Chapter 8
Understanding Addiction as a Pathological State of Multiple Decision Making Processes: A Neurocomputational Perspective

Mehdi Keramati, Amir Dezfouli, and Payam Piray

Abstract Theories of addiction in neuropsychology increasingly define addiction as a progressive subversion, by drugs, of the learning processes by which animals are equipped with, to adapt their behaviors to the ever-changing environment surrounding them. These normal learning processes, known as Pavlovian, habitual and goal-directed, are shown to rely on parallel and segregated cortico-striatal loops, and several computational models have been proposed in the reinforcement learning framework to explain the different and sometimes overlapping components of this network. In this chapter, we review some neurocomputational models of addiction originating from reinforcement learning theory, each of which explain addiction as a usurpation of one of the well-known models under the effect of addictive drugs. We try to show how each of these partially complete models can explain some behavioral and neurobiological aspects of addiction, and why it is necessary to integrate these models in order to have a more complete computational account for addiction.

8.1 Introduction

Addiction, including addiction to drugs of abuse, is defined as a compulsive orientation toward some certain behaviors, despite the heavy costs that might be followed (Koob and Le Moal 2005b). In the case of drug addiction, addicts are usually portrayed as people who seek and take drugs, even at the cost of adverse social, occupational and health consequences. Although a wide range of effects of drugs on different body and nervous system regions has been shown, it is progressively becoming accepted that the above definition of drug addiction arises from the pharmacological effects of drugs on the brain learning system, that is, the brain circuits involved in adaptively guiding animals' behaviors toward satisfying their needs (Everitt and Robbins 2005; Redish et al. 2008;

M. Keramati
Group for Neural Theory, Ecole Normale Superieure, Paris, France

A. Dezfouli (✉) · P. Piray
Neurocognitive Laboratory, Iranian National Center for Addiction Studies, Tehran University of Medical Sciences, Tehran, Iran
e-mail: a.dezfouli@ut.ac.ir

B. Gutkin, S.H. Ahmed (eds.), *Computational Neuroscience of Drug Addiction*,
Springer Series in Computational Neuroscience 10,
DOI 10.1007/978-1-4614-0751-5_8,

Belin et al. 2009). In fact, drugs of abuse are notorious for usurpation of the natural learning processes and consequently, understanding normal learning mechanisms has proven to be a prerequisite for understanding addiction as a pathological state of those underlying systems.

Conditioning literature in behavioral psychology has long studied animal behavior and has developed a rich and coherent framework for understanding associative learning by defining several components involved in decision making, most notably Pavlovian, habitual, and goal-directed systems (Dickinson and Balleine 2002). The neural underpinnings of these components and their competitive and collaborative interactions have also been well studied during the last 50 years (Balleine and O'Doherty 2010; Rangel et al. 2008), although there is still a long way to go. This psychological and neurobiological knowledge has paved the way for computational models of decision making to emerge. These models rephrase in a formal language, the developed concepts in the neuropsychology of decision making and thus, guarantee the coherency and self-consistency of the proposed computational theories, as well as quantitatively examining their validity using experimental data. The computational theory of reinforcement learning (RL) (Sutton and Barto 1998), which is the origin of all computational models reviewed in this chapter, is a putative formal framework that has captured many aspects of the psychological and neurobiological knowledge gathered around animal decision making. Within this framework, the "Q-learning" model explains the behavioral characteristics of the habitual process (Sutton and Barto 1998), which is believed to be neurally implemented in the sensorimotor cortico-striatal loop (Yin et al. 2004, 2008). The "actor-critic" models, on the other hand, explain collaboration between Pavlovian and habitual systems and are based on the integrity of limbic and sensorimotor loops (Joel et al. 2002). Finally, "dual-process" models, capture the interplay between habitual and goal-directed processes, and are based on the interaction between sensorimotor and associative loops, respectively (Daw et al. 2005; Keramati et al. 2011).

As addictive drugs are known to usurp the normal learning mechanisms, many of the computational models proposed to date for explaining addiction-like behaviors are based on the RL framework. In fact, each of the five computational models reviewed in this chapter (Redish 2004; Dezfouli et al. 2009; Dayan 2009; Piray et al. 2010; Keramati et al. 2011) explains addiction as a malfunction, due to the effect of drugs, of one of the variants of the RL theory mentioned above. As each model takes into account different, and sometimes overlapping components of the whole learning system, each of them can explain some limited, and sometimes overlapping, behavioral aspects of addiction.

In the following sections, we first briefly discuss some key concepts of the conditioning literature and its neural substrates. The main focus of the first section is on introducing Pavlovian and instrumental forms of associative learning and the multiple kinds of interaction between them, as well as the anatomically parallel and segregated closed loops in the cortico-basal ganglia system that underlie those different associative structures. Based on this literature, potential impairments in these systems induced by pharmacological effects of drugs, and their related behavioral

manifestations are explored in the next section. We then review five computational models of addiction, each of which has incorporated the pharmacological effects of drugs into a version of the computational theory of reinforcement learning. The first two models (Redish 2004; Dezfouli et al. 2009) are based on the Q-learning algorithm, which models the habitual decision making process. The second group of models (Dayan 2009; Piray et al. 2010) study the drug-induced pathological state of the actor-critic model, representing the interaction between Pavlovian and habitual process. And the last model (Keramati et al. 2011) relies on the dual-process theory of decision making. Finally, we discuss some open avenues for future theoretical efforts for explaining more behavioral and biological evidence on addiction in the RL framework.

8.2 Normal Decision Making Mechanism

Conditioning is an associative learning process by which animals learn to adapt their predictions and behaviors to the occurrence of different stimuli in the environment (Dickinson and Balleine 2002). This learning is made possible by representing the contingencies between different stimuli, responses, and outcomes, in brain associative structures. Psychologists have long made a distinction between Pavlovian and instrumental forms of conditioning. Pavlovian (or classical) conditioning is a form of associative learning where the animal learns that presentation of a neutral stimulus, known as conditioned stimulus (CS), predicts the occurrence of a salient event, known as unconditioned stimuli (US). For this reason, Pavlovian conditioning is also known as stimulus-stimulus (S-S) conditioning. Appearance of the US might evoke an innate, reflexive response called unconditioned response (UR). When this reflexive response is evoked by presenting the CS (which itself predicts the US), it is called a conditioned response (CR). Salivation in response to presentation or prediction of food is a famous example of conditioned or unconditioned responses, respectively. It is important to note that in Pavlovian conditioning, the animal has no control over the occurrence of events in the environment, but only observes. A computational model for learning these S-S associations is presented in Sect. 8.4.2.

In instrumental conditioning, in contrast, the animal learns to choose a sequence of actions so as to attain appetitive stimuli or to avoid aversive ones. At the early stages of exploring a new environment, the animal starts discovering the causal relations between specific actions and their consequent biologically significant outcomes. Based on this instrumental knowledge, at each state like s, the animal deliberates the consequences of different behavioral strategies and then, takes an action like a by which it reaches a desirable outcome like o. Regarding that this kind of instrumental behavior is aimed at gaining access to a certain outcome or goal, it is called goal-directed or stimulus-action-outcome (S-A-O) responding. A formal representation for this system is presented in Sect. 8.4.3. After the animal is extensively trained in the environment, it learns to habitually make a certain response, say a, whenever it finds itself in a certain state, like s, without considering the poten-

tial consequences that action might have. Not surprisingly, this type of instrumental behavior is called habitual or stimulus-response (S-R) responding. A computational model representing this type of learning is introduced in Sect. 8.4.1.

Although the three types of learning mechanisms (S-S, S-A-O, S-R) are defined operationally independent from each other, they both collaborate and compete to produce appropriate behavior. The S-S system mainly interacts with the S-R system (Yin and Knowlton 2006; Holland 2004). Conditioned reinforcement phenomenon and Pavlovian-to-instrumental transfer (PIT) are two demonstrations of this interaction, both playing a critical role in addiction to drugs. Conditioned reinforcement refers to the ability of a CS (e.g., a light associated with food) in gaining rewarding properties in order to support the acquisition of a new instrumental response (pressing a lever in order to turn the light on) (Mackintosh 1974). Actor-critic models, explained in Sect. 8.4.2 are proposed to model such an interaction between the two systems (but see Dayan and Balleine 2002). PIT, on the other hand, is a behavioral phenomenon in which non-contingent presentation of a CS markedly elevates responding for an outcome (Lovibond 1983; Estes 1948). Although PIT is suggested to play an important role in addictive behaviors, the computational accounts for the role of this phenomenon in addiction are still not well developed, and thus we do not discuss them in this chapter (see Dayan and Balleine 2002; Niv 2007 for computational models of PIT).

The so far studied interactions between the S-R and S-A-O systems, on the other hand, mainly focus on the competition between these two systems; i.e. these two systems compete for taking the control of behavior. As noted earlier, it has been demonstrated that at the early stages of learning, the behavior is governed by the S-A-O system, whereas extensive learning results in the S-R system winning the competition. The dual process models introduced in Sect. 8.4.3 are developed to model this interaction, and explain how drug-induced imbalance in the interaction between S-R and S-A-O systems can contribute to addictive behaviors.

The three different decision making processes discussed above are demonstrated to depend on topographically segregated, parallel cortico-striato-pallido-thalamo-cortical closed loops (Alexander et al. 1986, 1990; Alexander and Crutcher 1990). These loops include limbic, associative and sensory-motor loops, which are shown to mediate Pavlovian, goal-directed and habitual processes, respectively. Striatum is a central structure in this system, though it should be viewed as only a part of a bigger network. It receives glutamatergic projections from cortex, as well as dopaminergic inputs from Ventral Tegmental Area (VTA) and Substantia Nigra Pars Compacts (SNc). The striatum can be divided into anatomically and functionally heterogeneous subregions. Classically, the ventral subregion is shown to mediate Pavlovian conditioning, whereas the dorsal region is involved in instrumental conditioning (O'Doherty et al. 2004; Yin et al. 2008). Within the dorsal striatum, dorsolateral part and dorsomedial are demonstrated to mediate habitual and goal-directed processes, respectively (Yin et al. 2004, 2005, 2008).

