcement. RL models have contributed to- and continue to unveil the process underlying reinforcement from the neurochemical to systems levels. We present an overview of the neurobiological underpinnings of reinforcement learning and review the various ways in which the role of rewards has been implemented computationally. We point to the fact that reward signals act as reinforcers for multiple neural mechanisms, being at neurochemical levels or at the system levels. We review models of dopamine (DA) neuron activity elicited by reward-predicting stimuli, models of DA transmission, models of the effect of DA on synaptic transmission and plasticity, and models of the effect of DA on complex neural circuits mediating behavior.

Neurochemical level: models of dopamine neuron activity

A large body of experimental work has shown that reward-motivated behavior depends on the activity of DA neurons from the midbrain ventral tegmental area (VTA) and substantia nigra pars compacta (SNc; Ljungberg et al. 1992; O'Doherty et al. 2003; Seymour et al. 2004; Schultz 1998; Wise and Rompre 1989). Dopamine has been shown to play a role in motivation (Fibiger and Phillips 1986; Robbins and Everitt 1996; Wise and Hoffman 1992; Wise 2004, 2005), in the acquisition of appetitive conditioning tasks (Berridge and Robinson 1998; Everitt et al. 1999; Ikemoto and Panksepp 1999), in many aspects of drug addiction (Berke and Hyman 2000; Di Chiara 2002; Everitt and Robbins 2005; Kelley and Berridge 2002; Koob 1992; Robinson and Berridge 2008; Wise 1996a, b) and is involved in disorders such as Parkinson disease (Canavan et al. 1989; Voon et al. 2010; Frank 2005; Knowlton et al. 1996; Moustafa et al. 2008).

Although computational theories of reinforcement learning were in full force in the early 1980s, their popularity significantly increased later, in the early 1990s, with

the finding that DA cells were responsive to rewards and reward predictive stimuli (Mirenowicz and Schultz 1994, 1996; Schultz 1992, 1998; Waelti et al. 2001). Compatible with RL theories, these neurons transiently increased their activity in response to the presentation of unexpected rewards (i.e. time limited, in the order of 100 ms) and in response to cues predicting upcoming reward delivery (Hikosaka et al. 2008; Romo and Schultz 1990; Schultz et al. 1997). As illustrated in Fig. 2, during classical conditioning, unexpected reward initially triggered an increase in phasic activity of DA neurons, which then shifted with learning to the conditioned stimulus (Ljungberg et al. 1992; Mirenowicz and Schultz 1994; Pan et al. 2005). When an expected reward was omitted, the activity of DA neurons paused at the precise time when the reward should have been delivered (Ljungberg et al. 1992; Roesch et al. 2007; Satoh et al. 2003). This activity pattern suggests that DA neurons signal a 'reward prediction error'. After learning therefore, DA neurons do not respond to the reward itself, but the difference between expected and received reward. This notion of 'prediction error' is a central component of current RL theories.

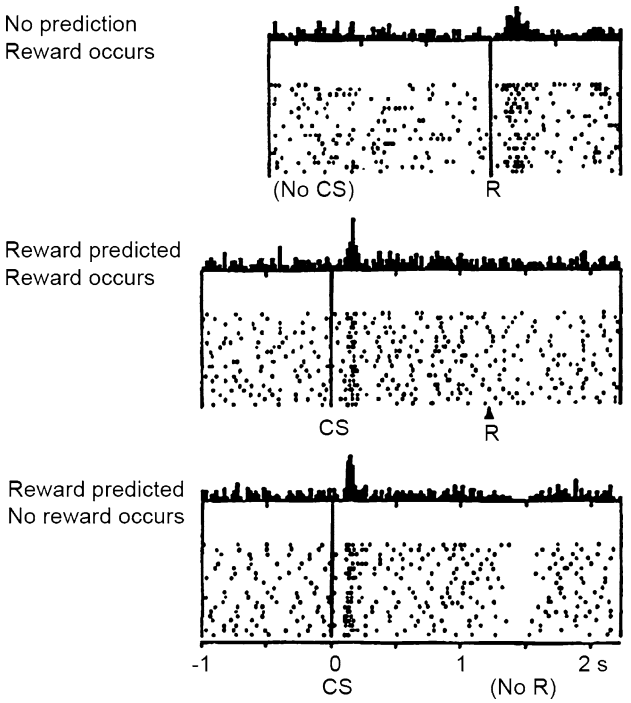

Fig. 2 Dopamine neurons report a reward prediction error. Peri-event time histogram of SNc neurons during an appetitive Pavlovian task. DA neurons show increased activations to unpredicted rewards (*R*; *Top*) and to learned conditioned stimuli predicting rewards (*CS*; *middle*). Once the stimulus-reward associations is learned, reward delivery no longer elicits an increase in the activity of DA neurons as it is fully expected (*middle*).When an expected reward is omitted, the activity of DA neurons drops at the moment where the reward should have been delivered (*bottom*). Reproduced with permission from Schultz et al. (1997)

Temporal difference model

The activity pattern of DA neurons represents the reward prediction error, which is central to the temporal difference (TD) model developed by Sutton and Barto (Sutton 1988; Sutton and Barto 1998). The TD model is based on a first-order Markovian process, meaning that the transition from one state to the next depends only on the current state and action and not on those observed previously. Markov decision processes (MDP) model the long-term optimality of taking certain actions depending on the current states (White 1993; Sutton and Barto 1998). Because knowledge of the sum of future rewards is not available to the agent when making a decision, Sutton and Barto adapted the TD method to compute bootstrapped estimates of this sum, which evolve as a function of the difference between temporally successive predictions. Specifically, the TD model calculates a prediction error $\delta(t)$ based on the temporal difference between the current discounted value function $\gamma V(s_{t+1})$ and that of the previous time step, $V(s_t)$.

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

In this equation, γ is a discount factor which allows rewards that arrive sooner to have a greater influence over delayed ones, and r_{t+1} represent the current reward (Sutton and Barto 1990, 1998). The prediction error is then used to update the value function estimate, which is a value of the long term desirability of a state (see Fig. 1)

$$V(s_t) = V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

where α is a 'step-size' parameter representing the learning rate. The TD model is used to guide learning so that future expectations are more accurate.

Numerous research groups have correlated the activity of DA neurons with the TD error originally proposed by the algorithm (Bayer and Glimcher 2005; Morris et al. 2004; Montague et al. 1996; Nakahara et al. 2004; Roesch et al. 2007; Schultz et al. 1997) and found that it indeed followed the model at many levels. For instance, the magnitude of the change in DA neuron activity was recently shown to be modulated by the size of the reward prediction error and by the length of the stimulus-reward interval, such that the most valuable reward induced the greatest increase in neural activity (Kobayashi and Schultz 2008; Roesch et al. 2007). This is in line with the TD model, which incorporates a discount factor to take these magnitude differences into account. Similarly, the duration of the pause of DA neuronal firing was shown to be modulated by the size of the negative reward prediction error (Bayer et al. 2007). However, not all aspects of DA cell activity are modeled by the traditional TD model. The following account presents some modifications to the TD model, as well as alternative models.

Incorporation of timing variability to the TD model

Rewards are not always delivered immediately after an action is taken, but are often delayed. The TD model typically does not account for such delays, thus new models have been developed to incorporate larger or variable delays between actions and rewards. One such model was designed to account for the change in activity of a VTA neuron during a delayed reward task (Montague et al. 1996). This model is referred to as the 'tapped delay line' or 'complete serial compound' model because it creates a different sensory stimulus representation for each time step following stimulus onset whether or not a reward has been delivered. This strategy gives flexibility in reward delivery times and also allows for the updating of the input weights at each time step. This model accounted for the time difference between cue and reward delivery by explicitly representing cortical inputs and reward outcomes at each time step. The modification of the synaptic weights was achieved using a classical Hebbian learning rule that accounted for the correlation between presynaptic stimulus-related activity and the reward-dependent prediction error.

