LETTER Communicated by A. David Redish

A Neurocomputational Model for Cocaine Addiction

Amir Dezfouli
a.dezfouli@ut.ac.ir
Payam Piray
piray@ut.ac.ir
Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

Mohammad Mahdi Keramati
mm_keramati@gsme.sharif.edu
School of Management and Economics, Sharif University of Technology, Tehran, Iran

Hamed Ekhtiari
h_ekhtiari@razi.tums.ac.ir
Neurocognitive Laboratory, Iranian National Center for Addiction Studies, Tehran, Iran

Caro Lucas
lucas@ut.ac.ir
Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

Azarakhsh Mokri
mokriazr@sina.tums.ac.ir
Department of Psychiatry, Tehran University of Medical Sciences, Tehran, and Department of Clinical Sciences, Iranian National Center for Addiction Studies, Tehran, Iran

Based on the dopamine hypotheses of cocaine addiction and the assumption of decrement of brain reward system sensitivity after long-term drug exposure, we propose a computational model for cocaine addiction. Utilizing average reward temporal difference reinforcement learning, we incorporate the elevation of basal reward threshold after long-term drug exposure into the model of drug addiction proposed by Redish. Our model is consistent with the animal models of drug seeking under punishment. In the case of nondrug reward, the model explains increased impulsivity after long-term drug exposure. Furthermore, the existence of a blocking effect for cocaine is predicted by our model.

1 Introduction

Cocaine is a powerfully addictive stimulant drug that is obtained from the leaves of the coca plant, *Erythroxylon coca*. It became popular in the 1980s and 1990s, and addiction to it is one of the society's greatest problems today.

Serious medical complications such as cardiovascular, respiratory, and neurological effects are associated with abuse of cocaine. However, its primary target of action is the central nervous system. Cocaine acts within neurochemical systems that are part of a motivational system neurocircuitry. The brain's motivational system enables a person to interpret and behaviorally respond to important environmental stimuli such as food, water, sex, or dangerous situations. After repeated drug use, pathological changes in a vulnerable brain due to the actions of cocaine lead to impaired behavioral responses to motivationally relevant events.

Certain elements of such maladaptive orientation to the environment seem to be shared across different drug classes: compulsive drug-seeking and drug-taking behavior even at the expense of adverse behavioral, social and health consequences (American Psychiatric Association, 2000) and a decrease in natural reward processing and reduced motivation for natural rewards (Kalivas & O'Brien, 2007). A comprehensive theory of addiction can explain these two cardinal features of drug addiction. The explanation must establish reductive links across neurobiology and behavior, assuming reductionism is a useful way of examining addiction.

The features noted result from pathological changes in motivation and choice (Kalivas & Volkow, 2005). A choice is produced by a decision-making process, that is, choosing an option or course of action from among a set of alternatives. In this study we use decision-making frameworks derived from reinforcement learning (RL) theory (Sutton & Barto, 1998) for an explanation of the features we have noted. The connection between RL models and neuroscience has been widely studied, which makes them suitable for developing neurocomputational models of decision making. And since computational models use mathematical language, the use of them provides explicit prediction capability and coherence in any description of structural and behavioral evidence. In the rest of this section, we introduce the RL framework. Next, based on a variant of RL models, we present the computational model of addiction proposed by Redish (2004). Finally, we discuss how well his model addresses the features of drug addiction.

RL deals with the problem of decision making in an uncertain environment. An RL agent perceives the environment in the form of states and rewards. In each state is a set of possible actions, and the agent should decide which one to take. In value-based decision-making frameworks (Rangel, Camerer, & Montague, 2008) such as RL, an agent chooses from among several alternatives on the basis of a subjective value that it has assigned to them. The value that the agent assigns to an action in a state represents the cumulative sum of all future rewards that the agent can gain by taking the

action. If we denote the value of action a_t in state s_t with $Q(s_t, a_t)$ we will have

$$Q(s_t, a_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \mid s_t, a_t] = E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \mid s_t, a_t\right]. \tag{1.1}$$

The factor $0 < \gamma < 1$, called the discount factor, determines the relative weighting of rewards earned earlier and those received later. r_i is the reward observed at time i.

In order to act optimally, the agent needs to know the estimated value of each action in each state. Estimated values of state-action pairs are learned from the rewards the agent gains after executing actions. A variant of RL known as temporal difference reinforcement learning (TDRL) uses an error signal, which is the difference between estimated value before taking an action and experienced value after its execution, for learning state-action values. Formally, if at time t the agent executes action a_t and leaves state s_t to state s_{t+1} with transition time d_t and receives reward r_{t+1}, then the error signal is calculated as follows:

$$\delta_t = \gamma^{d_t}(r_{t+1} + V(s_{t+1})) - Q(s_t, a_t), \tag{1.2}$$

where $V(s_{t+1})$ is the value of state s_{t+1}. Experiments show that $V(s_{t+1})$ can be considered in two ways: the maximum value of actions in state s_{t+1} (Roesch, Calu, & Schoenbaum, 2007) or the value of the action that the agent will take in state s_{t+1} (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006). In the former case, the learning method is called Q-learning (Watkins, 1989).

Taking the error signal into account, the updating rule for state-action values is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t, \tag{1.3}$$

where α is the learning rate parameter, which determines the extent to which new experiences affect subjective values.

Beside behavioral psychology accounts of TDRL, from a neurocomputational viewpoint TDRL has well-known neural substrates (Daw & Doya, 2006). It is now a classic observation that the error signal, δ_t, qualitatively corresponds to phasic activity of dopaminergic (DA) neurons in the ventral tegmental area (VTA) (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Based on this property of TDRL and taking into account the observation that cocaine produces a transient increase in nucleus accumbens (NA) and other VTA DA (Aragona et al., 2008; Kauer & Malenka, 2007), Redish (2004) proposed that drug-induced alterations can

be modeled by an additional term, $D(s_t)$. This modification causes the error signal to be always greater than $D(s_t)$:

$$\delta_t^c = \max(\gamma^{d_t}(r_{t+1} + V(s_{t+1})) - Q(s_t, a_t)) + D(s_t),\ D(s_t)). \tag{1.4}$$

The term $D(s_t)$ models the phasic activity of DA neurons in the NA induced by the pharmacological effects of cocaine. This activity of DA neurons probably corresponds to cocaine-induced spontaneous DA transients. These transients are correlated with brain cocaine concentrations and are not associated with operant response and environmental stimuli (Stuber, Wightman, & Carelli, 2005). Hence, it can be inferred that the term is uncompensatable during the learning process and explains why the error signal is always greater than $D(s_t)$.