8.3 Aspects of Addictive Behavior

In the general system-level framework within which the computational models of addiction are discussed in this chapter, three criteria for evaluating each model can be proposed. Each criterion is, in fact, a set of theories on which a system-level model of addiction is expected to be based on. Satisfying each of these criteria can improve either the behavioral explanatory power or relevancy to neurobiological reality of the corresponding model. These three criteria are: (1) being based on a model for the normal decision-making system, at both neurobiological and behavioral levels; (2) incorporating the pharmacological effects of drugs on neural systems into the structure of the computational model; and (3) explaining a set of well-known behavioral syndromes of drug addiction. In the previous section, we provided a conceptual framework for the normal decision-making system (basis 1), which will be later used as a basis for the addiction models introduced in this chapter. In this section, we focus on the third basis, and discuss some important behavioral aspects of drug addiction. Discussing the second basis is postponed until the description of computational models in Sect. 8.4.

8.3.1 Compulsive Drug Seeking and Taking

According to the current Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association 2000, p. 198) "*The essential feature of substance abuse is a maladaptive pattern of substance use manifested by recurrent and significant adverse consequences related to the repeated use of substances. . . . There may be repeated failure to fulfill major role obligations, repeated use in situations in which it is physically hazardous, multiple legal problems, and recurrent social and interpersonal problems.*" In other words, the fundamental characteristic of drug addiction is that the consumption of drug doesn't decrease proportionally when its costs (health costs, social costs, financial costs, etc.) increase. In behavioral economic terms, this type of behavior is referred to as inelastic consumption, as opposed to elastic consumption where decreases in demand are significant when price increases. In accordance with this feature, studies looking at the sensitivity of drug consumption to its price, demonstrate that the consumption of cigarettes and heroin among dependent individuals is less elastic (or sensitive) to price, compared to other reinforcers (Petry and Bickel 1998; Bickel and Madden 1999; Jacobs and Bickel 1999). Figure 8.1 presents a simplified environment for computationally investigating the sensitivity of the consumption of drugs to the associated costs (e.g., price). A decision maker (model) has two options: (a) to do nothing (C_1), which leads to the delivery of no reinforcer, and (b) to pay the cost of the drug (C_2), and then receive the drug reinforcer. The relative inelasticity of demand for drugs implies that the probability of selecting the second option (punishment-then-drug) by the model should be insensitive to the cost, as compared to a situation where the

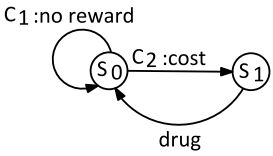

Fig. 8.1 The model has to choose between C_1 which brings no reward, and C_2. Choosing C_2 is followed by a cost, and then a drug reward

model receives a natural reinforcer instead of the drug (punishment-then-natural-reinforcer). This procedure is used to study the behavior of the model proposed in Sect. 8.4.1.

The compulsive nature of drug-seeking behavior in addicts is tried to be captured in animal models of addiction in various ways. In a variation of such experiments, rats are trained to respond on a seeking lever in order to get access to a taking lever, on which responding leads to the drug. Here, drug seeking and taking are separate actions. In the test phase, seeking responses are measured in the presence of a punishment-paired CS. In fact, during the test phase, the animal doesn't receive punishment nor drugs. Thus, its behavior is measured when no new training is provided and the animal should choose whether to continue going for the drug in the new condition or not (Vanderschuren and Everitt 2004). The formal representation of the procedure is similar to the one in Fig. 8.1: the animal can attenuate aversiveness of the expected electric shock by freezing (C_1), or alternatively, it can press the seeking lever in order to get access to the drug (C_2).

As another attempt to capture compulsivity in animal models, a CS is paired with an electric shock (electric shock plays the role of the cost associated with the drug) during the training phase. In the test phase, if the rat chooses to press the lever while the CS is present, it will receive the electric shock, which is then followed by the delivery of the drug (Deroche-Gamonet et al. 2004; Belin et al. 2008). This procedure is used to examine the behavior of the model proposed in Sect. 8.4.2.2 (a formal representation of the schedule is also provided there). In another experiment (Pelloux et al. 2007), half of the responses (i.e., lever presses) are followed by punishment (and not drug delivery), whereas the other half are followed by drug delivery (and not punishment). From an animal learning point of view, the benefit of this paradigm is that unlike the previous one, the assertiveness of the punishment will not attenuate through its association with the reward (see Pelloux et al. 2007 for more explanation). However, the exact difference between this paradigm and the previous ones from a modeling and behavioral economic point of view needs further investigation.

Intuitively, all the mentioned experiments are to investigate the degree to which the consumption of drugs is sensitive to the associated costs. However, the question of what degree of insensitivity to costs should be regarded as compulsive behavior is still unanswered. At least three types of criteria are used to distinguish between compulsive and non-compulsive drug seeking behavior: (1) Comparing the sensitivity of drug consumption to costs, with the sensitivity of the consumption of natural reinforcers (e.g., sucrose) to costs. Here, the experiments indicate that compared to natural rewards, drug consumption is less sensitive to punishments (Pelloux et al. 2007; Vanderschuren and Everitt 2004); (2) Comparing the behavior of different subpopulations of drug-exposed animals. In such experiments, animals are first divided into

groups based on a criterion like the degree of impulsivity, the degree of reactivity to novelty (Belin et al. 2008), or based on results of a test for reinstatement (Deroche-Gamonet et al. 2004). Next, the sensitivity of responses to a punishment is measured and compared between groups, and the group with the lowest sensitivity is considered to be compulsive. In this paradigm, individuals that exhibit compulsive behavior are considered as vulnerable individuals; and (3) Comparing the behavior of animals exposed to drug in different conditions and schedules of drug reinforcement. Here, the main finding is that the inelasticity in drug consumption progressively increases as the history of drug consumption increases. In fact, drug consumption becomes compulsive after a long-term drug exposure (Deroche-Gamonet et al. 2004; Pelloux et al. 2007; Vanderschuren and Everitt 2004).

In conclusion, appearance of compulsive behavior is a function of two independent factors: the degree of drug exposure (criterion 3) and the degree of vulnerability of the individual exposed to drug (criterion 2). This implies that, the more vulnerable the animal is, or the longer the period of exposure to the drug is, the insensitivity of drug consumption to punishments must increase, compared to a natural reward (criterion 1).

8.3.2 Impulsivity

Impulsivity is one of the behavioral traits that is closely related to addiction (Dalley et al. 2011). Addicts are generally characterized as impulsive individuals. They usually exhibit deficiency in response inhibition when it is necessary for reward acquisition, even in non-drug-related tasks. Impulsivity is a multidimensional construct, though two aspects of it seem to be more important: impulsive choice and impaired inhibition. Formal modeling and simulation of situations measuring impulsive choice is rather straight-forward (see below). However, modeling an environment for assessment of impaired inhibition (i.e., inability to inhibit maladaptive behaviors) is hard to achieve, and to our knowledge, there is no computational study on impaired inhibition. Thus, hereafter, we focus on the impulsive choice aspect and refer to it as impulsivity.

Impulsivity is defined as the selection of a small immediate reward over a delayed larger one (Dalley et al. 2011). Figure 8.2 illustrates an environment for the delay discounting task, which is commonly used for the assessment of impulsive behavior. As the figure shows, the model has two choices: one (C_1) leads to an immediate small reward, R_s, and the other (C_2) leads to a delayed (k time steps), but larger reward, R_l. In this environment, the impulsive individuals are those that have more tendency to small rewards, compared to other individuals. A wealth of evidence in human subjects suggests that drug-dependent individuals have more tendency to the small-reward choice, compared to non-dependent individuals (see Bickel and Marsch 2001; Reynolds 2006 for a review). In the same line, animal models report that chronic drug intake causes impulsive choice in rats, as they show less ability to delay gratification compared to control rats (Simon et al. 2007; Paine et al. 2003).

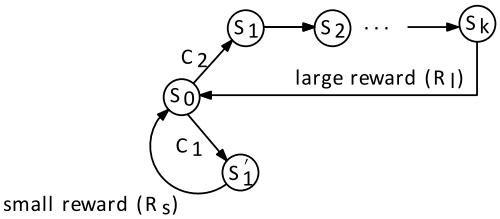

Fig. 8.2 Delay discounting task. The model has two choices, C_1 and C_2. Selection of C_1 leads to a small reward, R_s, only after one time step, whereas by choosing C_2, the model should wait for k time steps, and a large reward, R_l, will be delivered afterwards

It is still unclear whether this increased impulsivity in drug-dependent individuals is a determinant or only a consequence of drug use. However, in human, using self-report measures of impulsivity, it has been reported that youth with impulsive traits are more likely to initiate drug use (see de Wit 2009 for a review). In animals, high impulsivity measured by lack of behavioral inhibition predicts transition to compulsive behavior (Belin et al. 2008). Accordingly, it can be expected from a model of addiction that the more impulsive the model is (as measured by impaired inhibition), the more vulnerable it should be to develop compulsive drug-seeking. However, for concluding that choice impulsivity is also an indicator of vulnerability to develop compulsivity, there should at least be a strong correlation between these two measures of impulsivity. Although it is reported in some studies that impaired inhibition is significantly correlated with impulsivity measured in the delay discounting task (Robinson et al. 2009), other evidence suggest that these two behavioral constructs are not necessarily overlapping (de Wit 2009). Thus, for establishment of links between choice impulsivity and compulsivity, further computational works are needed on modeling impaired inhibition forms of impulsivity.