Daw and colleagues further extended the TD model to incorporate variability in reward timing, and to allow for a greater range of state representations, by using semi-Markovian dynamics and a Partially Observable Markov Decision Process (POMDP; Daw et al. 2006). In a semi-Markov process, state transitions occur probabilistically and the probability of transition depends on the current state and the amount of time spent in that state (referred to as 'dwell time'). This modification using semi-Markov dynamics added information about timing into the TD model. This model was further extended to allow for partial observability (e.g., when a reward was omitted). POMDP removes the one-to-one relationship between states and observations. Instead, it generates a probability distribution of states computed on the basis of estimates of the presence or absence of rewards. This modification in the model allowed the state representation, or state value to better reflect the expectations of the agent.

Finally, another way of representing variability in stimulus-reward intervals is based on the multiple model-based RL (MMRL) framework of Doya et al. (Fig. 3; Bertin et al. 2007; and for other application of the MMRL framework Doya et al. 2002; Samejima et al. 2003). Briefly, the MMRL model can be described as parallel modules, each containing a value estimator and a reward predictor. The value estimator module computes a value

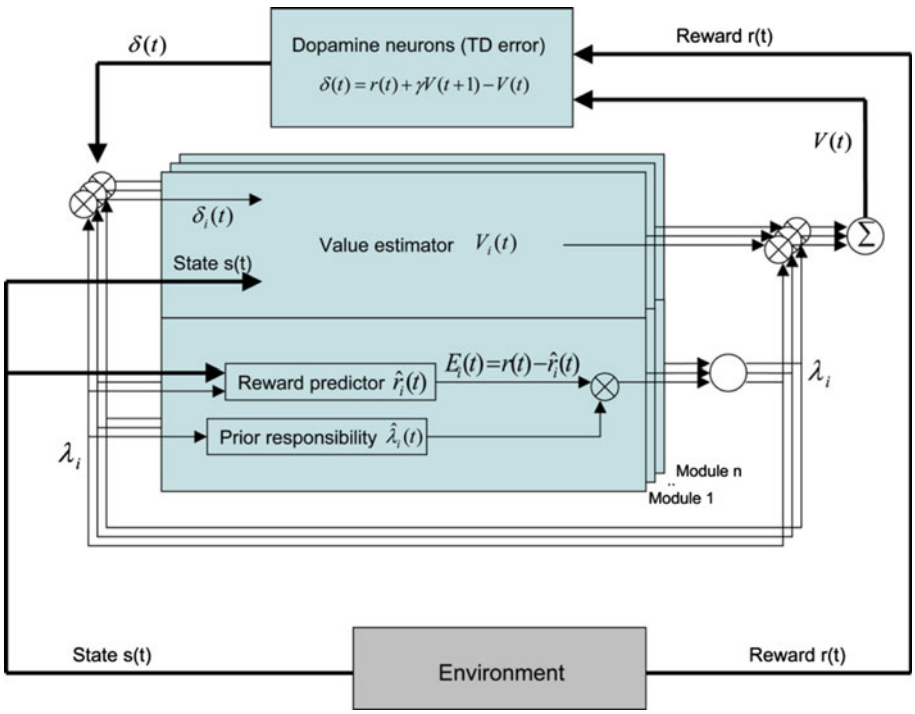

Fig. 3 Multiple Model Reinforcement Learning (MMRL) model. The activity of DA neurons is given by the global TD error $\delta(t)$. Each module is composed of a value estimator and a reward predictor. The value estimator outputs a reward prediction (presence or absence) at each time step. The reward predictor gives a vector of amounts of predicted reward at each time step following a conditioned stimulus. The prior responsibility predictor gates the reward predictor output based on prior knowledge about the module it is in. The reward predictor output is used to compute the new responsibility signal λ_i and gates the update of the reward predictor and value estimator, as well as the TD error for each module. At last, the TD error updates the weight vector of each value estimator. Reproduced with permission from Bertin et al. (2007)

function (based on the TD algorithm) and outputs the reward prediction. The reward predictor module generates a prediction of the amount of reward for each state, based on the amount of rewards received in the previous state. This module then generates a 'responsibility' signal which will gate the reward prediction, hence the TD error, for each module. The responsibility signal also serves to update the value estimator and reward predictor of the next module. Because this model is updated at every time step, it represents rewards delivered at different time intervals, similar to the tapped delay line model of Montague et al. (1996) mentioned above. The MMRL model also allows earlier rewards to weigh more than ones received later.

Together, these models explain how DA neurons may be dynamically representing prediction errors. However, DA neurons provide more information than a simple global prediction error signal. Indeed, the activity of VTA dopaminergic neurons was recently shown to predict the 'best reward option', in terms of size, delay and probability of delivery (reviewed in Hikosaka et al. 2008; Roesch et al. 2007). The activity of some DA neurons in monkey SNc was also known to reflect the choice of the future motor response (Morris et al. 2006). In addition, the activity of a population of DA neurons in the monkey SNc was shown to increase in response to stimuli predicting punishments (Matsumoto and Hikosaka 2009). Current computational models do not typically account for these data.

Dopamine is known to be released at more prolonged timescales than glutamate or GABA (seconds to minutes; Abercrombie et al. 1989; Arbuthnott and Wickens 2007; Louilot et al. 1986; Schultz 2002; Young et al. 1992). For this reason, it is possible that the activity of DA neurons does not reflect accurately its pattern of release at target structures or the timescale of its postsynaptic effects (but see Lapish et al. 2007). The following section reviews some experimental findings and models of DA release.

Post-synaptic level: models of the local control of the effects of dopamine on target brain areas

As presented above, it is now well established that the activity of the DA cells of the VTA/SNc correlates with prediction error. How is this phasic change in firing activity translated into DA release at target structures? Experimental data has shown that a single-pulse stimulation applied to the VTA or the median forebrain bundle leads to a DA release lasting a few seconds (Kilpatrick et al. 2000; Fields et al. 2007; Phillips et al. 2003; Robinson et al. 2002; Roitman et al. 2004; Yavich and MacDonald 2000). The probability of release of DA is however lower than that of glutamate and is varies greatly among DA synapses (Daniel et al. 2009). Furthermore, while the phasic increase in DA release is commonly referred to in the literature, evidences for tonic changes in DA concentrations (e.g. long lasting, timescales of seconds to minutes) have also been reported in vivo, in the basal ganglia and in prefrontal cortices (PFC; Bergstrom and Garris 2003; Floresco et al. 2003; Goto and Grace 2005; Grace 1991).

Model of dopamine release

To better understand the kinetics of DA delivery in the basal ganglia in response to stimulating pulses applied to the VTA/SNc, Montague et al. (2004b) created the "kick and relax" model. The authors used the Michaelis–Menten first order differential equation which is often used in biochemistry to describe enzymatic reactions. In this framework, the change in DA concentration was calculated as the difference between the rate of its release and that of its uptake. The authors adapted the equation to match their experimental finding which showed that stimulation trains applied to the VTA/SNc with short inter-burst intervals (2 s) induced a greater DA release in the caudate/putamen than with longer inter-burst intervals (5 s) that lead to a 'depression' in DA release. The model parameters included a 'kick' factor to facilitate and a 'relax' factor to depress the influence of input spikes so that the associated time constant of DA concentration decayed over time. The model suggested that dopamine release was in part a dynamical process controlled locally within the target structure.

Models of the effects of dopamine release

Part of understanding how incentive learning is processed in the brain involves knowing and understanding how DA neurons modulate the activity of their postsynaptic targets. The main projection of DA neurons is to the striatum (Lindvall and Bjorklund 1978). Early models proposed basic mechanisms by which catecholamines could change the input/output characteristics of neurons (Servan-Schreiber et al. 1990). This early work focused on general intrinsic membrane characteristics. More recent work on the specific actions of D1 and D2 dopamine receptors has refined the understanding of the role of dopamine on postsynaptic targets (Surmeier et al. 2007). While D1 activation is in general excitatory, it depends on the state of depolarization of the postsynaptic membrane potential; hyperpolarized neurons will respond to D1 activation by lowering their firing rate, while depolarized neurons will do the opposite. Thus, striatal neurons already activated by convergent excitatory glutamatergic inputs from the cortex will tend to be facilitated by D1 stimulation, whereas those firing spuriously may be inhibited. On the other hand, D2 activation is in general inhibitory.