The term r_t in equation 1.4 is a component of the drug reward, which models euphoria produced by cocaine. It depends on the drug's pure neuropharmacological effects, $D(s_t)$, but probably it is not identical to it. In fact, Chen et al. (2006) clearly showed that inhibition of dopamine transporter (DAT) is required for the rewarding effect of cocaine (Tilley, O'Neill, Han, & Gu, 2009) as measured by conditioned place preference. Although it can be inferred from this evidence that $D(s_t)$ is a determinant required for the drug reward, it cannot be inferred that it is the only determinant. Indeed, r_t cannot be defined completely with respect to $D(s_t)$, and other non-DA factors contribute to cocaine-induced euphoria. Hence, although $D(s_t)$ and r_t are related to each other, modeling drug reward with two components does not challenge the validity of the model.

The model satisfactorily establishes the desired link among the neural mechanism of decision making, the drug-induced alterations, and behavioral evidence. The definition of the error signal in equation 1.4 implies that with each experience of the drug, drug intake estimated value increases by at least $\alpha D(s_t)$ factor. Assuming that decisions are made based on their relative values, the model implies insensitivity of choices to the costs associated with drug abuse: by repeated drug abuse, the estimated value of drug consumption grows and outweighs the cost of harmful consequences associated with drug seeking and abuse. This implication is consistent with both animal models of compulsive drug seeking (Deroche-Gamonet, Belin, & Piazza, 2004; Mantsch, Ho, Schlussman, & Kreek, 2001; Paterson & Markou, 2003; Vanderschuren & Everitt, 2004) and human behavior.

An examination of the shortcomings of the model implies that the estimated value of drug intake grows unboundedly with drug consumption. This seems to be implausible where biological mechanisms limit the maximum value of the states (Redish, 2004). This problem is discussed by Redish, and to tackle the problem, the incorporation of a new factor, effectiveness of DA, into the model is proposed. This factor determines the

weighting of $D(s_t)$ and r_t in the learning of state-action values. Along with the drug consumption, the value of the factor is decreased and causes the value of states to remain finite. This factor is designed to model the biological mechanisms that limit the cocaine effect and solves the problem of implementing infinite values in the neural system. The benefit of this variable in the decision-making process at the algorithmic level is not clear, though.

Redish's model predicts that Kamin's blocking effect (Kamin, 1969) does not occur with cocaine. Blocking is a classical conditioning phenomenon that demonstrates that an animal is blocked from learning an association between a stimulus and a result if the stimulus is reinforced simultaneously with a different stimulus already associated with the result. Under the assumption that learning is based on the error signal, when a stimulus is already learned and completely predicts the result, then $\delta_t = 0$. Because the error signal is zero, the value of another stimulus reinforced with the previously learned stimulus is not updated and hence will not be associated with the result. Since cocaine always produces a positive error signal, its value is always better than its predicted value. Thus, a stimulus cannot block the association of another stimulus with the drug. This prediction of the model that cocaine does not show blocking effect proves wrong (Panlilio, Thorndike, & Schindler, 2007).

Most important, the model does not address the second feature of drug addiction. In situations where decision making is involved with natural reinforcers, the model predicts that it remains healthy by prior experience of the drug. This implication of the model is not consistent with evidence for the long-term changes in the processing of natural rewards in both cocaine-addicted animals and humans (Ahmed, 2004).

In this letter, we introduce a new computational model of cocaine addiction that is basically an extension to Redish's model. In order to address the problems we have noted, we have modified its structure based on neural evidence. In section 3, we validate our model by comparing its behavior with experimental data. We end by discussing the model's predictions, abilities and limitations and compare it with other models of drug addiction. We also present, the ways our model can be extended in future work.

2 The Model

Cocaine causes a transient increase in the level of DA, which leads to overvaluation of the drug intake. Beside this effect, evidence shows that chronic drug exposure causes long-lasting dysregulation of the reward processing system (Koob & Moal, 2005).

Ahmed and Koob (1998, 1999) studied the effect of limited versus extended access to cocaine self-administration on the brain reward threshold,

which is required for environmental events to activate DA cells. They observed that the brain reward threshold did not change in rats with limited access to cocaine, but it became progressively elevated in rats with extended access to cocaine. There are few systematic studies (Grace, 1995, 2000) that directly and convincingly report neural substrates of reward thresholds; however, Ahmed and Koob (2005) have suggested it may correspond to a specific tonus of accumbal DA. Tonic activity of DA neurons refers to baseline steady-state DA levels with slow firing at frequencies around 5 Hz, in contrast to phasic activity with fast burst firing at frequencies greater than 30 Hz.

Consistent with the elevation of the reward threshold after long-term drug abuse, Garavan et al. (2000) found that experienced cocaine addicts show impaired prefrontal activity in response to sexually evocative visual stimuli compared to normal human subjects. Moreover, decreased sensitivity to monetary reward in cocaine-addicted human subjects is reported (Goldstein et al., 2007).

Brain-imaging studies reveal another aspect of the long-term effects of drug abuse on the brain reward system. PET studies measuring DA D2 receptors have consistently shown long-lasting reduction in D2 DA receptor availability in the striatum (Volkow, Fowler, Wang, & Swanson, 2004; Nader et al., 2006). DA D2 receptors mediate the positive reinforcing properties of drugs of abuse and probably natural rewards (Volkow et al., 2004). This finding, coupled with evidence of decreased DA cell activity in cocaine addicts (Volkow, Fowler, Wang, Swanson, & Telang, 2007), would result in decreased output of DA circuits related to reward.

These studies suggest that long-term drug exposure causes an important alteration in the brain reward system. Due to this alteration, the level against which rewards are measured becomes abnormally elevated. In other words, long-term drug abuse elevates the basal reward level to a level that is higher than that of normal subjects. This elevation corresponds to the elevation of brain reward threshold, and it models a decrease in DA functioning.

From a computational modeling point of view, like values of state-action pairs, basal reward level can be considered an internal variable in the decision-making system. In order to investigate the effect of long-term drug abuse on this variable and its behavioral implications, four questions should be answered: (1) How do behavior and stimuli determine the value of basal reward level? (2) What is the role of this variable in decision making? (3) How is the variable implemented in the neural system? Finally, in order to investigate long-term effects of the drug abuse, we should consider a fourth question: (4) How can the effects of cocaine on the DA reward system be modeled by this variable? Fortunately, these questions can be addressed in the TD model of DA system proposed by Daw (2003). In the rest of this section, we introduce Daw's model in order to answer the first three questions. Then, to model cocaine effects on the brain DA reward system, we modify his framework.

Daw's model is based on average reward RL (Mahadevan, 1996). In this model, state-action values represent sums of differences between observed rewards and average reward:

$$Q(s_t, a_t) = E\left[\sum_{i=t}^{\infty}(r_i - \bar{R}_i) \mid s_t, a_t\right], \tag{2.1}$$

where $\bar{R}_t$ is the average reward per action. It is computed as an exponentially weighted moving average of experienced rewards:

$$\bar{R}_{t+1} \leftarrow (1-\sigma)\bar{R}_t + \sigma r_t, \tag{2.2}$$

where $\sigma \ll \alpha$.