8.3.3 Relapse

Although compulsive drug taking is an important defining feature of addiction, the most challenging clinical feature of addicts is that they remain vulnerable to relapse, even after long periods of withdrawal (Stewart 2008). Clinical and experimental studies have shown that non-contingent injections of drugs, re-exposure to drug-paired cues, and stress are three factors reinstating drug taking and seeking behavior (Shaham et al. 2003). In a typical reinstatement model of relapse, animals are first trained to acquire responses that lead to the drug (e.g., lever press in order to gain access to the drug). Next, they undergo "extinction training" in which, responses no longer result in the drug outcome. Once the behavior has extinguished, in a subsequent test phase, the effect of different factors triggering relapse (stress, drug priming, drug cues) on the extinguished behavior is determined.

According to these experimental procedures, developing formal representations of the tasks is straightforward. The challenging point, however, is the effect of pharmacological and environmental stimuli on the internal processes of a model, that

is, how the effect of drug priming or stress on the brain neurocircuitry can be represented in computational models. In Sect. 8.4.3, we return to these questions and suggest a potential way for modeling these manipulations.

8.4 Computational Accounts

8.4.1 S-R Models

Habit or S-R learning is the ability to learn adaptively to make appropriate responses when some certain stimuli are observed. According to this theory, given a situation or stimulus, if making a certain response produces a reward (a pleasant, biologically salient outcome), then the corresponding S-R association will be potentiated (reinforced) and thus, the probability of taking that response in similar circumstances in the future will increase. Inversely, a behavior will occur less frequently in the future, if it is followed by a punishment (an aversive outcome). In this manner, animals can be viewed as organisms that acquire appropriate behavioral strategies in order to maximize rewards and minimize punishments. This problem, faced by the animals, is analogous to the problem addressed in the machine learning theory of reinforcement learning (RL), which studies how an artificial agent can learn, by trial and error, to make actions to maximize rewards and minimize punishments. Indeed, in recent years, strong links have been forged between a method of RL, called Temporal Difference Reinforcement Learning (TDRL), animal conditioning literature and the potential underlying neural circuits of decision making. The developed neurocomputational models in this interdisciplinary field has provided as an appropriate basis for modeling drug addiction.

In the RL framework, stimulus and response are referred to as "state" and "action", respectively. At each time-step, t, the agent is in a certain state, say s_t, and among the several possible choices, it takes an action, say a_t, on the basis of subjective values that it has assigned to those alternatives through its past experiences in the environment. These assigned values are called Q-values. The more Q-value does an action have, the more likely that action is to be selected for performance. Denoting the probability of taking action a_t at state s_t by $\pi(a_t|s_t)$, the below equation known as the *Softmax* rule reflects this feature:

$$\pi(a_t|s_t) = e^{\beta Q(s_t,a_t)} \Big/ \sum_{b\in\mathbb{A}_{s_t}} e^{\beta Q(s_t,b)} \tag{8.1}$$

where $\mathbb{A}_{s_t}$ is the set of all available choices at state s_t. β is a free parameter determining the degree of dependence of the policy π on Q-values. In other words, this parameters adjust the exploration/exploitation trade-off.

For making optimal decisions, Q-values are aimed to be proportional to the discounted total rewards that are expected to be received after taking the action onward:

$$Q(s_t,a_t) = E\big[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots | s_t, a_t\big] = E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i | s_t, a_t\right] \tag{8.2}$$

Achieving this objective requires the animal to sufficiently explore the task environment. In the previous equation, $0 < \gamma < 1$ is the discount factor, which indicates the relative incentive value of immediate rewards compared to delayed ones.

To update the prior Q-values, a variant of RL known as TDRL calculates a prediction error signal each time the agent takes an action and receives a reward (as a feedback) from the environment. This prediction error is calculated by comparing the prior expected value of taking that action, $Q(s_t, a_t)$, with its realized value after receiving the reward r_t:

$$\delta_t = \gamma(r_{t+1} + V(s_{t+1})) - Q(s_t, a_t) \tag{8.3}$$

In this equation, $V(s_{t+1})$ is the maximum value of all feasible actions available at the state that comes after taking the action a_t. This prediction error is then utilized to update the estimated value for that action:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t \tag{8.4}$$

where $0 < \alpha < 1$ is the learning rate, determining the degree to which a new experience affects the Q-values. As a critical observation, the phasic activity of midbrain dopamine (DA) neurons is demonstrated to be significantly correlated with the prediction error signal that the TDRL model predicts (Schultz et al. 1997). In fact, dopamine neurons projecting to associative learning structures of the cortico-basal ganglia circuit are believed to carry a teaching signal that modulates the strength of S-R associations and thus, will increase the probability of taking an action in the future, if an unexpected reward has come as a consequence of that action.

TDRL provides a framework for the better understating of the S-R habit formation. In this framework, reinforcement of an association between stimulus s and response a after receiving a reward is equivalent to an increase in $Q(s_t, a_t)$. By utilizing the *softmax* action-selection rule, this will result in increasing the probability of taking that action in the future. By interpreting the TDRL model from another point of view, since only previously learned values accumulated through time determine which action the model takes in a certain state, the behavior of a TDRL model is not sensitive to sudden environmental changes. In other words, it takes several learning trials for the value of actions to be updated according to the new conditions. On the basis of this feature, the TDRL framework is behaviorally consistent with habitual (S-R) responding.

8.4.1.1 Redish's Model

If phasic dopamine activity corresponds to the reward prediction error signal, then after sufficient learning when predictions converge to their true values, the prediction error and thus phasic DA activity should converge to zero. In fact, this happens in the case of natural rewards: after adequate learning trials, the phasic activity of DA neurons vanishes. However, this is not true in the case of drugs such as cocaine and amphetamine. These drugs, through their neuropharmacological mechanisms, increase dopamine concentration within the striatum (Ito et al. 2002;

Stuber et al. 2005). This artificial build up of dopamine readily means that the error signal can not converge to zero in the course of learning and as a consequence, the experienced value of drug-related behaviors will grow more than expected. Based on this argument, a modified version of the TDRL algorithm is proposed Redish (2004) that can explain some behavioral aspects of addiction. Assuming that the pharmacological effect of drugs induces a bias with the magnitude of D on dopaminergic signalling, the error signal equation (8.3) can be rewritten as below when drug is available (Redish 2004):

$$\delta_t^c = \max(\gamma(r_t + V(s_{t+1})) - Q(s_t, a_t) + D(s_t), D(s_t)) \tag{8.5}$$

This implies that the prediction error signal will always be higher than D, as long as the drug's effect is available:

$$\delta_t^c \geq D(s_t) \tag{8.6}$$

Hence, by each drug consumption session the value that a decision-maker predicts for drug-seeking and -taking increases. This leads to the over-valuation of this behavior and explains why drug-associated behaviors become more and more insensitive to their harmful consequences through the course of addiction, as measured by the behavior of the model in the environment shown in Fig. 8.1. In fact, as drug-related S-R associations become more and more reinforced, only a more intense adverse event can cancel out the high estimated value of drug-seeking. This model, thus, explains how compulsive drug-seeking habits develop as a result of repeated drug abuse.

Thus, the model proposed in Redish (2004) provides an elegant explanation for progressive inelasticity of drug consumption as a function of the drug exposure history. However, this account does not propose explanations for other addictive behaviors such as impulsivity and relapse. Besides, some predictions of the model have proven inconsistent with some studies that have explicitly investigated the validity of the way in which the effect of drugs are modeled on the error signal.

Firstly, the model predicts that the true value of drug can never be predicted by environmental cues, because it is always better than expected. A behavioral implication of this property is that the "blocking" effect (Kamin 1969) should not occur for the case of drugs (Redish 2004). In fact, the "blocking" phenomenon occurs when a stimulus, as a result of sufficient training, can correctly predict the value of the upcoming outcome. In this case, if a new stimulus is paired with the old one after the training period, since the old stimulus can correctly predict the value of the outcome, no prediction error (teaching signal) should be generated and thus, no new learning will occur. Therefore, it is said that the old highly-trained stimulus blocks other stimuli to be associated with the outcome. However, as the model proposed in Redish (2004) assumes that drugs always induce non-compensable dopamine signalling, it predicts that the blocking effect should not be observed for stimuli that predict drugs. However, experimental results have shown that the "blocking" effect does occur in the case of drugs (Panlilio et al. 2007) and thus, the always-better-than-expected value formation for the drug is not a correct formulation. Secondly, the validity of this method of value learning is investigated even more explicitly. In

Marks et al. (2010), rats were first trained to press two levers in order to receive a large dose of cocaine. Then, the dose associated with one of the levers was decreased. Here, the theory predicts that the value associated with the low-dose lever will not decrease, because drug consumption always increases the value irrespective of the experienced dose (see Eq. (8.6)). At odds with this prediction, the result showed that the lever press performance for the reduced-dose lever has decreased, which indicates that the value of the drug has decreased.

8.4.1.2 Dezfouli et al.'s Model

Borrowing from the model proposed by Redish (2004) (the idea that drugs increase the error signal), we proposed another computational model for drug addiction (Dezfouli et al. 2009) that is based on the supplementary assumption that long-term exposure to drugs causes a long lasting dysregulation in the reward processing system (Koob and Le Moal 2005a). Consistent with behavioral findings, this persistent dysregulation causes less motivation in addicts toward natural rewards like sexually evocative visual stimuli, as well as secondary rewards like money (Garavan et al. 2000; Goldstein et al. 2007).