Models illustrating the possible functional roles for the differential activation of D1 and D2 receptors have been recently proposed. Frank and colleagues built on the work of Suri, Wickens and others, and investigated the effects of dopaminergic modulation of striatonigral and striatopallidal cells in the classical "direct" and "indirect" pathways of the basal ganglia (Frank 2005; for review and updates see Cohen and Frank 2009). There, dopamine increases the signal to noise ratio in striatonigral "Go" neurons expressing D1 receptors, modeled by an increase in the gain of the activation function and an increase in firing threshold. Conversely, dopamine inhibits the activity in striatopallidal "NoGo" neurons expressing D2 receptors. Together these effects determine action selection, whereby Go activity facilitates and NoGo activity suppresses, the selection of a cortical action via modulation of basal ganglia output nuclei and ultimately, thalamocortical activity (Fig. 4).

Together these experimental and theoretical findings indicate that the short term computational influences of DA depends in part on the firing rate of DA cells, the local dynamics of DA release and the exact nature of the activated postsynaptic receptors. The effect of DA on its target structures is however more complex than previously thought. For example, it involves intricate interactions with glutamate, which is also released by VTA/SNc cells, as well as from cortical afferents onto striatal neurons

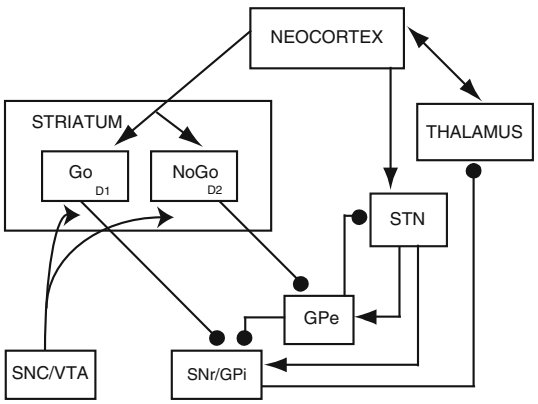

Fig. 4 Illustration of the striato-cortical loop. In this brain circuit, DA projections from the VTA/SNc target the striatal neurons and activate D1 and D2 receptors located on neurons of the direct (Go) and indirect (NoGo) pathways of the basal ganglia. The direct pathway refers to the projection of striatal neurons to the internal segment of the globus pallidus (GPi), which is the output of the basal ganglia. The indirect pathway refers to the projection of striatal neurons to the GPi, via the external segment of the globus pallidus (GPe). D1 receptor activation in the striatum leads to activation of the Go cells of the direct pathway and the consequent initiation of movements, whereas D2 receptor activation inhibits the NoGo cells of the striatum leading to the suppression of movements. The GPi and the substantia nigra pars reticulata (SNr) target the thalamic nuclei projection to the frontal cortex Adapted from Frank et al. 2005

(reviewed by Cepeda and Levine 1998; Kötter and Wickens 1995; Nicola et al. 2000; Reynolds and Wickens 2002). In the NAc, DA modulates glutamate co-transmission in a frequency-dependent manner, whereby only high-frequency stimulation of VTA neurons leads to DA facilitation of glutamate transmission (Chuhma et al. 2009). DA cells of the VTA are also known to co-release GABA (Gall et al. 1987; Maher and Westbrook 2008; Schimchowitsch et al. 1991). The network activity in the target brain areas also plays an important role in controlling local activity, and therefore in controlling the influence of DA on its postsynaptic cells. For instance, it was shown that dopamine could have both suppressing and enhancing effects depending on its concentration and on the 'UP' or 'DOWN' state of the target networks in vitro and in vivo (Kroener et al. 2009; Vijayraghavan et al. 2007). The intricacies of DA effects on transmission are so complex that it will fuel many more experimental and computational studies. Beyond its effects on synaptic transmission, DA can also have long term effects through its influence on synaptic plasticity at target structures including the basal ganglia, prefrontal cortex and amygdala.

Synaptic plasticity level: models of the role of dopamine in changing long-term synaptic strength

Behaviorally, DA is required in the early stages of Pavlovian conditioning paradigms, such as approach behaviors to food rewards or instrumental conditioning in general (reviewed in Wickens et al. 2007). This DA-dependency is due in part to a DA modulation of the plastic synaptic events that underlie reinforcement and associative learning. Synaptic plasticity has indeed been shown to occur at target structures of the VTA, including the NAc (Kombian and Malenka 1994; Li and Kauer 2004; Pennartz et al. 1993; Robbe et al. 2002; Taverna and Pennartz 2003; Thomas et al. 2001), PFC (Otani et al. 2003) and amygdala (Bauer and LeDoux 2004; Bauer et al. 2002; Fourcaudot et al. 2009; Huang and Kandel 1998, 2007; Humeau et al. 2003; Samson and Pare 2005). DA was shown to modulate plasticity in the striatum (Calabresi et al. 2007; Centonze et al. 2003; Kerr and Wickens 2001; Pawlak and Kerr 2008; Reynolds et al. 2001), the amygdala (Bissière et al. 2003), PFC (Huang et al. 2007; Kolomiets et al. 2009) and hippocampus (Frey et al. 1989, 1990, 1991; Otmakhova and Lisman 1996, 1998; Gurden et al. 1999) but not in the NAc. Behaviorally, it was also shown that pharmacological manipulations of DA levels in the brain shortly after performance can alter memory consolidation or reconsolidation (Dalley et al. 2005; Fenu and Di Chiara 2003; Hernandez et al. 2005; McGaugh 2002; Robertson and Cohen 2006).

Various reward-modulated plasticity models have been proposed (Florian 2007; Izhikevich 2007; Roberts et al. 2008; Thivierge et al. 2007). Most of these models use a form of synaptic modification known as spike timing dependent plasticity (STDP), which is based on the temporal relationship between pre-and postsynaptic activation (see Fig. 5b; Bi and Poo 1998; Markram et al. 1997). According to the STDP model, if a presynaptic spike occurs before a postsynaptic spike, long term potentiation (LTP) will occur, whereas the reverse order of pre- and postsynaptic spikes will induce long term depression (LTD). For plasticity to occur, the pre- and postsynaptic activation must happen within a short temporal window (less than 40 ms). The magnitude of the change in synaptic strengths depends on the time difference between pre- and postsynaptic activation and the direction of change depends on the order between pre-and postsynaptic activation.

In most reward-modulated STDP models, an RL component updates the synaptic weights based on a reward signal, eligibility trace and learning curve (Florian 2007; Izhikevich 2007; Roberts et al. 2008; Thivierge et al. 2007). A decaying eligibility trace determines which synapses are potentiated by DA and by how much depending on the size of the time interval between the stimulus and the reward (in this case, the time interval between the pre- and postsynaptic activations). The conceptual rationale is that in the early phases of learning, reward delivery leads to an increase activity in the VTA/SNc, thereby inducing DA release at target areas. DA will however only affect plasticity of the synapses that were active during the presentation of the predictive cue. This is because the eligibility trace has specifically 'tagged' the active synapses that underwent STDP, thereby solving the 'credit assignment' problem. In partial support for this model, Otmakhova and Lisman (1998) have shown that DA application after high-frequency stimulation leading to LTP can prevent the depotentiation of the synapses. Thus DA does not further potentiate synaptic plasticity as modeled, but prevents depotentiation leading to a similar overall result of having specific potentiated synapses. The applicability of the reward-modulated model was further tested with computer simulations using networks of leaky integrate-and-fire (LIF) neurons by Legenstein et al. (2008). The simulations showed the ability of this type of rule to predict spike times (rather than stimulus delivery times). As an alternative to STDP, Xie and Seung (2004) proposed a different learning rule based on the irregular nature of neural spiking and a reward-dependent eligibility trace. Other RL models demonstrated how neural networks could be trained to fire toward an optimum (Baras and Meir 2007; Potjans et al. 2009).