Similar to the framework introduced in section 1, this model uses TDRL method for learning state-action values. The error signal is computed as follows:

$$\delta_t = r_t + V(s_{t+1}) - Q(s_t, a_t) - \bar{R}_t. \tag{2.3}$$

It appears from equation 2.3 that for updating state-action values, $Q(s_t, a_t)$, the undiscounted value of state s_{t+1} is used in the learning process. In fact, in this TDRL method, unlike the one already presented, instead of discounting future rewards value exponentially, future rewards are discounted in a manner related to hyperbolic discounting. Such definition of the error signal does not imply that the value of a state is insensitive to the arrival time of future reward. Rather, in equation 2.3, the average reward is subtracted from $V(s_{t+1})$. By waiting in state s_t for one time step, the model loses an opportunity to gain a future reward. This opportunity loss is equal to $\bar{R}_t$ and is subtracted from the value of the next state. On the whole, the learning method guides action selection to the policy that maximizes the expected reward per step.

Returning to the questions, the value of the basal reward level is determined by experienced rewards from equation 2.2, and it corresponds to the average reward. It is incorporated into the decision-making process through value learning of the state-action pairs, as described in equation 2.3. Daw proposed that the tonic level of DA codes the average reward (Daw, 2003; Daw & Touretzky, 2002; Niv, Daw, Joel, & Dayan, 2007. Based on this suggestion, it is better to interpret $\bar{R}_t$ as the level against which the phasic activity of DA neurons, δ_t, appears. However, in the TDRL framework, the reinforcing efficacy of reward is mediated by the error signal, which makes it reasonable to consider $\bar{R}_t$ as the level against which rewards are measured.

Concerning the fourth question, the effect of cocaine abuse in the phasic activity of DA neurons can be incorporated into the error signal similar to Redish's model:

$$\delta_t^c = \max(r_t + V(s_{t+1}) - Q(s_t, a_t) + D(s_t),\, D(s_t)) - \bar{R}_t, \tag{2.4}$$

where $\bar{R}_t$ is out of the maximization operator because it is not related to the phasic activity of DA neurons.

With the above definition for the cocaine-induced error signal, the average reward cannot be updated only by the r_t term. Indeed, in order to preserve the consistency of the model, the effect of $D(s_t)$ should be reflected in the average reward. This can be done by rewriting experienced reward, r_t, with respect to the error signal using equation 2.3,

$$r_t = \delta_t - V(s_{t+1}) + Q(s_t, a_t) + \bar{R}_t, \tag{2.5}$$

and with substitution of the cocaine-induced error signal, δ_t^c, we will have

$$r_t^c = \delta_t^c - V(s_{t+1}) + Q(s_t, a_t) + \bar{R}_t. \tag{2.6}$$

In the case of the cocaine reward, the average reward is updated using r_t^c, and in the case of the natural reward, it is updated using r_t.

The other effect of cocaine, the long-term effect, is to elevate the basal reward level abnormally. According to the above framework, the normal level of the basal reward, which leads to optimal decision making, is derived from the average reward. Due to long-term drug abuse, the basal reward level deviates from its normal level. Thus, it is no longer equal to the average reward, $\bar{R}_t$, and we denote it with ρ_t. The deviation of the basal reward level, ρ_t, from its normal value can be modeled by a new term. The term denoted by κ_t represents the quantity of the deviation of basal reward level from the average reward. With the basal reward level biased by κ_t, we have

$$\rho_t = \bar{R}_t + \kappa_t. \tag{2.7}$$

Based on the above definition, the basal reward level has two components. The first component, equal to the average reward, $\bar{R}_t$, is considered the normal value of the basal reward level and guides the decision-making process to the optimal choices. The second component models the effect of long-term drug use on the reward system. This component, κ_t, shifts the basal reward level to a level higher than its normal value, which is the average reward. During drug exposure, κ_t grows and elevates the basal reward level abnormally. This elevation, modeled by κ_t, corresponds to the elevation of the brain reward threshold after long-term drug exposure. Furthermore, as κ_t is subtracted from the error signal corresponding to the

output of the DA system, its elevation models the decreased function of the DA system after long-term drug abuse. In a healthy decision-making system with no prior experience of the drug, the value of this parameter, κ_t, is stable at zero. On the contrary, each experience of the drug elevates it slowly:

$$\kappa_{t+1} \leftarrow (1-\lambda)\kappa_t + \lambda N, \tag{2.8}$$

where N represents the maximum level of deviation and λ controls the speed of deviation, which is smaller than σ, $\lambda \ll \sigma$. In contrast to the drug reward, with the experience of natural reward, the deviation declines gradually:

$$\kappa_{t+1} \leftarrow (1-\lambda)\kappa_t. \tag{2.9}$$

Regarding this modification to Daw's model, $\bar{R}_t$ is substituted with ρ_t in equations 2.3 to 2.6. Under the assumption that the basal reward level is coded by tonic DA, the abnormal elevation of basal reward level is neurally correlated with an increase in tonic DA activity. Whereas the ability of cocaine to increase brain DA levels is partially dependent on tonic DA activity, it is expected that with the increase in the activity of tonic DA after chronic drug exposure, cocaine loses its previous ability to increase brain DA levels (Mateo, Lack, Morgan, Roberts, & Jones, 2005; Volkow et al., 1997). The increase in brain DA levels due to the effects of cocaine is modeled by $D(s_t)$; thus, the effect of $D(s_t)$ on the error signal should decrease after chronic drug abuse. This can be seen in our model if we substitute $\bar{R}_t$ for ρ_t in equation 2.4 and rewrite it as follows:

$$\delta_t^c = max(r_t + V(s_{t+1}) - Q(s_t, a_t) + [D(s_t) - \kappa_t], [D(s_t) - \kappa_t]) - \bar{R}_t. \tag{2.10}$$

Thus, κ_t is subtracted from $D(s_t)$. With the elevation of κ_t after chronic drug abuse, the effect of $D(s_t)$ on the error signal decreases. Hence, the model is consistent with the fact that the ability of cocaine to increase brain DA levels decreases after long-term exposure to the drug.

This definition of the basal reward level is consistent with the evidence noted about the decrement of reward system sensitivity after prolonged drug exposure. Moreover, as the decision-making system is common in the natural and drug reinforcers, it is expected that deviation of the basal reward level from its normal value will have adverse effects on decision making in the case of natural rewards. In the next section, through simulations we will show the behavior of the model in different procedures.

3 Results

3.1 Value Learning. The model is simulated in the procedure shown in Figure 1 with the drug reward (see appendix A for details of the

Figure 1: Learning the value of a reward. Pressing the lever (PL) leads to the reward (R) delivery.