This dysregulation of the reward system can be modeled in a variant of the TDRL algorithm called "average-reward" TDRL (Mahadevan 1996). In this computational framework, before affecting the current strength of associations, rewards are measured against a level called "basal reward level" (Denoted by ρ_t). As a result, an outcome will have reinforcing effect only if the reward value is higher than the basal reward level. Otherwise, the reinforcing value of the outcome will be negative. The basal reward level, according to this framework, is equal to the average reward per step, which can be computed by an exponentially weighted moving average over experienced rewards (σ is the weight given to the most recent received reward):

$$\rho_t \leftarrow (1-\sigma)\rho_t + \sigma r_t \tag{8.7}$$

In fact, an outcome will reinforce the corresponding association only if it has a rewarding value higher than what the animal receives on average. In this formulation, the value of a state-action is the undiscounted sum of all future rewards measured against ρ_t:

$$Q(s_t, a_t) = E\left[\sum_{i=t}^{\infty}(r_i - \rho_i)|s_t, a_t\right] \tag{8.8}$$

These state-action values can be learned using the following error signal:

$$\delta_t = \gamma(r_{t+1} + V(s_{t+1})) - Q(s_t, a_t) - \rho_t \tag{8.9}$$

Using this error signal, Q-values are updated by the same rule of Eq. (8.4). The definition of the error signal in the average reward RL algorithm does not imply that the value of a state is insensitive to the arrival time of future rewards. In contrast, in Eq. (8.9), the average reward (ρ_t) is subtracted from $V(s_{t+1})$, meaning that by waiting in state s for one time step, the agent loses an opportunity to gain potential future

rewards. This opportunity cost is, in average, equal to ρ_t, and is subtracted from the value of the next state. This learning method guides action selection to a policy that maximizes the expected reward per step, rather than maximizing the sum of discounted rewards. As in the simple TDRL framework, the error signal computed by Eq. (8.9) corresponds to the phasic activity of DA neurons. The term ρ_t, on the other hand, is suggested to be coded by the tonic activity of DA neurons (Niv et al. 2007).

Roughly, long-term exposure to drugs causes two, perhaps causally related, effects on the dopamine-dependent reward circuitry. Firstly, chronic exposure to drugs affects the dopamine receptors availability within the striatum. Human subjects and non-human primates with a wide range of drug addictions have shown significant reductions in D2 receptor density within the striatum (Nader et al. 2002; Porrino et al. 2004a; Volkow et al. 2004b). This effect reduces the impact of normal dopamine release that carries the error signal and thus, results in a reduction in the magnitude of the error signal, compared to its normal value (Smith et al. 2006). Secondly, it is proposed that chronic drug abuse causes an abnormal increase in the tonic activity of dopamine neurons (Ahmed and Koob 2005). As the tonic DA activity is hypothesized to encode the ρ_t signal, this second effect of drugs can be modeled by abnormal elevation of the basal reward level. Thirdly, as mentioned earlier, chronic drug exposure causes decreased sensitivity of the reward system to natural rewards. This effect can be interpreted as an abnormal elevation of the level against which reward is measured. In other words, long-term drug abuse elevates the basal reward level to a level that is higher than that of normal subjects. This drug-induced elevation of the basal reward level, ρ_t, can be formally captured by adding a bias to it:

$$\rho_t^c = \rho_t + \kappa_t \tag{8.10}$$

Normally, κ_t is zero and therefore, rewards are measured against their average level (ρ_t). However, with drug use, κ_t grows and consequently, the basal reward level elevates abnormally to ρ_t^c. This modification covers all the three long-term effects of drugs discussed above. As adding a positive bias to ρ_t leads to a decrease in the error signal (see Eq. (8.9)), it is somehow reflecting the reduced availability of dopamine receptors. Alternatively, if ρ_t is related to the tonic activity of DA neurons, adding a bias to it corresponds to an increase in the tonic activity of these neurons.

According to the above modification to the average reward TDRL algorithm, we rewrite the error signal equation for the case of drugs as follows:

$$\delta_t^c = \max(\gamma(r_t + V(s_{t+1})) - Q(s_t, a_t) + D(s_t), D(s_t)) - \rho_t^c \tag{8.11}$$

Similar to the model proposed in Redish (2004), the maximization operator reflects the drugs' neuropharmacological effects, but unlike that model, the error signal is not always greater than zero. In this model, although drugs produce extra dopamine through direct pharmacological mechanisms, due to the increase in the basal reward level, the error signal will eventually converge to zero. This property ensures that the estimated value of the drug does not grow unboundedly, which makes the model more biologically plausible. Furthermore, as the prediction error

signal in this model can converge to zero after sufficient experience with drugs, no further learning will occur after extensive training. This will result the drug-predicting stimuli to block forming new associations. This is consistent with the report that the blocking effect is observed for the case of drugs (Panlilio et al. 2007).

It should be noted that because abnormal elevation of the basal reward level is a slow process, the error signal under the effect of drugs will be above zero for a relatively long time and thus, drug-seeking habits will be abnormally reinforced. This leads to insensitivity of drug consumption to drug associated punishment, as indicated by the tendency of the model toward C_2 in the environment shown in Fig. 8.1.

As the decision-making system is common for natural and drug reinforcers, deviation of the basal reward level from its normal value can also have adverse effects on decision making in the case of natural rewards. Within the framework proposed above, ρ_t^c determines the cost of waiting. Hence, high values of ρ_t^c in an environment indicate that waiting is costly and thus, guide the decision maker to options with a relatively faster reward delivery. In contrast, low values indicate that the delayed interval before reward delivery is not costly and it is worth waiting for a delayed but large reward. If chronic drug exposure leads to high values of ρ_t^c, then the model's behavioral strategy will shift abnormally toward more immediate rewards, even if their rewarding value is less than that of distant rewards. In other words, in the environment show in Fig. 8.2, preference of the model toward C_1 increases as the degree of prior exposure to drug increases. This is because the cost of waiting is relatively high and the decision-maker prefers to have immediate rewards. This explains why addicts become impulsive after chronic drug abuse (Logue et al. 1992; Paine et al. 2003; Simon et al. 2007).

As another deficit in the decision-making mechanism, since the basal reward level abnormally elevates in addicts, the model predicts that the motivation for natural reinforcers will decrease after long-term drug exposure. This prediction is consistent with behavioral evidence in human addicts (Garavan et al. 2000; Goldstein et al. 2007).

8.4.2 S-S and S-R Interaction: Actor-Critic Models

Actor-critic is a popular reinforcement learning model that subdivides the process of decision making into two subtasks: learning and action-selection (Sutton and Barto 1998). These two tasks are conducted by the "critic" and the "actor" components, respectively.

The critic component is responsible for adaptively predicting the value of states, $V(s_t)$, by utilizing the prediction error signal. Assuming that the agent leaves state s_t, enters state s_{t+1} and receives reward r_t at time t, the critic will compute the prediction error signal based on the received reward and the prior expectation of the agent:

$$\delta_t = \gamma(r_t + V(s_{t+1})) - V(s_t) \tag{8.12}$$

This prediction error is then used for updating predictions of the critic:

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \tag{8.13}$$

where, as before, α is the learning rate. As the critic only predicts the value of a state ($V(s_t)$), without caring about what action or external cause has led to it, it is suggested to be a model for S-S (Pavlovian) learning.

The actor component, on the other hand, is involved in making decisions about what action to perform at each state, based on its stored preferences for different actions, $P(s_t, a_t)$: the higher the preference toward an action, the higher the probability of taking that action by the actor. The preferences in the actor are learned based on the values learned by the critic: if taking an action by the actor in a state results in an increase in the value of the state (computed by the critic), the preference toward that action will also increase. The converse is also true: if taking an action leads to a decrease in the critic's value of the state, the probability that the actor takes the action again also decreases by decreasing the preference for that action.

For achieving this harmony, the critical feature of the actor-critic model is that the preferences in the actor are updated using the same prediction error signal that is produced and utilized by the critic component:

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + \alpha\delta_t \tag{8.14}$$

The fact that the actor uses the error signal generated by the critic can be viewed as an interaction between the S-S (critic) and the S-R (actor) systems. Behaviorally, conditioned reinforcement phenomenon implies that a CS which is associated with a reinforcer (e.g., a light associated with food) supports the acquisition of a new instrumental response (pressing a lever in order to turn the light on). Here, the association between the CS and the reinforcer can be learned by the critic component, that is, the value of the state in which the CS is presented (s_{CS}) increases as the reward in the subsequent state (reward delivery state) is experienced. Next, when several actions are available in a state (s_A), the action that leads to s_{CS} obtains a higher preference (learned by the actor), because taking that action leads to an increase in the value of s_A, as predicted by the critic.

In this respect, dissociating the functions of prediction and action-selection in the actor-critic model is reminiscent of the behavioral psychologist dissociation between Pavlovian and instrumental processes (Niv 2007; Joel et al. 2002). Consistently, a relatively rich body of experiments has shown the dissociable role of striatal subdivisions in prediction and action-selection (O'Doherty et al. 2004; Roesch et al. 2009). Based on these observations, critic and actor components can be thought to be neurally implemented by limbic and sensorimotor cortico-striatal loops, respectively.

Dopamine neurons are hypothesized to integrate information across parallel loops in the cortico-basal ganglia circuit (Haber et al. 2000; Haber 2003), by propagating the prediction error signal made by more limbic (ventral) regions toward associative (dorsomedial) and then motor (dorsolateral) areas of the striatum, via the spiral organization of dopamine neurons. By these spiral connections between the striatum and the VTA/SNc, the output of the accumbens shell can affect the

functioning of the core region and in the same way, the output of the accumbens core can influence more dorsal domains of the striatum, via SNc. These dopamine spirals that travel from the ventral to the dorsal regions of the striatum can account for the assumption of the model that the prediction error signal used for updating the actor's preferences is the same signal generated and used by the critic (Joel et al. 2002). These behavioral and neurobiological supports of the actor-critic model has made it a popular model for decision making analysis, and the central role that dopamine plays in it, has allowed addiction-modelers to employ it as a basis for their models.