The effect of DA on corticostriatal synaptic plasticity is complex (see Reynolds and Wickens 2002 for a review). For instance, the amount of DA released at the time of corticostrial activation influences the direction of the plasticity, low levels inducing LTD and higher levels LTP (see Fig. 6). A recent computational study represented the reward value by the change in extracellular concentration of DA relative to baseline levels (Thivierge et al. 2007). In this study, the extracellular DA dynamics due to DA diffusion as well as DA degradation were calculated. The former was calculated based on the Michaelis–Menten equation and the latter was based on diffusion equations, similar to the 'kick and relax' model of Montague et al. (2004b) presented in the previous section. The change in DA concentration was then introduced multiplicatively into the STDP learning rule to simulate experimental data. The change in synaptic weight was expressed as

$$\Delta w = F(\Delta t)\left([D(t)]_{\text{total}} - b\right)$$

with F representing the STDP rule, D the instantaneous DA concentration and b the baseline DA concentration. This multiplication of the STDP with the change in DA concentration allowed for the modeling of the biphasic effect

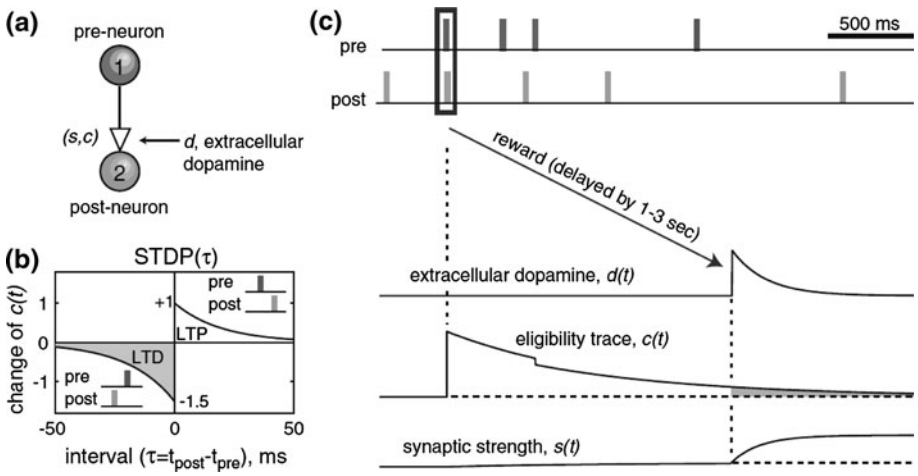

Fig. 5 A reward-modulated STDP model. **a** Schematic of DA modulation of corticostriatal synaptic plasticity. **b** STDP model illustrating the change in amplitude of the eligibility trace c(t), depending on the timing of pre- and postsynaptic activation τ. **c** Change in synaptic strength s(t) following overlapping STDP-induced eligibility trace activation and increased DA levels d(t) caused by reward delivery. Reproduced with permission from Izhikevich (2007)

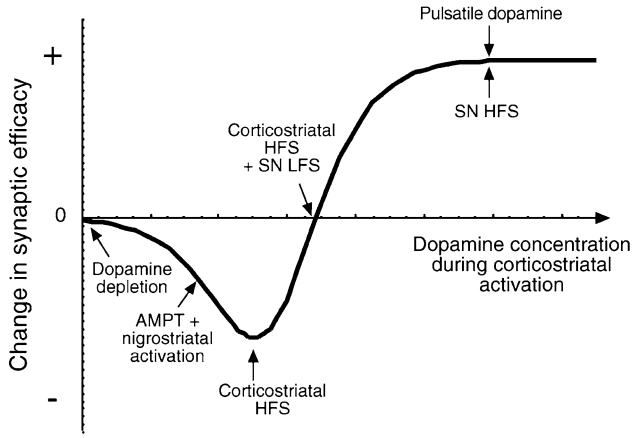

Fig. 6 The modulation corticostriatal plasticity by DA depends on its concentration. Low DA concentration during cortical high-frequency stimulation (HFS) leads to LTD, whereas higher DA concentration during HFS leads to LTP. Reproduced with permission from Reynolds and Wickens (2002)

of DA on corticostriatal synaptic plasticity (Fig 6, reviewed in Reynolds and Wickens 2002).

More complex interactions of DA and corticostriatal activation were also captured in the model of Frank et al. (2007; Fig. 7). DA bursts during positive prediction errors increase synaptic connection weights in subpopulations of Go-D1 cells that are concurrently excited by corticostriatal glutamatergic input, while also decreasing weights in NoGo-D2 cells. This enables the network to be more likely to repeat actions that are associated with positive reward prediction errors (Frank 2005), and is consistent with recent studies of synaptic plasticity showing D1-dependent LTP in striatonigral cells and D2-dependent LTD in striatopallidal cells (Shen et al. 2008). Moreover, in the model, DA pauses during negative prediction errors releasing NoGo cells from the tonic inhibition of DA onto high-affinity D2 receptors. The resulting transient activity increase in those NoGo cells that are concurrently excited by corticostriatal glutamatergic input is associated with activity-dependent synaptic strengthening (LTP), such that the system is more likely to suppress the maladaptive action in future encounters with the same sensory event. This effect is also consistent with studies in which a lack of D2 receptor stimulation (a "pause" in DA release) was associated with LTP in striatopallidal cells (Shen et al. 2008). The model has been used to simulate the effects of Parkinson's disease and pharmacological manipulations on learning and decision making processes in humans, and predicted that learning from positive and negative outcomes would be modulated in opposite directions by such treatments, as confirmed empirically (Frank et al. 2004; for review see Cohen and Frank 2009). More recently, the same model has accounted for some of the pharmacological effects of D2 receptor antagonists on the acquisition, extinction and renewal of motivated behavior in rats (Wiecki et al. 2009). Of course, several biophysical details are omitted in these models of corticostriatal circuitry. For example, future modeling work should incorporate STDP within the context of the basal ganglia circuitry to examine the role of spike-timing, together with the roles of other neurochemicals, such as adenosine, on corticostriatal plasticity (Shen et al. 2008).

The influence of dopamine on synaptic plasticity depends on numerous factors including the targeted cell type, the specific receptors activated and the brain region under consideration (Bissière et al. 2003; Calabresi et al. 2007; Otani et al. 2003; Shen et al. 2008; Wickens 2009; Yao et al. 2008). Reinforcement learning results from the dynamical interaction between numerous brain networks mediating the acquisition of conditioning tasks and action selection. We next present RL models at the system level.

Behavioral level: system level models of dopamine modulation of action selection

In reinforcement, numerous brain regions contribute to the association of cues with rewards, the learning of the actions that lead to these rewards and the adaptation to changes in reward values or in task contingencies (Dayan and Balleine 2002). The difficulty in associating reward processing with neurotransmitter systems is that typically these systems have very diffuse and often reciprocal projections to many if not all parts of the brain, so that a general understanding of the neural substrate of reward processing amounts to understanding the entire brain. Indeed, besides the VTA/SNc system, a number of brain regions have been shown to also respond to rewards or reward predictive cues, including the amygdala (Hatfield et al. 1996; Holland and Gallagher 1999) and orbitofrontal cortex (Furuyashiki and Gallagher 2007; Furuyashiki et al. 2008; Gallagher et al. 1999; O'Doherty et al. 2002, 2003; Schoenbaum et al. 1999; Tremblay and Schultz 2000a, b). A few behavioral models of incentive leaning are presented here. These models are based on either Q-learning or on the actor-critic algorithms, both representing extensions of the TD models that incorporate action selection. One interesting feature of these models is that they all require the division of the task into several modules in order to account for the various behavioral characteristics observed experimentally. This further emphasizes the need for parallel networks underlying reinforcement.

Q-learning algorithm and the actor-critic model

The TD model relies on differences between expected and received rewards of temporally separated events or states.

