Chapter 5
Dual-System Learning Models and Drugs of Abuse

Dylan A. Simon and Nathaniel D. Daw

Abstract Dual-system theories in psychology and neuroscience propose that a deliberative or goal-directed decision system is accompanied by a more automatic or habitual path to action. In computational terms, the latter is prominently associated with model-free reinforcement learning algorithms such as temporal-difference learning, and the former with model-based approaches. Due in part to the close association between drugs of abuse and dopamine, and also between dopamine, temporal-difference learning, and habitual behavior, addictive drugs are often thought to specifically target the habitual system.

However, although many drug-taking behaviors are well explained under such a theory, evidence suggests that drug-seeking behaviors must leverage a goal-directed controller as well. Indeed, one exhaustive theoretical account proposed that drugs may have numerous, distinct impacts on both systems as well as on other processes.

Here, we seek a more parsimonious account of these phenomena by asking whether the apparent profligacy of drugs' effects might be explained by a single mechanism of action. In particular, we propose that the pattern of effects observed under drug abuse may reveal interactions between the two controllers, which have typically been modeled as separate and parallel. We sketch several different candidate characterizations and architectures by which model-free effects may impinge on a model-based system, including sharing of cached values through truncated tree search and bias of transition selection for prioritized value sweeping.

5.1 Introduction

Dual-system theories of decision making—involving, for instance, a deliberative "goal-directed" controller and a more automatized or "habitual" one—are ubiq-

D.A. Simon
Department of Psychology, New York University, New York, NY, USA

N.D. Daw (✉)
Center for Neural Science and Department of Psychology, New York University, New York, NY, USA
e-mail: nathaniel.daw@nyu.edu

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_5,

uitous across the behavioral sciences (Blodgett and McCutchan 1947; Dickinson 1985; Verplanken et al. 1998; Kahneman and Frederick 2002; Loewenstein and O'Donoghue 2004; Daw et al. 2005; Wood and Neal 2007). Many theories of drug abuse draw on this sort of framework, proposing that the compulsive nature of abuse reflects a transition of behavioral control from the voluntary system to the habitual one (Tiffany 1990; Ainslie 2001; Everitt et al. 2001; Vanderschuren and Everitt 2004; Everitt and Robbins 2005; Bechara 2005). Such a characterization may explain many drug-taking behaviors that become stereotyped and automatic, and dovetails naturally with models of the function of the neuromodulator dopamine (a ubiquitous target of drugs of abuse) suggesting a specific role for this neuromodulator in reinforcing habits (Di Chiara 1999; Redish 2004). However, the view of abusive behaviors as excessively automatized stimulus-response habits cannot easily explain many sorts of drug-*seeking* behaviors, which can involve novel and often increasingly inventive goal-directed acquisition strategies (Tiffany 1990). Such theories also do not speak to more cognitive phenomena such as craving.

Drug abuse is a dysfunction of decision making, acquired through learning. In this domain, theories are often formalized in terms of reinforcement learning (RL) algorithms from artificial intelligence (Sutton and Barto 1998). By providing a quantitative characterization of decision problems, RL theories have enjoyed success in behavioral neuroscience as methods for direct analysis and interpretation of trial-by-trial decision data, both behavioral and neural (Schultz et al. 1997; Daw and Doya 2006). Importantly, these theories also offer a putative computational counterpart to the goal-directed vs. habitual distinction, which may be useful for characterizing either system's role in drug abuse. In these terms, the more automatic, habitual behaviors are typically associated with so-called *model-free* RL, notably temporal-difference (TD) methods such as the actor/critic, in which successful actions are reinforced so that they may be repeated in the future. However, it has more recently been proposed that goal-directed behaviors can be captured with a categorically distinct type of RL known as *model-based*, in which actions may be planned based on a learned associative model of the environment (Doya 1999; Daw et al. 2005; Tanaka et al. 2006; Hampton et al. 2006; Pan et al. 2007; Redish and Johnson 2007; Rangel et al. 2008; Gläscher et al. 2010).