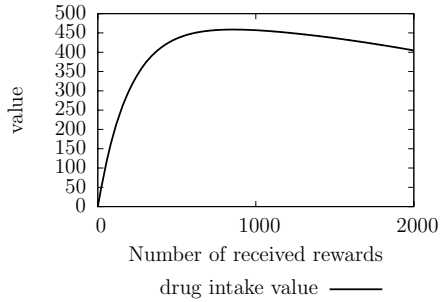

Figure 2: Estimated value of the drug. Estimated value of the drug reward does not increase unboundedly with consumption. A decrease in the estimated value of the drug after long-term abuse is due to an abnormal elevation of the basal reward level.

simulations). In this procedure, the model learns that the action of pressing the lever leads to the drug delivery. As Figure 2 illustrates, the estimated value of the drug reward does not increase unboundedly as drug exposure duration increases. Additionally, the figure shows that after long-term drug abuse, the estimated value of the drug decreases. This is because of the abnormal elevation of the basal reward level due to an increase of κ_t.

Figure 3 shows the error signal during value learning of a natural reward in the procedure shown in Figure 1. As the figure illustrates, an increase in the duration of prior drug exposure decreases the value of the error signal, and this decrease leads to a decline in the estimated value of the reward. Hence, under the assumption that state-action values and the error signal are correlated with the activity of specific human brain regions (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), decreased activity of the regions in response to a motivational stimulus is predicted. Also, as the figure shows, the prediction error falls to near zero after repeated drug intake, due to the compensation mechanism in the model realized by the basal reward level. The elevation of the basal reward level is not limited to the drug; it plays the same role in the case of natural rewards. Even in the absence of deviation of the basal reward level after long-term drug abuse, an elevated basal reward level will cancel out the effect of $D(s_t)$ and r_t on the error signal.

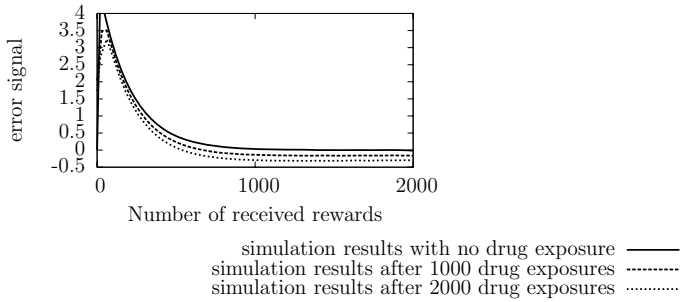

Figure 3: Effect of different durations of drug exposure on the error signal during the value learning process of a natural reward. The longer the duration of prior drug exposure is, the lower the error signal will be.

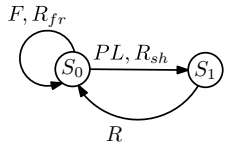

Figure 4: Simulation of compulsive drug seeking. The model has to choose between freezing action (F) and pressing the lever (PL). Choosing PL is followed by the shock punishment, R_{sh}, and the reward (R) afterward. Choosing F, the model receives a small punishment R_{fr} and will not receive the reward. The procedure is simulated for the drug reward and a natural reward separately.

3.2 Compulsive Drug Seeking. Based on unlimited access to drug SA, different animal models have been proposed for increased insensitivity toward punishments associated with drug seeking (Deroche-Gamonet et al., 2004; Mantsch et al., 2001; Paterson & Markou, 2003; Vanderschuren & Everitt, 2004). For example, Vanderschuren and Everitt (2004) studied how an adverse consequence affects drug-seeking behavior after limited and prolonged cocaine SA. They found that an aversive conditioned stimulus (CS) that had been independently paired with a shock suppresses drug seeking after limited drug SA; after prolonged drug SA, the aversive CS did not affect the seeking responses. This pattern of drug seeking was not observed in the case of sucrose seeking as a natural reinforcer; even after prolonged sucrose consumption, aversive CS suppressed the seeking behavior.

We simulate the model in the procedure shown in Figure 4. In the procedure, the model has to choose between freezing action (F) and pressing the lever (PL). Choosing PL is followed by the shock punishment, R_{sh}, and the reward (R) afterward. Choosing F, it receives a small punishment, R_{fr}, and the model will not receive the reward.

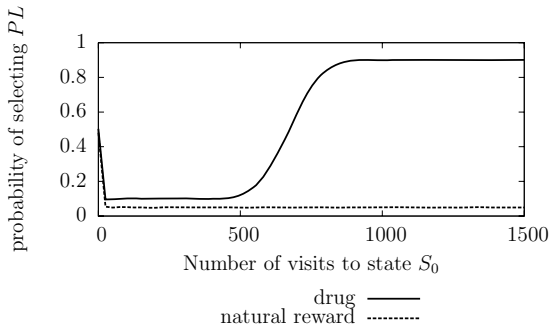

Figure 5: Probability of pressing the lever (PL) when responses are punished severely. Shock suppresses drug taking after limited access to cocaine. But after prolonged use, the punishment cannot suppress drug taking. In the case of the natural reward, the shock suppresses responses after limited and prolonged natural reward consumption.

The model is simulated for the drug reward and a natural reward (e.g., sucrose) separately. Figure 5 shows the probability of selecting PL. In the case of drug reward, because the estimated value of PL is less than F in limited use, the probability of selecting PL is low. But after extended use, the adverse consequence (i.e., shock) cannot suppress drug taking, and thus the probability of selecting PL becomes high despite the punishment of the shock. Additionally, the figure shows the behavior of the model in the case of the natural reward. Responding to the natural reinforcer even after prolonged consumption is suppressed by the shock punishment.

Relative insensitivity of drug seeking to punishment is due to an increase in the estimated value of drug intake after repeated drug use. In our model, unlike Redish's, an increase in the estimated value of drug intake is not because the term $D(s_t)$ is uncompensatable. Indeed, the assumption of a high reward value for the drug (produced by inhibition of DAT) is necessary for the emergence of a developing insensitivity to punishment. This is because the final estimated value for the drug is an increasing function of $D(s_t)$ and r_t. For sufficiently small values of them, the final estimated value of the drug will not have a high value and hence cannot cancel out the cost of punishment in decision making.

In Redish's model, the assumption of a high value for $D(s_t)$ is not required for modeling developing insensitivity to punishment. Even if the term has a small value, after sufficiently long drug exposure, its estimated value will be high. In his model, if we make the term $D(s_t)$ compensatable by removing the maximization operator in equation 1.4, the drug's estimated value becomes stable after a few learning steps. Hence, under this condition, the behavior of the model does not differ after long-term drug exposure and

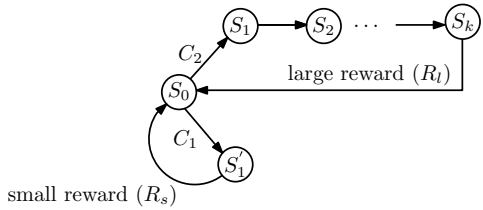

Figure 6: Delayed discounting task. A decision maker has two choices, C_1 and C_2. Selection of C_1 after one time step is followed by a small reward, R_s, while choosing C_2, the decision maker should wait k time steps, each of which is in one state leading to k time steps. After k time steps, a large reward, R_l, will be delivered.

short-term drug exposure. In our model, the term $D(s_t)$ is compensatable. In spite of this, estimated value of the drug does not stabilize after a few learning steps (limited use); only after prolonged use does it reach a high level. This is because in our model, the growth of the estimated value of the drug stops; that is, $\delta_t < 0$, when the basal reward level meets a level at least equal to $D(s_t)$. This fact, coupled with considering a slow increase in the basal reward level, explains why the drug's estimated value does not saturate after limited use.