In and of itself, this model is insufficient to explain how new behaviors emerge as a result of conditioning, or how goal-directed actions are performed. To that aim, extensions of the TD model have been designed to incorporate action selection. Two frequently used theoretical frameworks are Q-learning and the actor-critic algorithms. Both paradigms use the TD error to update the state value. Q-learning is based on the TD algorithm, and optimizes the long term value of performing a particular action in a given state by generating and updating a state-action value function Q (Sutton and Barto 1998; Watkins and Dayan 1992). This model assigns a Q-value for each action-state pair (rather than simply for each state as in standard TD). For an action a_t performed at given state s_t, this value can be expressed as

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(\delta(t))$$

where the parameter α is the learning rate as in the TD rule. The prediction error $\delta(t)$ is expressed as

$$\delta(t) = R + \gamma \max_i(s_{t+1}, a_i) - Q(s_t, a_t)$$

and can be interpreted as the difference between the previous estimate $Q(s_t,a_t)$ and the new estimate after taking action a_t. The new estimate incorporates any experienced reward R together with the expected future reward (discounted by γ), assuming that the action with greatest Q value is taken in the subsequent state s_{t+1}. The value of the state-action pair therefore increases only if $\delta(t)$ is greater than 0, i.e. if the new estimate is greater than the previous estimate (due to an unexpected reward R or a higher expectation of future reward by having achieved state s_{t+1}). If so, the value of the pair will increase and more accurately reflect the expected future reward for taking action a, such that this action will be more likely to be selected again the next time the agent will be in this particular state s_t.

The probability of selecting a given action a in state s is typically a sigmoidal function comparing the Q values of that action to all others:

$$P(a_i) = \exp(\beta Q a_i) / \sum_j \exp\left(\beta Q a_j\right),$$

where β is the slope of the sigmoid determining the degree of exploration vs. exploitation; i.e., high values are associated with deterministic selection of the action with the highest Q value, whereas lower β values allow for probabilistic selection of lower Q values.

With proper parameter tuning, this Q learning model has the advantage of allowing for the exploration of potentially less beneficial options (i.e. sometimes selecting the actions with lower Q values in s_{t+1}) without impacting the Q-value of the prior state-action pair when the outcomes of these exploratory actions are not fruitful. An extension to the Q-learning algorithm was recently described by Hiraoka and colleagues (2009).

The actor-critic algorithm, on the other hand, optimizes the policy directly by increasing the probability that an action is taken when a reward is experienced without computing the values of actions for a given state (Barto 1995; Joel et al. 2002; Konda and Borkar 1999; Takahashi et al. 2008). The actor-critic model has been successfully used to understand how coordinated movements arise from reinforced behaviors (Fellous and Suri 2003; Suri et al. 2001).

How do neurons integrate information about rewards and sensory stimuli to produce coordinated outputs to the thalamus and induce movement? In this formulation, the 'critic' implements a form of TD learning related to dopamine neuron activity, while the 'actor' implements the learning occurring during sensory-motor associations (See Fig. 1). Otherwise stated, the critic estimates the state values based on the TD learning rule (without computing the value of particular actions) and the actor updates the policy using the reward prediction error δ(t) resulting from achieving a particular state. Accordingly, the policy parameter is updated as follows

$$\pi(s_t, a_t) = \pi(s_t, a_t) + \alpha\delta(t).$$

where the parameter α is the learning rate, as in the TD rule and π is the policy.

The algorithms above, in particular Q learning, have been applied to understand reinforcement-based decision making in humans (e.g., O'Doherty et al. 2004; Frank et al. 2007; Voon et al. 2010). In two recent studies, human subjects learned to select among symbols presented on a computer screen, where the selection of some symbols were associated with a greater probability of reward (points or financial gains), and others were associated with a greater probability of negative outcome (losses) (Frank et al. 2007; Voon et al. 2010). A modified Q-learning algorithm was used to model the probability that an individual would select a particular symbol in a given trial as a function of reinforcement history. This abstract formulation was further related to the posited mechanisms embedded in the neural model of the basal ganglia Go/NoGo dopamine system described earlier. Thus, rather than having a single learning rate to adapt Q-values for action as a function of reward prediction error, this model utilized separate learning rates for positive and negative prediction errors, corresponding to striatal D1 and D2 receptor mechanisms respectively. The authors found that the best-fitting learning rate parameters which provided the best objective fit to each participant's choices (using maximum likelihood), were predicted by variations in genetic factors affecting striatal D1 and D2 receptor mechanisms (Frank et al. 2007). Others have used a similar approach to show that these learning rate parameters are differentially modulated by dopaminergic medications administered to

patients with Parkinson's (Voon et al. 2010). Finally, another recent study replicated the striatal D1/D2 genetic variation effects on positive and negative learning rates in the context of a task that required participants to adjust response times to maximize rewards, which had also been shown to be sensitive to dopaminergic manipulation (Moustafa et al. 2008; Frank et al. 2009). Moreover, in addition to learning, this model also included a term to account for exploration, which was predicted to occur when participants were particularly uncertain about the reward statistics due to insufficient sampling. It was found that individual differences in this uncertainty-driven exploration were strongly affected by a third genetic variation known to regulate prefrontal cortical, rather than striatal, dopamine. Although no mechanistic model has been proposed to account for this latter effect, this result highlights again that dopamine may be involved in multiple aspects of reinforcement learning by acting via different receptors, time courses, and brain areas.

Once behavioral responses are learned, it can be challenging to modify behaviors to changes in stimulus-reward contingencies. Learning that a cue no longer predicts the delivery of a reward involves learning to refrain making a particular action because it no longer has a purpose. This form of learning is called extinction. Extinction is not the result of unlearning or forgetting (Bouton 2004); rather, it is a form of learning that involves different neural networks from those recruited during the original learning of the stimulus-outcome associations. Extinction is associated with spontaneous recovery when placed back in the original context and the action renewal rate is faster than the rate of the initial acquisition. The TD model would represent this situation as a decrease in the prediction error, thereby indicating that no reward is expected from presentation of the stimulus. Since the TD model in itself cannot incorporate all the particularities of extinction learning, Redish and colleagues (2007) built a model of extinction that can also explain spontaneous recovery and renewal rate. This model is divided in two components; a TD component acquired the state value and a situation-categorization component (state-classification) differentiated the associative phase from the extinction phase. The TD component was implemented using the Q-learning algorithm to measure the value of taking an action in a particular state (Sutton and Barto 1998). During extinction, since rewards are omitted, the prediction error decreased and the probability of changing state increased. The state classification component allowed for the alternation between parallel state spaces. A threshold determined whether a new state should be created. A low tonic value-prediction error term was represented by an exponentially decaying running average of recent negative prediction-error signals. Low values produced a change in state representation, for the 'extinction state'. A subsequent positive value-prediction error term modeled renewal (spontaneous recovery when the animal is returned to the first context). While this study did not explicitly model neural mechanisms, but behavioral data, the requirement of having parallel state representations illustrated the way parallel neural networks might be working together to form different stimulus representations depending on the situation. Other related examples of extinction and renewal have been modeled in basal ganglia networks to account for the effects of dopaminergic pharmacological agents on the development and extinction of motor behaviors in rats (Wiecki et al. 2009). Finally, fear conditioning is another well documented conditioning paradigm amenable to extinction studies (reviewed in Ehrlich et al. 2009; Quirk and Mueller 2007). The computational role of dopamine in this paradigm is however still unclear.

The schedule of reinforcement during training is critical for maintaining goal-directed response. With extended training, responses become habitual, thereby insensitive to changes in the tasks contingencies or reward values. This change with training from goal-directed to habitual behavior is correlated with changes in the brain networks involved in these behaviors. The dorsolateral striatum is involved in habitual responding while the prefrontal cortex mediates goal-directed behavior (Yin et al. 2008). Using outcome uncertainty as the distinguishing factor mediating goal-directed versus habitual behaviors, Daw et al. created two models, one for each type of behavior (Daw et al. 2005). These models, based on Q-learning algorithms, were used to represent each system; the dorsolateral striatum network was represented by a model-free 'cache' (slow learning integrative TD-like) system, and the prefrontal network by a model-based 'tree-search' system. The uncertainty was introduced in the model-free system using Bayesian probability distributions in the Q-value estimates. When a reward is devalued through satiation, only the Q-value of the state-action pair associated with the reward state is reduced in the model-free system, such that extended training would be required to adapt all Q values leading up to this state. In contrast, in the model-based 'tree search' system, a change in the reward function immediately impacted that of all other states. This allowed for the 'tree-search' model to remain goal-directed and adapt to changes in the reward value, but not the 'cache' system which continued to respond despite the devaluation of the reinforcer, thereby simulating habit responding.