8.4.2.1 Dayan's Model

Recently, inspired by the model proposed in Redish (2004), Dayan proposed an actor-critic model for addiction (Dayan 2009). The model is based on a variant of the actor-critic model called "advantage learning" (Dayan and Balleine 2002) in which, the critic module has the same algorithm as the classical actor-critic model explained above. Thus, the critic module produces a prediction error signal (δ_V) and uses it for both updating its own value predictions (as in Eq. (8.13)) and also feeding it into the actor component. Rather than learning the preference toward actions, the actor component learns the advantage of taking that action over all other actions that has been previously taken in that state. This "advantage" is denoted by $A(s_t, a_t)$. To learn this "advantage", the actor uses a transformed error signal δ_A:

$$\delta_A = \delta_V - A(s_t, a_t) \tag{8.15}$$

This signal is then used to update the expected advantages:

$$A(s_t, a_t) \leftarrow A(s_t, a_t) + \alpha\delta_A \tag{8.16}$$

The actor utilizes advantages instead of classic preferences to choose among different possible actions. After sufficient learning, as the best action will be the action that the agent takes frequently, its advantage over previously taken actions will tend to zero and the advantage of other alternatives will become negative in their steady levels.

The basis of this model is a hypothesis suggested in Everitt and Robbins (2005) that explains addiction, at a behavioral level, as a transition from voluntary control over drug consumption at the early stages to rigid habitual and compulsive behavior in later stages. Specifically, the hypothesis indicates that this behavioral shift is based on a transition of control over drug-seeking behavior from limbic structures, such as prefrontal cortex (PFC) and nucleus accumbens (NAc), to more motor structures, particularly dorsal striatum. Neurobiological evidence has suggested that this shift is mainly mediated by striatal-midbrain spiraling network that connects the ventral regions of the striatum to more the dorsal parts (Belin and Everitt 2008). According to the Dayan's model, the pharmacological effect of drugs on the dopamine spirals will not only affect the actor indirectly through its effect on the critic's error signal, δ_V, but will also directly affect the actor's updating mechanism due to its effect on δ_A.

In fact, if the pharmacological effect of drugs is assumed to be equal to D, then it will be augmented to the critic's error signal and thus, the critic's value for a drug state will converge to $\gamma(r_D + V(s_{t+1})) + D$. Similarly, due to the effect of drugs on δ_A, the advantage of drug-related actions will increase by D units. This abnormality has been interpreted as a reason to explain why drug-seeking behavior becomes compulsive. The model in Dayan (2009) can also explain how addictive drugs can induce abnormal drug-seeking behavior without abnormally affecting the addict's expectations stored in the critic.

8.4.2.2 Piray et al.'s Model

So far, we have described models that have explained addiction as a disease that is pervasively augmented by drug experience. However, like other diseases, addiction requires a suitable host, that is, a susceptible individual, to spread (Nader et al. 2008). Indeed, overwhelming evidence has shown that only a subpopulation of humans, as well as animals, that have experienced drugs, show symptoms of addiction (compulsive drug seeking and taking) (O'Brien et al. 1986). Some behavioral traits and neural vulnerabilities have been hypothesized to predispose addiction (Koob and Le Moal 2005a; Everitt et al. 2008; Nader et al. 2008). Importantly, a large body of literature suggests a crucial role for dopamine receptors in predisposition to exhibit addiction-like behavior. For example, Dalley and colleagues have shown that lower density of D2 receptors in NAc, but not dorsal striatum, of rats, predicts higher tendency to cocaine self-administration and also addiction-like behavior (Dalley et al. 2007; Belin et al. 2008). Similar results have been reported in non-human primates' neuroimaging studies (Nader et al. 2008), as well as in human studies (Volkow et al. 2008). Moreover, it has been reported recently that low D1 receptor availability within NAc predisposes tendency to cocaine self-administration (Martinez et al. 2009).

In a similar line, a wealth of evidence has shown the important role of dopamine receptors in the development of obesity (Johnson and Kenny 2010) and pathological gambling (Steeves et al. 2009). This is computationally important because a common framework for these diseases and drug addiction, as suggested by Volkow et al. (2008) and Potenza (2008), cannot be constructed only by focusing on the direct pharmacological effects of drugs (Ahmed 2004), but instead, there should be a model that some elements of it bootstrap abnormal and compulsive tendency to rewarding stimuli.

Recently, we proposed a simple actor-critic like model to capture this feature of addiction (Piray et al. 2010). The model relies on three assumptions motivated by neurobiological evidence: (1) VTA dopamine neurons encode action-dependent prediction error (Roesch et al. 2007; Morris et al. 2006) and ventral striatal neurons encode action-dependent values (Roesch et al. 2009; Nicola 2007; Ito and Doya 2009), (2) lower co-availability of D1 and D2 receptors, that is, lower availability of either D1 or D2, in NAc is a necessary condition for addiction to both drug and food to develop (Hopf et al. 2003; Ikemoto et al. 1997; Dalley et al. 2007;

Johnson and Kenny 2010; Martinez et al. 2009), and (3) the first leg of the spiral, that is, posteromedial VTA to NAc shell, is involved in appetitive but not aversive learning (Ford et al. 2006; Ikemoto 2007).

The translation of these assumptions to actor-critic components is straightforward. The first assumption can be interpreted as an action-dependent value representation in the critic, $V(s_t, a_t)$, and also action-dependent prediction error, instead of action-independent ones (see Eq. (8.12):

$$\delta_t = \gamma(r_t + V(s_{t+1})) - V(s_t, a_t) \tag{8.17}$$

$V(s_{t+1})$ is again the value of the best available choice at state s_{t+1} (see Piray et al. 2010 for further discussion).

To model the second assumption, we need to suppose a role for dopamine receptors in terms of the actor-critic model. In line with previous studies (Rutledge et al. 2009; Frank et al. 2007), we have assumed that the availability of dopamine receptors modulates the learning rate (see Smith et al. 2006; Dezfouli et al. 2009 for other ways of modeling the function of dopamine receptors in RL models). Thus, a slight modification in the critic's learning rule, Eq. (8.13), is required:

$$V(s_t, a_t) \leftarrow V(s_t, a_t) + \kappa_c \alpha \delta_t \quad \text{if } r > 0 \tag{8.18}$$

where κ_c corresponds to the availability of dopamine receptors in the NAc. In this formulation, the second assumption can be realized by normalizing the parameter κ_c to one for a healthy subject, and setting it to a value less than one ($\kappa_c < 1$) for individuals who are susceptible to addiction. Finally, the third assumption implies that only appetitive, but not aversive, learning is modulated by the availability of dopamine receptors. Thus, Eq. (8.18) should only be used for appetitive learning; and for learning the value of aversive outcomes ($r < 0$), the prediction error computed by Eq. (8.17) will be used directly.

The behavior of the model can be examined in the task introduced in Deroche-Gamonet et al. (2004). In this experiment, animals learn to self-administer drugs by performing a lever-press action firstly. In the next phase, the lever-press action gets paired with an acute shock punishment. It has been reported that only a proportion of rats, almost 20 percent, that had prolonged experience with drugs, show compulsive behavior.

Figure 8.3 illustrates the behavior of the model in an environment that models the mentioned experiment. As the figure shows, the simulated individual selects the drug-related lever, even after removing the drug reward and instead, giving an acute punishment (phase 2). Since the critic's value is updated with $\kappa_c \alpha \delta$, but the actor's preference is updated by $\alpha \delta$, when $\kappa_c < 1$, the preference toward action a in phase 1 increases abnormally, whereas it increases in a normal way in the critic. In phase 2, however, both the value and the preference are updated by an equal amount and thus, as the figure shows, the amount of drop in both the critic's value and the actor's preference is equal. This drop is sufficient for the critic's value, $V(s, a)$, to converge to r_{sh}, but is not enough to make the preference, $P(a, s)$, negative. For action b, as the reward associated with it is zero, its value and preference remain zero. Hence, in phase 2, while the value of action a falls below the value of action

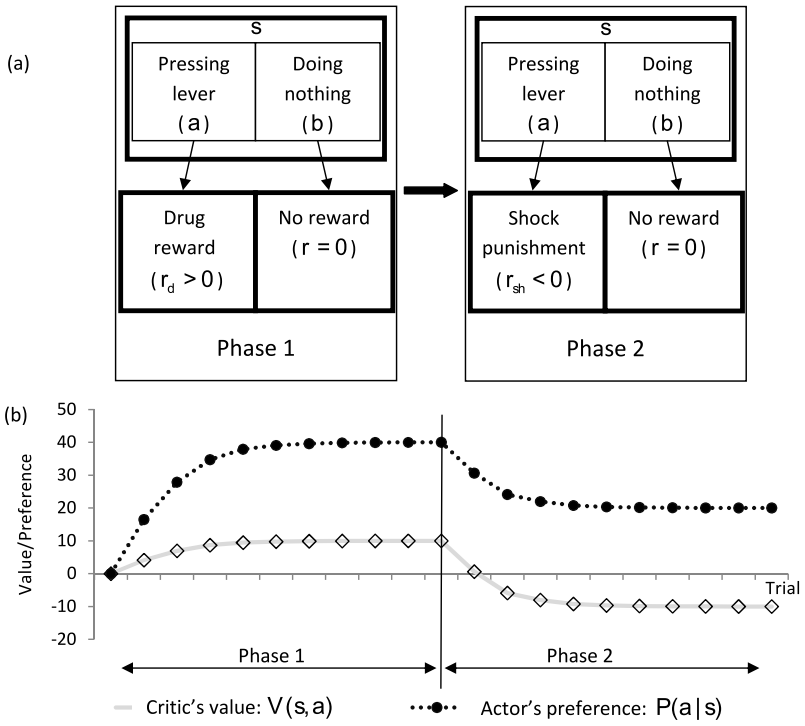

Fig. 8.3 (**a**) A model with vulnerability to addiction ($\kappa_c = 0.25$) performs the task illustrated in the figure. In state s, the model chooses between two actions. Action a results in a drug reward (drug taking action, $r_d = 10$) and action b results in no reward. After sufficient learning in this phase, the drug reward will be removed and action a is paired with a shock punishment, r_{sh} (phase 2). (**b**) The performance of the model proposed in Piray et al. (2010) in the mentioned task. While the optimal behavior in phase 2 is choosing b, the vulnerable model chooses a. This is because the preference toward a in phase 1 is abnormally exaggerated in the actor, while its value is normal in the critic. Moreover, since learning from punishment is required in this phase, both value and preference will be updated equally and thus, the amount of drop in both the critic's value and the actor's preference is equal. After a while, the critic's value converges to r_{sh} and thus, the prediction error by performing a converges to zero. As a result, no more changes in the value and the preference associated with a occurs. This effect will decrease the critic's value for a to a level below that of b (zero), but is not sufficient to make the preference toward a negative. The origin of this behavior is the abnormal increase in the actor's preference (habit) toward a in phase 1

b, the preference toward action a is still above that of action b. In fact, when the critic's value converges to r_{sh}, the prediction error of performing a converges to zero and so, no change in the value and preference associated with a occurs.