3.3 Impulsive Choice. Impulsivity is a type of human behavior characterized by "actions that are poorly conceived, prematurely expressed, unduly risky, or inappropriate to the situation and that often result in undesirable outcomes" (Daruna & Barnes, 1993). It is a multifaceted construct, though two facets of it seem prominent in drug addiction: impulsive choice and impaired inhibition. Impulsive choice is exemplified by choosing a small immediate reward in preference to a larger, but delayed, reinforcer. Impaired inhibition refers to the inability to inhibit inappropriate or maladaptive behaviors. The direction of causation between drug addiction and increased impulsivity is not clear. However, the hypothesis that an abusive drug increases impulsive choice is supported by several studies (Paine, Dringenberg, & Olmstead, 2003). Moreover, chronic cocaine administration produces an increase in impulsive choice (Logue et al., 1992; Paine et al., 2003; Simon, Mendez, & Setlow, 2007). Here we show that the abnormal elevation of the basal reward level due to long-term cocaine consumption leads to the emergence of impulsive behavior from the model.

Impulsive choice is typically measured using delayed discounting tasks (DDT). In the delayed discounting paradigm, a subject is asked to choose between two options, one of which is associated with a small reward delivered immediately and the other with a larger reward delivered after a delay. Figure 6 shows the task configuration. A decision maker has two choices: C_1 and C_2. Selection of C_1 is followed by a small reward, R_s, after one time

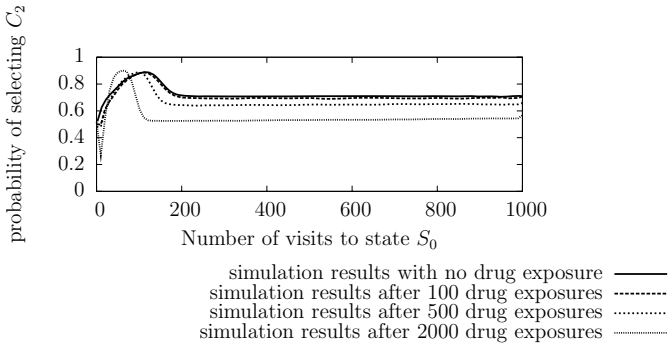

Figure 7: Probability of selecting the delayed large reward in instances of the model with different durations of prior drug exposure. With an increase in the duration of prior drug exposure, the models select large delayed reward with lower probability.

step. In choosing C_2, the decision maker waits k time steps, each of which is in one state, leading to k time steps. After k time steps have elapsed, a large reward, R_l, will be delivered.

The behavior of the model in the task is significantly under the influence of the basal reward level, ρ_t. In fact ρ_t determines the cost of waiting. High values of ρ_t in a task indicate that waiting is costly and thus guide the decision maker to the choice with a relatively faster reward delivery. In contrasty, low values indicate that losing time before reward delivery is not costly and it is worth waiting for a large reward. In a healthy decision-making system, the value of ρ_t changes during the learning process according to stimuli and actions. As result, it guides the action selection policy to the actions that maximize expected reward per step.

Now we assume that after chronic cocaine consumption, the basal reward level has an abnormally high value. Based on the discussion, it is reasonable that the model behavior shifts abnormally in favor of immediate rewards because the cost of waiting is relatively high and the decision maker prefers to choose the option that leads to immediate reward. The behavior of the model in DDT after different durations of prior drug exposure is shown in Figure 7. As the figure shows, along with an increase in the duration of prior drug exposure, the model chooses the large reward, C_2, with lower probability. This behavior of the model indicates that impulsivity increases with drug abuse increase. As the deviation declines after repeated natural reward use, the behavior of the model converges to that of a healthy model.

3.4 Blocking Effect. Panlilio et al. (2007) investigated whether the blocking effect occurs with cocaine. In their experiment, rats were divided into two groups: a blocking group and a nonblocking group. Initially, in the

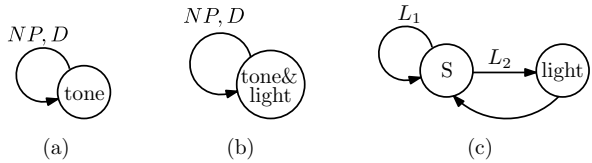

Figure 8: Simulation of the blocking effect. NP: nose poke, D: drug reward.

blocking group, cocaine SA was paired with a tone. In the nonblocking group, cocaine SA was paired with no stimulus. Next, in both groups, cocaine SA was paired with both the tone and a light. Finally, in the testing phase, the reinforcing effect of the light stimulus was measured in the two groups. The results showed that the light stimulus was associated with the drug reward in the nonblocking group of rats but not in the blocking group, and thus the tone blocked conditioning to the light.

In order to simulate blocking, a procedure similar to the one in Panlilio et al. (2007) experiment is used (see appendix B for details). Two instances of the model are simulated separately, each instance corresponding to one group of rats. The simulation has four steps:

1. Training. Both instances are trained with cocaine reward in the procedure shown in Figure 1.
2. Training. The blocked instance learns cocaine reward in a state in which the tone stimulus is presented (see Figure 8a). The nonblocked instance continues to learn the cocaine reward in step 1.
3. Training. Both instances learn cocaine reward in a state in which tone and light stimuli are presented (see Figure 8b).
4. Test. The two instances were simulated in the condition where two levers are available. Pressing the first one (L_1) has no effect, while pressing the second one (L_2) leads to a state in which the light stimulus is presented (see Figure 8c).

In the test step, if the blocked instance selects the second lever with less probability than the nonblocked instance, then the light stimulus is blocked by the tone stimulus because it reveals that the value of drug intake is not associated with the light stimulus. Figure 9 illustrates the behavior of each instance in the testing phase. As the figure shows, the blocked instance has no tendency to the second lever. This means the tone stimulus has blocked the light stimulus. Hence, the behavior of the model is consistent with the experiment about the occurrence of blocking with cocaine.