Most decision making experiments rely on learning to perform visuomotor tasks a type of cross-modal learning. In a system level model, a network representing the dorsolateral prefrontal cortex and the anterior basal ganglia was responsible for receiving visual inputs, and a second network representing the supplementary motor area and the

posterior basal ganglia was responsible for generating a motor output (Nakahara et al. 2001). Both networks were modeled differently with the visual network containing working memory capability allowing for rapid initial learning of the task and for task reversal. The motor network had a slower learning rate, but once the task was learned, it allowed for faster and more accurate motor actions than the visual network. Parallel loops were modeled for each network and a coordinator module was added to adjust the respective contribution of each network. The general model followed the actor-critic architecture mentioned above and in Fig. 1, in which the critic generated prediction errors by changing the weight matrices of the visual and motor network (both are considered actors in this model). During learning, the weight matrices of each network changed so as to maximize the sum of future rewards. Simulations confirmed that these parallel networks fits the data more accurately in terms of learning rates and adaptation to changes in the task sequence than either of the two networks working independently.

The psychological theories of reinforcement highlight the fact that this form of learning is intricate but amenable to modeling. The acquisition and performance of goal-directed behaviors are known to be sensitive to the reinforcement schedule, stimulus-reward contingencies, context and reward values. Due to their complexity, and as presented here, computational models need the integration of parallel computational structures to explain how these behaviors adapt to variable experimental and environmental conditions.

Conclusions

Reinforcement learning is at the basis of many forms of adaptation of the organism to its environment. More than simply a form of stimulus-outcome association, RL is intrinsically designed to optimize. Optimization can be that of an amount of rewards, a speed of learning, or that of robustness to perturbing events. This feature of the RL paradigm is puzzling because RL relies on relatively information-poor feedback signals. We reviewed some of the main models of RL and outlined how they can be understood and implemented by the nervous system at multiple levels, from synaptic to system. Rare are the models that attempt multi-level integration. It is however in the integration between these levels that lay the biological success of RL. Finally, we note that this review did not cover the many artificial intelligence approaches that have attempted to address reinforcement at more cognitive and abstract levels. Future work should be done to bridge and contrast the artificial intelligence and computational neuroscience approaches.

Acknowledgments The authors wish to thank Dr. Ian Fasel, Nathan Insel and Minryung Song for useful comments on the manuscript. R.D.S. was supported by the Canadian Institute of Health Research SIB 171357.

References

Abercrombie ED, Keefe KA et al (1989) Differential effect of stress on in vivo dopamine release in striatum, nucleus accumbens, and medial frontal cortex. J Neurochem 52:1655–1658

Arbuthnott GW, Wickens J (2007) Space, time and dopamine. Trends Neurosci 30:62–69

Baras D, Meir R (2007) Reinforcement learning, spike-time-dependent plasticity, and the bcm rule. Neural Comput 19:2245–2279

Barto AG (1995) Adaptive critics and the basal ganglia. Models of information processing in the basal ganglia. 215–232

Barto AG, Sutton RS, Anderson C (1983) Neuron-like adaptive elements that can solve difficult learning control problems, IEEE transactions on systems, man, and cybernetics. SMC 13:834–846

Bauer EP, LeDoux JE (2004) Heterosynaptic long-term potentiation of inhibitory interneurons in the lateral amygdala. J Neurosci 24:9507–9512

Bauer EP, Schafe GE, LeDoux JE (2002) NMDA receptors and L-type voltage-gated calcium channels contribute to long-term potentiation and different components of fear memory formation in the lateral amygdala. J Neurosci 22:5239–5249

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141

Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. J Neurophysiol 98:1428–1439

Bergstrom BP, Garris PA (2003) Passive stabilization of striatal extracellular dopamine across the lesion spectrum encompassing the presymptomatic phase of Parkinson's disease: a voltammetric study in the 6-OHDA-lesioned rat. J Neurochem 87:1224–1236

Berke JD, Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. Neuron 25:515–532

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) 191:391–431

Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Brain Res Rev 28:309–369

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

Bi G, Poo M (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18:10464–10472

Bissière S, Humeau Y, Luthi A (2003) Dopamine gates ltp induction in lateral amygdala by suppressing feedforward inhibition. Nature Neurosci 6:587–592

Bouton ME (2004) Context and behavioral processes in extinction. Learn Mem 11:485–494

Calabresi P, Picconi B, Tozzi A, Di Filippo M (2007) Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends Neurosci 30:211–219

Camerer CF (2008) Neuroeconomics: opening the gray box. Neuron 60:416–419

Canavan AG, Passingham RE et al (1989) The performance on learning tasks of patients in the early stages of Parkinson's disease. Neuropsychologia 27:141–156

Cardinal RN, Parkinson JA et al (2002) Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev 26:321–352

Centonze D, Grande C et al (2003) Distinct roles of D1 and D5 dopamine receptors in motor activity and striatal synaptic plasticity. J Neurosci 23:8506–8512

Cepeda C, Levine MS (1998) Dopamine and *N*-methyl-D-aspartate receptor interactions in the neostriatum. Dev Neurosci 20:1–18

Chuhma N, Choi WY et al (2009) Dopamine neuron glutamate cotransmission: frequency-dependent modulation in the mesoventromedial projection. Neuroscience 164:1068–1083

Cohen MX, Frank MJ (2009) Neurocomputational models of basal ganglia function in learning, memory and choice. Behav Brain Res 199:141–156

Dalley JW, Laane K et al (2005) Time-limited modulation of appetitive Pavlovian memory by D1 and NMDA receptors in the nucleus accumbens. Proc Natl Acad Sci USA 102:6189–6194

Daniel JA, Galbraith S et al (2009) Functional heterogeneity at dopamine release sites. J Neurosci 29:14670–14680

Daw ND, Niv Y et al (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711

Daw ND, Courville AC, Touretzky DS (2006) Representation and timing in theories of the dopamine system. Neural Comput 18:1637–1677

Dayan P, Abbott LF (2001) Theoretical neuroscience. Computational and mathematical modeling of neural systems. The MIT Press, Cambridge

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8:429–453

Di Chiara G (2002) Nucleus accumbens shell and core dopamine: differential role in behavior and addiction. Behav Brain Res 137:75–114

Doya K, Samejima K et al (2002) Multiple model-based reinforcement learning. Neural Comput 14:1347–1369

Ehrlich I, Humeau Y et al (2009) Amygdala inhibitory circuits and the control of fear memory. Neuron 62:757–771

Everitt BJ, Parkinson JA et al (1999) Associative processes in addiction and reward. The role of amygdala-ventral striatal subsystems. Ann N Y Acad Sci 877:412–438

Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 8:1481–1489

Fellous J-M, Suri RE (2003) The roles of dopamine. The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 361–365

Fenu S, Di Chiara G (2003) Facilitation of conditioned taste aversion learning by systemic amphetamine: role of nucleus accumbens shell dopamine D1 receptors. Eur J Neurosci 18:2025–2030

Fibiger HC, Phillips AG (1986) Reward, motivation, cognition, psychobiology of mesotelencephalic dopamine systems. Handbook of physiology—The nervous system IV. F E Bloom Baltimore, Williams and Wilkins, 647–675

Fields HL, Hjelmstad GO et al (2007) Ventral tegmental area neurons in learned appetitive behavior and positive reinforcement. Annu Rev Neurosci 30:289–316

Floresco SB, West AR et al (2003) Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nature Neurosci 6:968–973

Florian RV (2007) Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Comput 19: 1468–1502