Notably, if we assume that by chronic administration of drugs, the availability of receptors will further decrease, which is supported by neurobiological data (Nader et al. 2002; Porrino et al. 2004a; Volkow et al. 2004b), the discrepancy between the values and the preferences in the appetitive system will further increase through learning (see Piray et al. 2010 for details). Hence, the insensitivity of addicts to

the negative consequences of drug-taking will increase after a prolonged experience with drugs.

This model has two major behavioral implications. First, the compulsivity only appears in vulnerable individuals. Second, compulsivity does not depend on the pharmacological effect of drugs and thus, the model can explain compulsive tendency to natural rewards, such as palatable foods and gambling, in a common framework with compulsive drug taking. The important neurobiological implication of the model is that compulsivity depends on abnormally strong actor's preferences toward drugs; however, it is the critic's deficit that is the origin of this abnormal behavior. Thus, the model accounts for the progressive shift of behavior control during drug consumption from ventral to dorsal striatum, which is initiated by the ventral striatal vulnerabilities and mediated by the dopaminergic spiralling network (Everitt et al. 2008; Porrino et al. 2004b).

8.4.3 S-R and S-A-O Interaction: Dual-Process Models

Whereas actor-critic models have tried to model some properties of the limbic and sensorimotor loops as well as their interaction, dual-process models are focused on the sensorimotor and associative loops, responsible for making habitual and goal-directed decisions, respectively, as well as competitive and collaborative interactions between them (Daw et al. 2005; Keramati et al. 2011). In this section we explain a dual-process model that we have proposed recently (Keramati et al. 2011), which we believe has important implications for explaining some aspects of addiction.

In this model, similar to the seminal dual-process model (Daw et al. 2005), the fundamental nature of the habitual system is the same as a simple TDRL model discussed in previous sections. This system is capable of enforcing or weakening an association between a state and an action (denoted by $Q^H(s_t, a_t)$, hereafter), based on the prediction error signal, which is hypothesize to be carried by the phasic activity of dopamine neurons (Schultz et al. 1997). At the time of decision making, the established associations can be exploited, and as all the information needed for making a choice between several alternatives is accumulated in S-R associations from previous experiences, the habitual responses can be made within a short interval after the stimulus is presented. However, this speed in action selection doesn't come without cost: because many learning trials are required for the outcomes of an action to affect an association, the strength of associations are low-elastic to the outcomes, making the habitual responses inaccurate, particularly under changing motivational or environmental conditions.

In contrast, the goal-directed system is hypothesized to learn through experience the causal relationship between actions and outcomes, so that it has access to a decision tree at the time of decision making and can deliberate the consequences of different alternatives. Denoting the learned dynamics of the environment by $\hat{p}_T(s \xrightarrow{a} s')$ (indicating the probability of traveling from state s to s' by taking action a), and

the reward function by $\hat{p}_R(r|s,a)$ (indicating the probability of receiving reward r by taking action a at state s), the estimated value of a state-action pair can be calculated by the below recursive equation. This algorithm is intuitively equivalent to a full-depth search in a decision tree for finding the maximum attainable reward by taking each of the available choices:

$$\hat{Q}^G(s_t,a_t) = E\left[\hat{p}_R(r|s_t,a_t)\right] + \gamma \sum_{s'} \hat{p}_T\left(s_t \xrightarrow{a_t} s'\right).\hat{V}\left(s'\right) \tag{8.19}$$

Although this system can estimate the value of actions more accurately and more optimally, it is not as fast as the habitual system because of the cognitive load (tree search) required for value estimation. Thus, the animals' decision making machinery is always confronted with a trade-off between speed and accuracy; that is, whether to make a fast, but inaccurate habitual response, or to wait for the goal-directed system to make a more optimal decision. This trade-off is hypothesized to be based on a cost-benefit analysis. Assuming that the time needed for the goal-directed system to accurately calculate the estimated value of each available action is τ, the cost of deliberating for each response will be $\bar{R}\tau$, where $\bar{R}$ is the amount of reward that the animal is expected to receive at each unit of time. This variable can be simply computed by taking an average over the rewards obtained through time in the past, as in Eq. (8.7). As discussed before, this average reward signal is hypothesized to be carried by the tonic activity of dopamine neurons (Niv et al. 2007). Thus, if for whatever reason the tonic firing rate of dopamine neurons elevates, the model predicts that the cost of goal-directed responding will increase and consequently, decisions will be made more habitually.

The benefit of deliberation for a certain action, on the other hand, is equal to how much the animal estimates that having the exact value of that action will help it improve its decision policy. This parameter, called "value of perfect information (*VPI*)", is computable using the estimated Q-values and their corresponding uncertainties cached in the habitual system. Without going into details of the algorithm, one critical prediction of the model is that if the values of two competing actions, estimated by the habitual system, are very close together, then knowing their exact values would greatly help the animal make the optimal decision between those two choices. In contrast, if at a certain state, the estimated value of one of the feasible choices is markedly greater than other actions, and its uncertainty is low, then it can be inferred by the animal that it is less likely that having perfect information about the value of actions will change its initial conjecture about the best choice, made by the habitual system. Thus, under such conditions, the goal-directed system will not contribute to the decision making process.

As the consistency of the model with behavioral and neuronal findings is discussed in the original paper (Keramati et al. 2011), we focus here on the implications of the model for addiction.

All the previous computational theories of addiction discussed in this chapter explain how drug-seeking and drug-taking habits consolidate through the course of addiction as a result of neuroplasticity in different regions of the cortico-basal ganglia circuit, under the effect of dopamine bursts. Although these models have

proven fruitful to some degrees, many theories of addiction emphasize on impairment of top-down cognitive control as the essential source of compulsivity. Inability of addicts in breaking habits that have evident adverse consequences is attributed to dysfunctional prefrontal cortical executive control over abnormally strong maladaptive habits. In fact, the evolution of control over behavior from ventral to dorsal striatum, discussed in the previous section, is followed by a shift within the dorsal striatum from action-outcome to stimulus-response mechanisms (Pierce and Vanderschuren 2010; Belin et al. 2009).

Taking into account the effect of drugs on phasic dopamine, the dual-process model discussed above can explain how addictive drugs, by over-reinforcing stimulus-response associations, result in the estimated value of the habitual system for drug-seeking choices becoming maladaptively high. As a consequence, the *VPI* signal (benefit of deliberation) for those actions will be very low after long-term drug consumption and thus, the individual will make habitual and automatic responses, without considering the possible consequences. Consistent with this prediction, it has been reported that short-term drug seeking is a goal-directed behavior, whereas after prolonged drug exposure, drug seeking becomes habitual (Zapata et al. 2010). According to the models introduced in the previous sections, this is equivalent to the insensitivity of drug consumption to harmful consequences.

Beside the direct effect of addictive drugs on reinforcing drug-seeking S-R associations through their pharmacological effect on the dopaminergic circuit, they also pathologically subvert higher level learning mechanisms responsible for suppressing inflexible responses. Protracted exposure to drugs of abuse is widely reported to associate with behavioral deficits in tasks that require cognitive areas of the brain to be involved (Rogers and Robbins 2001; Almeida et al. 2008). Reduction in the activity of the PFC regions in abstinent addicts is also reported in many imaging studies (Goldstein and Volkow 2002; Volkow et al. 2004a). Interestingly, extended access to cocaine has shown to induce long-lasting impairments in working memory-dependent tasks, accompanied with decreased density of neurons in dorsomedial PFC (George et al. 2008). Considering the role of this region in goal-directed decision making, the atrophy of the associative cortex induced by drugs can further disrupt the balance between the goal-directed and habitual systems in favor of the latter. One simple way to model these morphological neuroadaptations in the dual-process framework is to assume that debility of the goal-directed system corresponds to its weakness in searching for the accurate estimated value of actions in the decision tree. Thus, reaching an acceptable level of accuracy (searching deep enough) to obtain "perfect information" will require more time (τ) in addicts, compared to healthy individuals. The assumption that the low performance of addicts in cognitive tasks can be modelled by a higher-than-normal τ can be tested by comparing the addicts' reaction time with that of healthy individuals, at the early stages of learning when responding is still goal-directed. Furthermore, since the deliberation time constitutes the cost of deliberation, another prediction that comes from this assumption is that addicts, because of having higher-than-normal deliberation cost, are less prone to deliberate and thus, more prone to make habitual responses than normal subjects. Thus, habitual responding for natural rewards must appear earlier

in addicts, compared to non-addict subjects. In addition, addicts must be more vulnerable than healthy subjects to commit actions with catastrophic consequences, not only in drug-associated cases, but also in other aspects of their daily lives.