4 Prediction

Perhaps the least intuitive part of equation 2.4 is that the value of the state that the drug consumption leads to, $V(s_{t+1})$, is ignored in the cases where

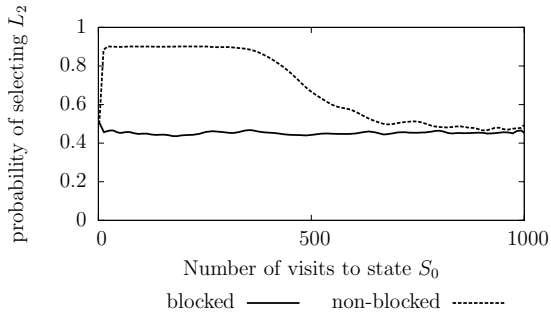

Figure 9: Probability of selecting L_2 in blocked and nonblocked instances of the model. In the blocked instance, the model shows no tendency to press L_2, which is associated with the light stimulus. The nonblocked instance selects L_2 with probability equal to L_1 and thus the value of the drug is not associated with the light stimulus.

the first operand of the maximization operator becomes bigger than the second one. Fortunately, this definition of the error signal can be tested experimentally using drug SA. Due to the mentioned property, the behavior of the model is sensitive to the temporal order of drug delivery and punishment. The model shows different behaviors between a task in which a punishment is followed by the drug reward and another one in which the drug reward is followed by the punishment. These two conditions can be described as follows:

1. The model has two choices: pressing the lever and an action that we call "other," meaning actions other than pressing the lever. Pressing the lever leads to drug delivery, and then, contingent on pressing the lever, the model receives punishment. Taking "other" action leads to a reward with a value near zero and it is not followed by the punishment (see Figure 10a). The value of pressing the lever is updated using the error signal in equation 2.4. Under the assumption that the punishment is very aversive and thus $V(s_{t+1})$ has a large negative value, we have

$$\max(r_t + V(s_{t+1}) - Q(s_t, a_t) + D(s_t), D(s_t)) = D(s_t), \tag{4.1}$$

and thus the error signal for updating value of pressing the lever will be

$$\delta_t = D(s_t) - \bar{R}_t. \tag{4.2}$$

The value of "other" action is updated by the error signal computed in equation 2.3 equal to $R_0 - \bar{R}_t$. R_0 is a reward with a value near zero, and hence $D(s_t) > E[R_0]$. As pressing the lever is updated by a greater value ($D(s_t) - \bar{R}_t$) in comparison to "other" action ($R_0 - \bar{R}_t$), pressing the level will gain a higher value in relation to "other" action.

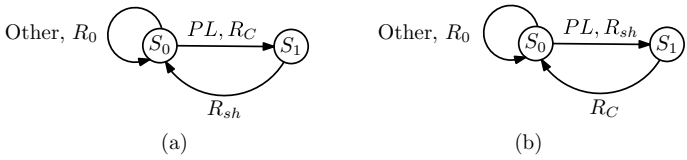

Figure 10: Two scenarios of drug SA for investigation of the effect of different temporal ordering of drug reward and punishment on the behavior of the model. (a) Pressing the lever (PL) leads to the drug delivery (R_C), and then, contingent on PL, the model receives shock punishment (R_{sh}). R_0: reward with zero mean. Other: actions other than PL. (b) PL leads to the punishment and the drug delivery afterward.

Therefore, the model chooses pressing the lever even after limited drug consumption and regardless of how much the punishment is aversive. This model behavior is related to neither the high estimated value of the drug nor the myopia for future punishment. It is because of that that the output of the maximization operator, a source of the error signal, is always greater than $D(s_t)$ regardless of the value of the next state.

2. Pressing the lever leads to the punishment and drug delivery afterward (see Figure 10b). The value of pressing the lever is updated by equation 2.3, and the scenario is similar to the procedure that we use for simulating compulsive drug seeking. The model behavior is sensitive to the magnitude of the punishment and the stage of the addiction. If the punishment is aversive enough or the model is not sufficiently trained by the drug reward, it will not choose pressing the lever.

Although we did not explicitly investigate this, it seems that the prediction in the first situation is questionable. If the prediction is wrong, this may stem from the strictness of the maximization operator, and thus its substitution with a softer averaging operator seems to solve the problem. Another important possibility is that the aversiveness of the states after drug consumption influences the value of drug consumption state through a pathway other than the DA one we model here. This possibility is consistent with the models that involve other pathways in the learning mechanism (Daw, Kakade, & Dayan, 2002). Involvement of other pathways causes the maximization operator not to cancel out the value of the states that drug consumption leads to, and the value of the states can influence the value of the drug consumption state through a different mechanism.

5 Discussion

We considered three assumptions on cocaine addiction: (1) the phasic activity of DA neurons codes prediction error, (2) cocaine causes an artificial

buildup of DA, and (3) prolonged exposure to cocaine causes abnormal elevation in the basal reward level. Based on these assumptions, we propose a model for cocaine addiction. We use the average reward TDRL model, which has the property that neural substrates of its internal signals and variables are known. The model predicts a decrease in reward system sensitivity and increased impulsivity after long-term drug exposure. Compulsive drug seeking is also explained by the high value of drug intake. Nevertheless, the model does not imply unbounded value for drug taking even after long-term exposure to the drug. It also predicts the existence of the blocking effect for the drug reward.

Previously, based on the neuronal network dynamical approach, Gutkin, Dehaene, and Changeux (2006) modeled nicotine addiction by highlighting the effect of nicotine on different timescales. In that model, a slow opponent process plays a critical role in drug addiction. It is assumed that the DA signal governs the gating of memory. On the other hand, long-term exposure to nicotine causes the DA signal to fall below a certain threshold needed for efficient learning. The model explains decreased harm avoidance based on an impaired learning mechanism. After long-term drug abuse, the model is unable to learn that drug seeking and use is followed by harmful consequences. Among the advantages of the model is that it does not imply unbounded value for drug intake. In comparison with our model, it is more concrete and explains the process at a neuronal level.

The model of Gutkin et al. (2006) predicts that when an animal is in the extinction phase after long-term exposure to a drug, the learning should be completely impaired. Therefore, the model cannot account for the unlearning of seeking behavior and reinstatement (relapse to drug use after extinction) after prolonged drug exposure as well. Our model addresses the unlearning of seeking behavior, but it predicts that states leading to drug use will lose their value when the drug is removed. Therefore, reinstatement cannot be described by our model either. This problem can be solved by facilitating the model with a state expansion mechanism (Redish, Jensen, Johnson, & Kurth-Nelson, 2007).

Because there is no explicit representation of estimated values in the Gutkin et al. (2006) model, the value of nondrug reinforcers before and after drug consumption cannot be compared structurally. Considering the behavior of the model, it learns a nondrug reinforcer more slowly after long-term drug consumption. But after being learned, the behavior of the model does not differ from what it was before the chronic drug abuse. This property of the model differs from our's, which does not imply slower learning and predicts a decrease in the value of environmental motivational stimuli and increased impulsivity after extended drug exposure.

Since drug dose does not correspond to any variable in our model, the model cannot be validated against the experiments that report a relationship between drug dose and response rates in different drug SA schedules. Moreover, it cannot be validated against simple schedules (e.g., fixed ratio

schedule with no time-out period in which responding for the drug is prevented). This is because in the currently available framework proposed by Niv et al. (2007), which considers response rate in RL, it is assumed that animals have a constant motivation level for the reinforcer during an experimental session. This assumption is not held in the case of a simple schedule, where each injection is followed by short-term satiety and a decrease in motivation for the drug. This limitation of our model makes it hard to compare its descriptive ability with pharmacological models of drug addiction (Ahmed & Koob, 2005; Norman & Tsibulsky, 2006), which are developed under the presence of the satiety effect. However, analyzing the behavior of the model in second-order schedules, a progressive schedule, and fixed ratio schedules that enforce a time-out period between successive injections is possible. Validation of the model against this experiment, especially investigation of the pattern of responses before and after chronic drug abuse, is an important step toward a more elaborate model of drug addiction.