Fourcaudot E, Gambino F et al (2009) L-type voltage-dependent Ca(2+) channels mediate expression of presynaptic LTP in amygdala. Nat Neurosci 12:1093–1095

Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17:51–72

Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in Parkinsonism. Science 306:1940–1943

Frank MJ, Moustafa AA et al (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci USA 104:16311–16316

Frank MJ, Doll BB et al (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neurosci 12:1062–1068

Freeman WJ (2007) Definitions of state variables and state space for brain-computer interface: part 1. Multiple hierarchical levels of brain function. Cogn Neurodyn 1:3–14

Frey U, Hartmann S et al (1989) Domperidone, an inhibitor of the D2-receptor, blocks a late phase of an electrically induced long-term potentiation in the CA1-region in rats. Biomed Biochim Acta 48:473–476

Frey U, Schroeder H et al (1990) Dopaminergic antagonists prevent long-term maintenance of posttetanic LTP in the CA1 region of rat hippocampal slices. Brain Res 522:69–75

Frey U, Matthies H et al (1991) The effect of dopaminergic D1 receptor blockade during tetanization on the expression of long-term potentiation in the rat CA1 region in vitro. Neurosci Lett 129:111–114

Furuyashiki T, Gallagher M (2007) Neural encoding in the orbitofrontal cortex related to goal-directed behavior. Ann N Y Acad Sci 1121:193–215

Furuyashiki T, Holland PC et al (2008) Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behavior. J Neurosci 28:5127–5138

Gall CM, Hendry SH et al (1987) Evidence for coexistence of GABA and dopamine in neurons of the rat olfactory bulb. J Comp Neurol 266:307–318

Gallagher M, McMahan RW et al (1999) Orbitofrontal cortex and representation of incentive value in associative learning. J Neurosci 19:6610–6614

Glimcher PW, Rustichini A (2004) Neuroeconomics: the consilience of brain and decision. Science 306:447–452

Goto Y, Grace AA (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. Nat Neurosci 8:805–812

Grace AA (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience 41:1–24

Gurden H, Tassin JP et al (1999) Integrity of the mesocortical dopaminergic system is necessary for complete expression of in vivo hippocampal–prefrontal cortex long-term potentiation. Neuroscience 94:1019–1027

Hatfield T, Han JS et al (1996) Neurotoxic lesions of basolateral, but not central, amygdala interfere with pavlovian second-order conditioning and reinforcer devaluation effects. J Neurosci 16:5256–5265

Hernandez PJ, Andrzejewski ME et al (2005) AMPA/kainate, NMDA, and dopamine D1 receptor function in the nucleus accumbens core: a context-limited role in the encoding and consolidation of instrumental memory. Learn Mem 12:285–295

Hikosaka O, Bromberg-Martin E et al (2008) New insights on the subcortical representation of reward. Curr Opin Neurobiol 18:203–208

Hiraoka K, Yoshida M, Mishima T (2009) Parallel reinforcement learning for weighted multi-criteria model with adaptive margin. Cogn Neurodyn 3:17–24

Holland PC, Gallagher M (1999) Amygdala circuitry in attentional and representational processes. Trends Cogn Sci 3:65–73
Huang Y-Y, Kandel ER (1998) Postsynaptic induction and PKA-dependent expression of LTP in the lateral amygdala. Neuron 21:169–178
Huang Y-Y, Kandel ER (2007) Low-frequency stimulation induces a pathway-specific late phase of LTP in the amygdala that is mediated by PKA and dependent on protein synthesis. Learn Mem 14:497–503
Huang CC, Lin HJ et al (2007) Repeated cocaine administration promotes long-term potentiation induction in rat medial prefrontal cortex. Cereb Cortex 17:1877–1888
Humeau Y, Shaban H et al (2003) Presynaptic induction of heterosynaptic associative plasticity in the mammalian brain. Nature 426:841–845
Ikemoto S, Panksepp J (1999) The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res Brain Res Rev 31:6–41
Izhikevich EM (2007) Solving the distal reward problem through linkage of stdp and dopamine signaling. Cereb Cortex 17:2443–2452
Joel D, Niv Y et al (2002) Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw 15:535–547
Kaelbling LP, Littman ML et al (1996) Reinforcement learning: a survey. JAIR 4:237–285
Kelley AE, Berridge KC (2002) The neuroscience of natural rewards: relevance to addictive drugs. J Neurosci 22:3306–3311
Kerr JN, Wickens JR (2001) Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. J Neurophysiol 85:117–124
Kilpatrick MR, Rooney MB et al (2000) Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. Neuroscience 96:697–706
Knowlton BJ, Mangels JA et al (1996) A neostriatal habit learning system in humans. Science 273:1399–1402
Kobayashi S, Schultz W (2008) Influence of reward delays on responses of dopamine neurons. J Neurosci 28:7837–7846
Kolomiets B, Marzo A et al (2009) Background dopamine concentration dependently facilitates long-term potentiation in rat prefrontal cortex through postsynaptic activation of extracellular signal-regulated kinases. Cereb Cortex 19:2708–2718
Kombian SB, Malenka RC (1994) Simultaneous LTP of non-NMDA- and LTD of NMDA-receptor-mediated responses in the nucleus accumbens. Nature 368:242–246
Konda VR, Borkar VS (1999) Actor-critic—type learning algorithms for markov decision processes. SIAM J Control Optim 38:94–123
Koob GF (1992) Drugs of abuse: anatomy, pharmacology and function of reward pathways. Trends Pharmacol Sci 13:177–184
Kötter R, Wickens J (1995) Interactions of glutamate and dopamine in a computational model of the striatum. J Comput Neurosci 2:195–214
Kroener S, Chandler LJ et al (2009) Dopamine modulates persistent synaptic activity and enhances the signal-to-noise ratio in the prefrontal cortex. PLoS One 4:e6507
Lapish CC, Kroener S et al (2007) The ability of the mesocortical dopamine system to operate distinct temporal modes. Psychopharmacology (Berl) 191:609–626
Legenstein R, Pecevski D et al (2008) A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. PLoS Comput Biol 4:e1000180
Li Y, Kauer JA (2004) Repeated exposure to amphetamine disrupts dopaminergic modulation of excitatory synaptic plasticity and neurotransmission in nucleus accumbens. Synapse 51:1–10
Lindvall O, Bjorklund A (1978) Anatomy of the dopaminergic neuron systems in the rat brain. Adv Biochem Psychopharmacol 19:1–23
Ljungberg T, Apicella P et al (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophys 67:145–163
Louilot A, Le Moal M et al (1986) Differential reactivity of dopaminergic neurons in the nucleus accumbens in response to different behavioral situations. An in vivo voltammetric study in free moving rats. Brain Res 397(2):395–400
Mackintosh NJ (1983) Conditioning and associative learning. Oxford University Press, New York
Maher BJ, Westbrook GL (2008) Co-transmission of dopamine and GABA in periglomerular cells. J Neurophysiol 99:1559–1564
Markram H, Lübke J et al (1997) Regulation of synaptic effocacy by coincidence of postsynaptic aps and epsps. Science 275:213–215
Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459:837–841
McGaugh JL (2002) Memory consolidation and the amygdala: a systems perspective. Trends Neurosci 25:456
Mirenowicz J, Schultz W (1994) Importance of unpredictability for reward responses in primate dopamine neurons. J Neurophys 72:1024–1027
Mirenowicz J, Schultz W (1996) Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature 379:449–451
Montague PR, Dayan P et al (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci 16:1936–1947
Montague PR, Hyman SE et al (2004a) Computational roles for dopamine in behavioural control. Nature 431:760–767
Montague PR, McClure SM et al (2004b) Dynamic gain control of dopamine delivery in freely moving animals. J Neurosci 24:1754–1759
Morris G, Arkadir D et al (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 43:133–143
Morris G, Nevet A et al (2006) Midbrain dopamine neurons encode decisions for future action. Nature Neurosci 9:1057–1063
Moustafa AA, Cohen MX et al (2008) A role for dopamine in temporal decision making and reward maximization in parkinsonism. J Neurosci 28:12294–12304
Nakahara H, Doya K et al (2001) Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences—a computational approach. J Cogn Neurosci 13:626–647
Nakahara H, Itoh H et al (2004) Dopamine neurons can represent context-dependent prediction error. Neuron 41:269–280
Nicola SM, Surmeier J et al (2000) Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. Annu Rev Neurosci 23:185–215
O'Doherty JP, Deichmann R et al (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826
O'Doherty J, Dayan P et al (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337
O'Doherty J, Dayan P et al (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454
Otani S, Daniel H et al (2003) Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb Cortex 13:1251–1256
Otmakhova NA, Lisman JE (1996) D1/d5 dopamine receptor activation increases the magnitude of early long-term potentiation at ca1 hippocampal synapses. J Neurosci 16:7478–7486