The arbitration between the two systems is not only under the effect of long-lasting brain adaptations (like the two mechanisms described above: drug affects *VPI* and τ signals), but some transient changes in the brain decision making variables might also affect the arbitration between the two systems for a short period of time. For example, if for any reason the tonic dopamine, which is assumed to encode the average reward signal, increases for a certain period, the model predicts that the cost of deliberation will increase and thus, the individual will be more susceptible to make habitual responses during that period. This prediction of the model can explain why drug relapse is often precipitated by exposure to drug-associated cues, non-contingent drug injection, or stress (Shaham et al. 2003; Kalivas and McFarland 2003). In fact, the model explains that after prolonged abstinence, these three triggers of relapse revive the habitual system by increasing tonic dopamine and therefore, result in the dormant maladaptive habits to drive the behavior again toward drug consumption. Stress, as a potent trigger of relapse, has shown to increase extracellular concentration of dopamine in cortical and subcortical brain regions in both animal models of addiction (Thierry et al. 1976; Mantz et al. 1989) and humans (Montgomery et al. 2006). Intermittent tail-shock stress, for example, increases extracellular dopamine relative to the baseline by 39% and 95% in nucleus accumbens and medial frontal cortex, respectively (Abercrombie et al. 1989). Interestingly, protracted exposure to stress, similar to the effect of chronic drug consumption, results in the atrophy of the medial prefrontal cortex and the associative striatum, as well as hypertrophy of the sensorimotor striatum. These structural changes are accompanied with progressive behavioral insensitivity to the outcome of responses (Dias-Ferreira et al. 2009). This phenomenon can be explained in a similar argument proposed for explaining the long-lasting effect of drugs on the associative loop. Exposure to drug cues and drug-priming (non-contingent injection of drugs), as other triggers of relapse, are also well-known to increase extracellular dopamine for a considerable period of time (Di Chiara and Imperato 1988; Ito et al. 2002).

In sum, the dual-process model proposed above, explains the story of addiction in a scenario like this: at the early stages of drug self-administration, similar to responding for natural rewards, responding for drugs is controlled by the goal-directed system. After extensive training, as a result of a decrease in the *VPI* signal, as well as an increase in the average reward signal and deliberation time (as described before), the habitual behavior takes control over behavior. At this stage, as no drug is delivered to the animal anymore (extinction period), the average reward signal will drop significantly and thus, the goal-directed system will again take control over behavior. Finally, when a relapse trigger is experienced by the animal, the average reward signal increases again and thus, the habitual system can again come to the scene. Hence, the high values assigned to drug-seeking behavior by the habitual system will make the animal motivated to start responding for the drug again.

As explained above, this scenario is based on the assumption that the goal-directed system doesn't predict maladaptively high values for drug-seeking behaviors. This assumption implies that animals with inactivated brain regions underlying the habitual system should not develop compulsive behavior.

Furthermore, it is assumed that after the extinction period, the habitual system assigns a high value to drug seeking and taking behavior when the animal is exposed to relapse-triggering conditions. This property cannot be explained by the computational models of the S-R system introduced previously. This is because during the extinction phase, drug taking action is not followed by a drug reward and thus, drug seeking and taking actions lose their assigned values. This implies that the habitual system will not exhibit a compulsive behavior after extinction training. To explain the fact that drug-related behaviors regain high values after the animal faces relapse-triggers, it is necessary to incorporate more complicated mechanisms into the habitual system to represent the effect of relapse-triggers on habitual responding (see Redish et al. 2007 for the case of cue-induced relapse).

Finally, the explained scenario predicts that the reinstatement of drug seeking behavior is due to the transition of control from the goal-directed system to the habitual system. However, it is still unclear whether the drug seeking response after reinstatement is under the control of the habitual system or the goal directed system (Root et al. 2009).

8.5 Conclusion

Drug addiction is definitely a much more complicated phenomenon, both behaviorally and neurally, than the simplified image presented in this chapter. Neurally, different drugs have different sites of action and even for a certain drug like cocaine, the dopaminergic system is not the only circuit that is under the pharmacological effect. For example, serotonergic (Dhonnchadha and Cunningham 2008; Bubar and Cunningham 2008) and glutamatergic (Kalivas 2009) systems are also shown to be affected by drugs. However, the computational theory of reinforcement learning has proven to be an appropriate framework to approach this complex phenomenon. The great advantage of this framework is in its ability to bridge between behavioral and neural findings. Moreover, modeling DA receptors' availability within the actor-critic framework (Piray et al. 2010), as an example, shows that the RL framework is also potentially capable of modeling at least some of the detailed neuronal mechanisms.

There are still many steps to be taken in order to improve the current RL-based models of addiction. On important step is to develop an integrated model that can have all the three learning processes (Pavlovian, habitual and goal-directed) at the same time. Such a model would be expected to explain several behavioral aspects of addiction like loss of cognitive control, as well as the influence that Pavlovian predictors of drugs can exert on habitual and/or goal-directed systems (the role of PIT in cue-triggered relapse).

References

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

Ahmed SH (2004) Addiction as compulsive reward prediction. Science (New York) 306:1901–1902

Ahmed SH, Koob GF (2005) Transition to drug addiction: a negative reinforcement model based on an allostatic decrease in reward function. Psychopharmacology 180:473–490

Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13:266–271

Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9:357–381

Alexander GE, Crutcher MD, De Long MR (1990) Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. Prog Brain Res 85:119–146

Almeida PP, Novaes MAFP, Bressan RA, de Lacerda ALT (2008) Executive functioning and cannabis use. Rev Bras Psiquiatr (São Paulo, 1999) 30:69–76

American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders: DSM-IV-TR, 4th edn, Washington, DC

Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35:48–69

Belin D, Everitt BJ (2008) Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum. Neuron 57:432–441

Belin D, Mar AC, Dalley JW, Robbins TW, Everitt BJ (2008) High impulsivity predicts the switch to compulsive cocaine-taking. Science (New York) 320:1352–1355

Belin D, Jonkman S, Dickinson A, Robbins TW, Everitt BJ (2009) Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. Behav Brain Res 199:89–102

Bickel WK, Madden GJ (1999) A comparison of measures of relative reinforcing efficacy and behavioral economics: cigarettes and money in smokers. Behav Pharmacol 10:627–637

Bickel WK, Marsch LA (2001) Toward a behavioral economic understanding of drug dependence: delay discounting processes. Addiction (Abingdon) 96:73–86

Bubar MJ, Cunningham KA (2008) Prospects for serotonin 5-HT2R pharmacotherapy in psychostimulant abuse. Prog Brain Res 172:319–346

Dalley JW, Fryer TD, Brichard L, Robinson ESJ, Theobald DEH et al (2007) Nucleus accumbens d2/3 receptors predict trait impulsivity and cocaine reinforcement. Science (New York) 315:1267–1270

Dalley JW, Everitt BJ, Robbins TW (2011) Impulsivity, compulsivity, and top-down cognitive control. Neuron 69(4):680–694

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

Dayan P (2009) Dopamine, reinforcement learning, and addiction. Pharmacopsychiatry 42(1):S56–S65 Suppl

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

de Wit H (2009) Impulsivity as a determinant and consequence of drug use: a review of underlying processes. Addict Biol 14:22–31

Deroche-Gamonet V, Belin D, Piazza PV (2004) Evidence for addiction-like behavior in the rat. Science (New York) 305:1014–1017

Dezfouli A, Piray P, Keramati MM, Ekhtiari H, Lucas C et al (2009) A neurocomputational model for cocaine addiction. Neural Comput 21:2869–2893

Dhonnchadha BAN, Cunningham KA (2008) Serotonergic mechanisms in addiction-related memories. Behav Brain Res 195:39–53

Di Chiara G, Imperato A (1988) Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. Proc Natl Acad Sci USA 85:5274–5278

Dias-Ferreira E, Sousa JC, Melo I, Morgado P, Mesquita AR et al (2009) Chronic stress causes frontostriatal reorganization and affects decision-making. Science (New York) 325:621–625

Dickinson A, Balleine BW (2002) The role of learning in motivation. In: Gallistel CR (ed) Steven's handbook of experimental psychology: learning, motivation, and emotion, 3rd edn, vol 3. Wiley, New York, pp 497–533

Estes WK (1948) Discriminative conditioning; effects of a pavlovian conditioned stimulus upon a subsequently established operant response. J Exp Psychol 38:173–177

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

Everitt BJ, Belin D, Economidou D, Pelloux Y, Dalley JW et al (2008) Neural mechanisms underlying the vulnerability to develop compulsive drug-seeking habits and addiction. Philos Trans R Soc Lond B, Biol Sci 363:3125–3135

Ford CP, Mark GP, Williams JT (2006) Properties and opioid inhibition of mesolimbic dopamine neurons vary according to target location. J Neurosci 26:2788–2797

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci USA 104:16311–16316

Garavan H, Pankiewicz J, Bloom A, Cho JK, Sperry L et al (2000) Cue-induced cocaine craving: neuroanatomical specificity for drug users and drug stimuli. Am J Psychiatry 157:1789–1798

George O, Mandyam CD, Wee S, Koob GF (2008) Extended access to cocaine self-administration produces long-lasting prefrontal cortex-dependent working memory impairments. Neuropsychopharmacology 33:2474–2482

Goldstein RZ, Volkow ND (2002) Drug addiction and its underlying neurobiological basis: neuroimaging evidence for the involvement of the frontal cortex. Am J Psychiatry 159:1642–1652

Goldstein RZ, Alia-Klein N, Tomasi D, Zhang L, Cottone LA et al (2007) Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? Am J Psychiatry 164:43–51

Haber SN (2003) The primate basal ganglia: parallel and integrative networks. J Chem Neuroanat 26:317–330

Haber SN, Fudge JL, McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J Neurosci 20:2369–2382