In the structure of the model, other regions of brain, such as the amygdala, prefrontal cortex (PFC), and ventral pallidum, which are known to be involved in addictive behavior (Kalivas & O'Brien, 2007), are absent. Furthermore, transition of control from goal-directed declarative circuitry to habit circuitry (Everitt & Robbins, 2005) is not modeled. It seems that PFC can be added to the model using the multiprocess frameworks of decision making (Doya, 1999; Rangel et al., 2008; Redish, Jensen, & Johnson, 2008) as in the computational framework proposed by Daw, Niv, and Dayan (2005). Investigation of how control of behavior switches between the dorsolateral-striatal system and PFC in various stages of addiction will be an important step toward a more plausible model of drug addiction.

Experiments show that a subpopulation of rats is resistant to punishment after extended drug consumption (Deroche-Gamonet et al., 2004; Pelloux, Everitt, & Dickinson, 2007). Such individual difference and susceptibility to compulsive drug seeking is not reflected in our model. The behavior of the model is governed by some free parameters, which are partially dependent on the biological properties of the organism. For example, the estimated value of a drug after long-term drug abuse is dependent on the values of $D(s_t)$ and r_t. This means that for sufficiently small values of them, compulsive drug seeking will not emerge from the model due to a low estimated value of the drug. Some kinds of individual differences can be modeled and reduced to such factors.

Appendix A: Simulation Details

We define the reward of a natural reinforcer as a normally distributed variable as

$$R_N \sim N(\mu_N, \sigma_N^2). \tag{A.1}$$

Table 1: Simulation Parameters' Values.

Parameter	Value
σ	0.005
$D(s_t)$	15
α	0.2
μ_N	5
σ_N	0.02
μ_{fr}	-2
σ_{fr}	0.02
μ_{sh}	-200
σ_{sh}	0.02
μ_c	2
σ_c	0.02
C_u	6
λ	0.0003
N	2
μ_s	1
σ_s	0.02
μ_l	15
σ_l	0.02
ϵ	0.1
k	7ts

The punishment of the shock is molded by a normally distributed random variable with a large negative mean value ($\mu_{sh} \ll 0$):

$$R_{sh} \sim N\big(\mu_{sh}, \sigma_{sh}^2\big), \tag{A.2}$$

and the effect of freezing on the reduction of the shock punishment, with a normally distributed value with a mean less than μ_{sh}, ($\mid \mu_{fr} \mid \ll \mid \mu_{sh} \mid$),

$$R_{fr} \sim N\big(\mu_{fr}, \sigma_{fr}^2\big), \tag{A.3}$$

the cocaine reward with

$$R_c \sim N\big(\mu_c, \sigma_c^2\big), \tag{A.4}$$

and in a similar way, R_s, R_l with normal random variables:

$$\begin{aligned} R_s &\sim N\big(\mu_s, \sigma_s^2\big) \\ R_l &\sim N\big(\mu_l, \sigma_l^2\big). \end{aligned} \tag{A.5}$$

Model parameter values are presented in Table 1.

Actions are selected using an ϵ-greedy action selection policy in which the action with the highest estimated value is selected with probability $1-\epsilon$ (nonexploratory action), and with probability ϵ, an action is selected at random, uniformly (exploratory action). Average reward is computed over nonexploratory actions.

The model is simulated 1000 times, and the probability of selecting each action is calculated by the fraction of times the model has selected the action. The results are smoothed using Bezier curves. In the case of impulsive choice, for simplicity of implementation we use a semi-Markov version of average reward RL algorithm (Das, Gosavi, Mahadevan, & Marchalleck, 1999). The delay in all states is assumed to be 1 time step (ts).

Appendix B: Blocking Effect

In order to simulate the blocking effect, it is necessary to represent multiple stimuli. For this purpose, similar to Montague et al. (1996), we use a linear function approximator for the state-action values (Melo & Ribeiro, 2007):

$$Q(s_t, a_t) = \xi(s_t).w(a_t) \tag{B.1}$$

Here, we use a binary representation for $\xi(s_t)$. That is, ith element of vector $\xi(s_t)$ is one if and only if ith stimulus be presented at time t, and zero otherwise. The weights are updated using the TD learning rule,

$$w(a_t) \leftarrow w(a_t) + \alpha\delta_t\xi(s_t), \tag{B.2}$$

and the error signal is computed using equation 2.4 for the cocaine reward and equation 2.3 for the natural reward with state-action values computed from equation B.1.

Simulation of the blocking effect has several steps; thus, an issue arises in the blocked instance of the model: how a value associated with a stimulus in a prior step should be carried into a new step. For example, the value associated with the light stimulus in step 2 of the simulation should be carried into the light value in step 3 at the beginning of the step. To address this issue, we simply initialize the model with previously learned values for state-action pairs and average reward (because values of state-action pairs are relative to average reward) at the beginning of step 3.

Acknowledgments

We are very grateful to Serge H. Ahmed, Yael Niv, Nathaniel Daw, and David Redish for helpful discussions and Laleh Ghadakpour, Morteza Dehghani, Azin Moallem, Mohammad Mahdi Ajallooeian, Habib Karbasian, and Sadra Sade for comments on the manuscript. Also, we acknowledge

useful comments from the anonymous reviewers who helped to greatly improve the letter.

References

Ahmed, S. H. (2004). Neuroscience: Addiction as compulsive reward prediction. *Science, 306*(5703), 19011902.

Ahmed, S. H., & Koob, G. F. (1998). Transition from moderate to excessive drug intake: Change in hedonic set point. *Science, 282*(5387), 298–300.

Ahmed, S. H., & Koob, G. F. (1999). Long-lasting increase in the set point for cocaine self-administration after escalation in rats. *Psychopharmacology, 146*(3), 303–312.

Ahmed, S. H., & Koob, G. F. (2005). Transition to drug addiction: A negative reinforcement model based on an allostatic decrease in reward function. *Psychopharmacology, 180*, 473–490.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Aragona, B. J., Cleaveland, N. A., Stuber, G. D., Day, J. J., Carelli, R. M., & Wightman, R. M. (2008). Preferential enhancement of dopamine transmission within the nucleus accumbens shell by cocaine is attributable to a direct increase in phasic dopamine release events. *Journal of Neuroscience, 28*(35), 8821–8831.

Chen, R., Tilley, M. R., Wei, H., Zhou, F., Zhou, F., Ching, S., et al. (2006). Abolished cocaine reward in mice with a cocaine-insensitive dopamine transporter. *Proceedings of the National Academy of Sciences, 103*(24), 9333–9338.