Otmakhova NA, Lisman JE (1998) D1/d5 dopamine receptors inhibit depotentiation at ca1 synapses via camp-dependent mechanism. J Neurosci 18:1270–1279
Pan WX, Schmidt R et al (2005) Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. J Neurosci 25:6235–6242
Pawlak V, Kerr JN (2008) Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. J Neurosci 28:2435–2446
Pennartz CM, Ameerun RF et al (1993) Synaptic plasticity in an in vitro slice preparation of the rat nucleus accumbens. Eur J Neurosci 5:107–117
Phillips PE, Stuber GD et al (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422:614–618
Potjans W, Morrison A, Diesmann M (2009) A spiking neural network model of an actor-critic learning agent. Neural Comput 21:301–339
Quirk GJ, Mueller D (2007) Neural mechanisms of extinction learning and retrieval. Neuropsychopharmacology 33:56–72
Redish AD, Jensen S et al (2007) Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. Psychol Rev 114: 784–805
Reynolds JN, Wickens JR (2002) Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw 15:507–521
Reynolds JN, Hyland BI et al (2001) A cellular mechanism of reward-related learning. Nature 413:67–70
Robbe D, Alonso G et al (2002) Role of p/q-Ca2+ channels in metabotropic glutamate receptor 2/3-dependent presynaptic long-term depression at nucleus accumbens synapses. J Neurosci 22:4346–4356
Robbins TW, Everitt BJ (1996) Neurobehavioural mechanisms of reward and motivation. Curr Opin Neurobiol 6:228–236
Roberts PD, Santiago RA et al (2008) An implementation of reinforcement learning based on spike timing dependent plasticity. Biol Cybern 99:517–523
Robertson EM, Cohen DA (2006) Understanding consolidation through the architecture of memories. Neuroscientist 12:261–271
Robinson TE, Berridge KC (2008) Review. The incentive sensitization theory of addiction: some current issues. Philos Trans R Soc Lond B Biol Sci 363:3137–3146
Robinson DL, Heien ML et al (2002) Frequency of dopamine concentration transients increases in dorsal and ventral striatum of male rats during introduction of conspecifics. J Neurosci 22:10477–10486
Roesch MR, Calu DJ et al (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neurosci 10:1615–1624
Roitman MF, Stuber GD et al (2004) Dopamine operates as a subsecond modulator of food seeking. J Neurosci 24:1265–1271
Rolls ET (2000) Precis of the brain and emotion. Behav Brain Sci 23:177–191 discussion 192–233
Romo R, Schultz W (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J Neurophysiol 63:592–606
Samejima K, Doya K et al (2003) Inter-module credit assignment in modular reinforcement learning. Neural Netw 16:985–994
Samson RD, Pare D (2005) Activity-dependent synaptic plasticity in the central nucleus of the amygdala. J Neurosci 25:1847–1855
Satoh T, Nakai S et al (2003) Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci 23:9913–9923
Schimchowitsch S, Vuillez P et al (1991) Systematic presence of GABA-immunoreactivity in the tubero-infundibular and tubero-hypophyseal dopaminergic axonal systems: an ultrastructural immunogold study on several mammals. Exp Brain Res 83:575–586
Schoenbaum G, Chiba AA et al (1999) Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. J Neurosci 19:1876–1884
Schultz W (1992) Activity of dopamine neurons in the behaving primate. Semi Neurosci 4(2):129–138
Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1–27
Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263
Schultz W, Dayan P et al (1997) A neural substrate of prediction and reward. Science 275:1593–1599
Servan-Schreiber D, Printz H et al (1990) A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. Science 249:892–895
Seymour B, O'Doherty JP et al (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667
Shen W, Flajolet M et al (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848
Suri RE, Bargas J et al (2001) Modeling functions of striatal dopamine modulation in learning and planning. Neuroscience 103:65–85
Surmeier DJ, Ding J et al (2007) D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends Neurosci 30:228–235
Sutton RS (1984) Temporal credit assignment in reinforcement learning Ph.D. dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA. Published as COINS Technical Report 84-2
Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44
Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. MIT Press, Cambridge
Sutton RS, Barto AG (1998) Reinforcement learning, an introduction. MIT Press, Cambridge
Takahashi Y, Schoenbaum G et al (2008) Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front Neurosci 2:86–99
Taverna S, Pennartz CM (2003) Postsynaptic modulation of AMPA- and NMDA-receptor currents by group III metabotropic glutamate receptors in rat nucleus accumbens. Brain Res 976:60–68
Thivierge JP, Rivest F et al (2007) Spiking neurons, dopamine, and plasticity: timing is everything, but concentration also matters. Synapse 61:375–390
Thomas MJ, Beurrier C et al (2001) Long-term depression in the nucleus accumbens: a neural correlate of behavioral sensitization to cocaine. Nat Neurosci 4:1217–1223
Tremblay L, Schultz W (2000a) Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. J Neurophysiol 83:1877–1885
Tremblay L, Schultz W (2000b) Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. J Neurophysiol 83:1864–1876
Vijayraghavan S, Wang M et al (2007) Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. Nat Neurosci 10:376–384
Voon V, Reynolds B, Brezing C et al (2010) Impulsive choice and response in dopamine agonist-related impulse control behaviors. Psychopharmacology (Berl) 207:645–659
Waelti P, Dickinson A et al (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412: 43–48
Watkins C, Dayan P (1992) Q-learning. Mach Learning 8:279–292
White DJ (1993) Markov decision processes. Willey, New York

Wickens JR (2009) Synaptic plasticity in the basal ganglia. Behav Brain Res 199:119–128

Wickens JR, Horvitz JC et al (2007) Dopaminergic mechanisms in actions and habits. J Neuroscience 27:8181

Wiecki TV, Riedinger K et al (2009) A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: Context dependency, extinction, and renewal. Psychopharmacology 204:265–277

Wise RA (1996a) Addictive drugs and brain stimulation reward. Annu Rev Neurosci 19:319–340

Wise RA (1996b) Neurobiology of addiction. Curr Opin Neurobiol 6:243–251

Wise RA (2004) Dopamine, learning and motivation. Nat Rev Neurosci 5:483–494

Wise RA (2005) Forebrain substrates of reward and motivation. J Comp Neurol 493:115–121

Wise RA, Hoffman DC (1992) Localization of drug reward mechanisms by intracranial injections. Synapse 10:247–263

Wise RA, Rompre PP (1989) Brain dopamine and reward. Annu Rev Psychol 40:191–225

Worgotter F, Porr B (2005) Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. Neural Comput 17:245–319

Xie X, Seung HS (2004) Learning in neural networks by reinforcement of irregular spiking. Phys Rev E Stat Nonlin Soft Matter Phys 69:041909

Yao WD, Spealman RD et al (2008) Dopaminergic signaling in dendritic spines. Biochem Pharmacol 75:2055–2069

Yavich L, MacDonald E (2000) Dopamine release from pharmacologically distinct storage pools in rat striatum following stimulation at frequency of neuronal bursting. Brain Res 870:73–79

Yin HH, Ostlund SB et al (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks. Eur J Neurosci 28:1437–1448

Young AM, Joseph MH et al (1992) Increased dopamine release in vivo in nucleus accumbens and caudate nucleus of the rat during drinking: a microdialysis study. Neuroscience 48:871–876