Holland PC (2004) Relations between pavlovian-instrumental transfer and reinforcer devaluation. J Exp Psychol, Anim Behav Processes 30:104–117

Hopf FW, Cascini MG, Gordon AS, Diamond I, Bonci A (2003) Cooperative activation of dopamine d1 and d2 receptors increases spike firing of nucleus accumbens neurons via g-protein betagamma subunits. J Neurosci 23:5079–5087

Ikemoto S (2007) Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. Brains Res Rev 56:27–78

Ikemoto S, Glazier BS, Murphy JM, McBride WJ (1997) Role of dopamine d1 and d2 receptors in the nucleus accumbens in mediating reward. J Neurosci 17:8580–8587

Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci 29:9861–9874

Ito R, Dalley JW, Robbins TW, Everitt BJ (2002) Dopamine release in the dorsal striatum during cocaine-seeking behavior under the control of a drug-associated cue. J Neurosci 22:6247–6253

Jacobs EA, Bickel WK (1999) Modeling drug consumption in the clinic using simulation procedures: demand for heroin and cigarettes in opioid-dependent outpatients. Exp Clin Psychopharmacol 7:412–426

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

Johnson PM, Kenny PJ (2010) Dopamine d2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats. Nat Neurosci 13:635–641

Kalivas PW (2009) The glutamate homeostasis hypothesis of addiction. Nat Rev, Neurosci 10:561–572

Kalivas PW, McFarland K (2003) Brain circuitry and the reinstatement of cocaine-seeking behavior. Psychopharmacology 168:44–56

Kamin L (1969) Predictability, surprise, attention, and conditioning. In: Campbell BA, Church RM (eds) Punishment and aversive behavior. Appleton-Century-Crofts, New York, pp 279–296

Keramati M, Dezfouli A, Piray P (2011) Speed-accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol 7(5):1–25

Koob GF, Le Moal M (2005a) Plasticity of reward neurocircuitry and the 'dark side' of drug addiction. Nat Neurosci 8:1442–1444

Koob GF, Le Moal M (2005b) Neurobiology of addiction. Academic Press, San Diego

Logue A, Tobin H, Chelonis J, Wang R, Geary N et al (1992) Cocaine decreases self-control in rats: a preliminary report. Psychopharmacology 109:245–247

Lovibond PF (1983) Facilitation of instrumental behavior by a pavlovian appetitive conditioned stimulus. J Exp Psychol, Anim Behav Processes 9:225–247

Mackintosh NJ (1974) The psychology of animal learning. Academic Press, London

Mahadevan S (1996) Average reward reinforcement learning: foundations, algorithms, and empirical results. Mach Learn 22:159–195

Mantz J, Thierry AM, Glowinski J (1989) Effect of noxious tail pinch on the discharge rate of mesocortical and mesolimbic dopamine neurons: selective activation of the mesocortical system. Brain Res 476:377–381

Marks KR, Kearns DN, Christensen CJ, Silberberg A, Weiss SJ (2010) Learning that a cocaine reward is smaller than expected: a test of redish's computational model of addiction. Behav Brain Res 212:204–207

Martinez D, Slifstein M, Narendran R, Foltin RW, Broft A et al (2009) Dopamine d1 receptors in cocaine dependence measured with PET and the choice to self-administer cocaine. Neuropsychopharmacology 34:1774–1782

Montgomery AJ, Mehta MA, Grasby PM (2006) Is psychological stress in man associated with increased striatal dopamine levels?: a [11C]raclopride PET study. Synapse (New York) 60:124–131

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057–1063

Nader MA, Daunais JB, Moore T, Nader SH, Moore RJ et al (2002) Effects of cocaine self-administration on striatal dopamine systems in rhesus monkeys: initial and chronic exposure. Neuropsychopharmacology 27:35–46

Nader MA, Czoty PW, Gould RW, Riddick NV (2008) Positron emission tomography imaging studies of dopamine receptors in primate models of addiction. Philos Trans R Soc Lond B, Biol Sci 363:3223–3232

Nicola SM (2007) The nucleus accumbens as part of a basal ganglia action selection circuit. Psychopharmacology 191:521–550

Niv Y (2007) The effects of motivation on habitual instrumental behavior. PhD thesis, The Hebrew University of Jerusalem, Interdisciplinary Center for Neural Computation

Niv Y, Daw ND, Joel D, Dayan P (2007) Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology 191:507–520

O'Brien CP, Ehrman R, Ternes J (1986) Classical conditioning in human opioid dependence. In: Goldberg SR, Stolerman IP (eds) Behavioral analysis of drug dependence, 1st edn. Academic Press, London, pp 329–356

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K et al (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science (New York) 304:452–454

Paine TA, Dringenberg HC, Olmstead MC (2003) Effects of chronic cocaine on impulsivity: relation to cortical serotonin mechanisms. Behav Brain Res 147:135–147

Panlilio LV, Thorndike EB, Schindler CW (2007) Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. Pharmacol Biochem Behav 86:774–777

Pelloux Y, Everitt BJ, Dickinson A (2007) Compulsive drug seeking by rats under punishment: effects of drug taking history. Psychopharmacology 194:127–137
Petry NM, Bickel WK (1998) Polydrug abuse in heroin addicts: a behavioral economic analysis. Addiction (Abingdon) 93:321–335
Pierce RC, Vanderschuren LJMJ (2010) Kicking the habit: the neural basis of ingrained behaviors in cocaine addiction. Neurosci Biobehav Rev 35:212–219
Piray P, Keramati MM, Dezfouli A, Lucas C, Mokri A (2010) Individual differences in nucleus accumbens dopamine receptors predict development of addiction-like behavior: a computational approach. Neural Comput 22:2334–2368
Porrino LJ, Daunais JB, Smith HR, Nader MA (2004a) The expanding effects of cocaine: studies in a nonhuman primate model of cocaine self-administration. Neurosci Biobehav Rev 27:813–820
Porrino LJ, Lyons D, Smith HR, Daunais JB, Nader MA (2004b) Cocaine self-administration produces a progressive involvement of limbic, association, and sensorimotor striatal domains. J Neurosci 24:3554–3562
Potenza MN (2008) The neurobiology of pathological gambling and drug addiction: an overview and new findings. Philos Trans R Soc Lond B, Biol Sci 363:3181–3189
Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev, Neurosci 9:545–556
Redish AD (2004) Addiction as a computational process gone awry. Science (New York) 306:1944–1947
Redish AD, Jensen S, Johnson A, Kurth-Nelson Z (2007) Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol Rev 114:784–805
Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31:415–437; discussion 437–487
Reynolds B (2006) A review of delay-discounting research with humans: relations to drug use and gambling. Behav Pharmacol 17:651–667
Robinson ESJ, Eagle DM, Economidou D, Theobald DEH, Mar AC et al (2009) Behavioural characterisation of high impulsivity on the 5-choice serial reaction time task: specific deficits in 'waiting' versus 'stopping'. Behav Brain Res 196:310–316
Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615–1624
Roesch MR, Singh T, Brown PL, Mullins SE, Schoenbaum G (2009) Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. J Neurosci 29:13365–13376
Rogers RD, Robbins TW (2001) Investigating the neurocognitive deficits associated with chronic drug misuse. Curr Opin Neurobiol 11:250–257
Root DH, Fabbricatore AT, Barker DJ, Ma S, Pawlak AP et al (2009) Evidence for habitual and goal-directed behavior following devaluation of cocaine: a multifaceted interpretation of relapse. PLoS ONE 4:e7170
Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA et al (2009) Dopaminergic drugs modulate learning rates and perseveration in parkinson's patients in a dynamic foraging task. J Neurosci 29:15104–15114
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science (New York) 275:1593–1599
Shaham Y, Shalev U, Lu L, Wit HD, Stewart J (2003) The reinstatement model of drug relapse: history, methodology and major findings. Psychopharmacology 168:3–20
Simon NW, Mendez IA, Setlow B (2007) Cocaine exposure causes long-term increases in impulsive choice. Behav Neurosci 121:543–549
Smith AJ, Li M, Becker S, Kapur S (2006) Linking animal models of psychosis to computational models of dopamine function. Neuropsychopharmacology 32:54–66
Steeves TDL, Miyasaki J, Zurowski M, Lang AE, Pellecchia G et al (2009) Increased striatal dopamine release in parkinsonian patients with pathological gambling: a [11C] raclopride PET study. Brain 132:1376–1385

Stewart J (2008) Psychological and neural mechanisms of relapse. Philos Trans R Soc Lond B, Biol Sci 363:3147–3158

Stuber GD, Wightman RM, Carelli RM (2005) Extinction of cocaine self-administration reveals functionally and temporally distinct dopaminergic signals in the nucleus accumbens. Neuron 46:661–669

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Thierry AM, Tassin JP, Blanc G, Glowinski J (1976) Selective activation of mesocortical DA system by stress. Nature 263:242–244

Vanderschuren LJMJ, Everitt BJ (2004) Drug seeking becomes compulsive after prolonged cocaine self-administration. Science (New York) 305:1017–1019

Volkow ND, Fowler JS, Wang G (2004a) The addicted human brain viewed in the light of imaging studies: brain circuits and treatment strategies. Neuropharmacology 47(1):3–13 Suppl

Volkow ND, Fowler JS, Wang G, Swanson JM (2004b) Dopamine in drug abuse and addiction: results from imaging studies and treatment implications. Mol Psychiatry 9:557–569

Volkow ND, Wang G, Fowler JS, Telang F (2008) Overlapping neuronal circuits in addiction and obesity: evidence of systems pathology. Philos Trans R Soc Lond B, Biol Sci 363:3191–3200

Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. Nat Rev, Neurosci 7:464–476

Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181–189

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22:513–523

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

Zapata A, Minney VL, Shippenberg TS (2010) Shift from Goal-Directed to habitual cocaine seeking after prolonged experience in rats. J Neurosci 30:15457–15463