Daruna, J. H., & Barnes, P. A. (1993). A neurodevelopmental view of impulsivity. In W. G. McCown, J. L. Johnson, & M. B. Shure (Eds.), *The impulsive client: Theory, research, and treatment* (p. 23). Washington, DC: American Psychological Association.

Das, T. K., Gosavi, A., Mahadevan, S., & Marchalleck, N. (1999). Solving semi-Markov decision problems using average reward reinforcement learning. *Management Science, 45*, 560–574.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications*. Unpublished doctoral dissertation, Carnegie Mellon University.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16*(2), 199–204.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks, 15*(4–6), 603–616.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*, 876–879.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation, 14*(11), 2567–2583.

Deroche-Gamonet, V., Belin, D., & Piazza, P. V. (2004). Evidence for addiction-like behavior in the rat. *Science, 305*, 1014–1017.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks, 12*, 961–974.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience, 8*(11), 1481–1489.

Garavan, H., Pankiewicz, J., Bloom, A., Cho, J. K., Sperry, L., Ross, T. J., et al. (2000). Cue-induced cocaine craving: Neuroanatomical specificity for drug users and drug stimuli. *American Journal of Psychiatry, 157*(11), 1789–1798.

Goldstein, R. Z., Alia-Klein, N., Tomasi, D., Zhang, L., Cottone, L. A., Maloney, T., et al. (2007). Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *American Journal of Psychiatry, 164*(1), 43–51.

Grace, A. A. (1995). The tonic/phasic model of dopamine system regulation: Its relevance for understanding how stimulant abuse can alter basal ganglia function. *Drug and Alcohol Dependence, 37*(2), 111–129.

Grace, A. A. (2000). The tonic/phasic model of dopamine system regulation and its implications for understanding alcohol and psychostimulant craving. *Addiction, 95* (Suppl. 2), S119–128.

Gutkin, B. S., Dehaene, S., & Changeux, J. (2006). A neurocomputational hypothesis for nicotine addiction. *Proceedings of the National Academy of Sciences of the United States of America, 103*(4), 1106–1111.

Kalivas, P. W., & O'Brien, C. (2007). Drug addiction as a pathology of staged neuroplasticity. *Neuropsychopharmacology, 33*(1), 166–180.

Kalivas, P. W., & Volkow, N. D. (2005). The neural basis of addiction: A pathology of motivation and choice. *American Journal of Psychiatry, 162*(8), 1403–1413.

Kamin, L. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts.

Kauer, J. A., & Malenka, R. C. (2007). Synaptic plasticity and addiction. *Nature Reviews Neuroscience, 8*(11), 844–858.

Koob, G. F., & Moal, M. L. (2005). Plasticity of reward neurocircuitry and the "dark side" of drug addiction. *Nature Neuroscience, 8*(11), 1442–1444.

Logue, A., Tobin, H., Chelonis, J., Wang, R., Geary, N., & Schachter, S. (1992). Cocaine decreases self-control in rats: A preliminary report. *Psychopharmacology, 109*(1), 245–247.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning, 22*(1), 159–195.

Mantsch, J. R., Ho, A., Schlussman, S. D., & Kreek, M. J. (2001). Predictable individual differences in the initiation of cocaine self-administration by rats under extended-access conditions are dose-dependent. *Psychopharmacology, 157*(1), 31–39.

Mateo, Y., Lack, C. M., Morgan, D., Roberts, D. C. S., & Jones, S. R. (2005). Reduced dopamine terminal function and insensitivity to cocaine following cocaine binge self-administration and deprivation. *Neuropsychopharmacology, 30*(8), 1455–1463.

Melo, F. S., & Ribeiro, M. I. (2007). Q-learning with linear function approximation. *Proceedings of the 20th Annual Conference on Learning Theory* (pp. 308–322). Berlin: Springer.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16*, 1936–1947.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9*(8), 1057–1063.

Nader, M. A., Morgan, D., Gage, H. D., Nader, S. H., Calhoun, T. L., Buchheimer, N., et al. (2006). PET imaging of dopamine D2 receptors during chronic cocaine self-administration in monkeys. *Nature Neuroscience, 9*(8), 1050–1056.

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

Norman, A. B., & Tsibulsky, V. L. (2006). The compulsion zone: A pharmacological theory of acquired cocaine self-administration. *Brain Research, 1116*(1), 143–152.

Paine, T. A., Dringenberg, H. C., & Olmstead, M. C. (2003). Effects of chronic cocaine on impulsivity: Relation to cortical serotonin mechanisms. *Behavioural Brain Research, 147*(1–2), 135–147.

Panlilio, L. V., Thorndike, E. B., & Schindler, C. W. (2007). Blocking of conditioning to a cocaine-paired stimulus: Testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. *Pharmacology, Biochemistry, and Behavior, 86*(4), 774–777.

Paterson, N. E., & Markou, A. (2003). Increased motivation for self-administered cocaine after escalated cocaine intake. *Neuroreport, 14*(17), 2229–2232.

Pelloux, Y., Everitt, B. J., & Dickinson, A. (2007). Compulsive drug seeking by rats under punishment: Effects of drug taking history. *Psychopharmacology, 194*, 127–137.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience, 9*(7), 545–556.

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science, 306*, 1944–1947.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences, 31*(4), 415–487.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review, 114*(3), 784–805.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10*(12), 1615–1624.

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

Simon, N. W., Mendez, I. A., & Setlow, B. (2007). Cocaine exposure causes long-term increases in impulsive choice. *Behavioral Neuroscience, 121*(3), 543–549.

Stuber, G. D., Wightman, R. M., & Carelli, R. M. (2005). Extinction of cocaine self-administration reveals functionally and temporally distinct dopaminergic signals in the nucleus accumbens. *Neuron, 46*(4), 661–669.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tilley, M. R., O'Neill, B., Han, D. D., & Gu, H. H. (2009). Cocaine does not produce reward in absence of dopamine transporter inhibition. *Neuroreport, 20*(1), 9–12.

Vanderschuren, L. J. M. J., & Everitt, B. J. (2004). Drug seeking becomes compulsive after prolonged cocaine self-administration. *Science, 305*, 1017–1019.

Volkow, N. D., Fowler, J. S., Wang, G., & Swanson, J. M. (2004). Dopamine in drug abuse and addiction: Results from imaging studies and treatment implications. *Molecular Psychiatry, 9*(6), 557–569.

Volkow, N. D., Fowler, J. S., Wang, G., Swanson, J. M., & Telang, F. (2007). Dopamine in drug abuse and addiction: Results of imaging studies and treatment implications. *Archives of Neurology, 64*(11), 1575–1579.

Volkow, N. D., Wang, G. J., Fowler, J. S., Logan, J., Gatley, S. J., Hitzemann, R., et al. (1997). Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature, 386*(6627), 830–833.

Watkins, C. (1989). *Learning from Delayed Rewards*. Unpublished doctoral dissertation, King's College, Cambridge, U.K.

Received October 13, 2008; accepted March 5, 2009.