Such theories hypothesize that goal-directed and habitual behaviors arise from largely separate and parallel RL systems in the brain: model-based and model-free. Model-free RL forms the basis for a well-known account of dopamine neurons in the midbrain, as well as BOLD activity in dopamine targets in the basal ganglia (Houk et al. 1994; Schultz et al. 1997; Berns et al. 2001; O'Doherty et al. 2003; McClure et al. 2003). Since it is well established that drugs of abuse affect the function of these systems, and that other problem behaviors such as compulsive gambling show evidence of related effects, it has been a natural and fruitful line of research to apply TD-like theories to drug abuse (Di Chiara 1999; Redish 2004). However, these more computational theories pose the same puzzle as their psychological counterparts: how to account for the role of more flexible, drug-seeking behaviors apparently associated with goal-directed (in this case,

model-based) control. Here, we consider how these behaviors might be understood in terms of the less well characterized model-based system. We follow Redish et al. (2008) in this endeavor, but focus more on what drug abuse phenomena suggest about potential variants or elaborations of the standard model-based account. In particular, we relate these issues to a range of other data suggesting that the two hypothesized RL systems are not as separate as they have been envisioned, but may instead interact in some respects. We consider how different sorts of interaction might be captured in modified forms of these theories in order to extend the computational account of drug abuse.

5.2 Background: Reinforcement Learning and Behavior

As a framework for formalizing theories of drug abuse, this section lays out the basics of RL, the study of learning optimal decisions through trial and error. For a more detailed description of this branch of computer science, see Sutton and Barto (1998) or, for its applications to psychology, Balleine et al. (2008).

5.2.1 The Markov Decision Process

Most decision problems in RL are based on Markov decision processes (MDPs), which formalize real-world problems as a sequence of steps, each of which involves a choice between actions affecting the resulting reward and the situation going forward. Formally, an MDP is a set of *states*, $\mathcal{S}$, and *actions*, $\mathcal{A}$, which occur in some sequence, s_t and a_t over timesteps t, such that s_{t+1} depends stochastically on s_t and a_t, but on no other information. This dependence is described by a *transition function* specifying the probability distribution over possible next states given the current state and chosen action:

$$T(s, a, s') = \mathrm{P}\left[s_{t+1} = s' \middle| s_t = s, a_t = a\right]$$

Rewards are similarly described by a stochastic *reward function* mapping each state to the quantity of reward received in that state: $R : \mathcal{S} \rightarrow \mathbb{R}$, such that $r_t = R(s_t)$. The transition and reward functions thus fully describe the process.

5.2.2 Values and Policies

The goal of an agent in an MDP is to select actions so as to maximize its reward, and more specifically, to *learn* to do so by trial and error, using only information about the underlying process observed during behavior (i.e., samples from the transition and reward functions). Specifically, at a state, s, an agent aims to pick the action,

a, that will maximize the cumulative, temporally discounted rewards that will be received in the future, in expectation over future states and actions:

$$Q(s,a) = \mathrm{E}\left[\sum_{i=1} \gamma^{i-1} r_{t+i} | s_t = s, a_t = a\right]$$

where $\gamma < 1$ is an exponential time discounting factor. This quantity is known as the state-action value function, and many approaches to RL involve estimating it, either directly or indirectly, so as to choose the action maximizing it at each state.

A key aspect of MDPs (indeed, what makes them difficult), is their sequential nature. An agent's future value prospects depend not only on the current state and action, but on future choices as well. Formally, consider a *policy* by which an agent selects actions, that is, a (possibly stochastic) function describing the action to take in each state: $\pi : \mathcal{S} \rightarrow \mathcal{A}$. From this, we can define the expected value of taking action a in state s, and then following policy π thereafter:

$$Q^{\pi}(s,a) = \sum_{s'} T\left(s,a,s'\right)\left[\mathrm{E}\left[R\left(s'\right)\right] + \gamma \sum_{s''} T\left(s',\pi\left(s'\right),s''\right)\left[\mathrm{E}\left[R\left(s''\right)\right] + \cdots\right]\right] \tag{5.1}$$

This value depends on the sequence of future expected rewards that will be obtained, averaged over all possible future trajectories of states, $s, s', s'', \ldots$, according to the policy and transition function. One way of framing the goal, then, is to determine the optimal policy, known as π^*, that will maximize $Q^{\pi^*}(s_t, \pi^*(s_t))$ at each step.

A key insight relevant to solving this problem is that the state-action value may be written recursively:

$$Q^{\pi}(s,a) = \sum_{s'} T\left(s,a,s'\right)\left[R\left(s'\right) + \gamma Q^{\pi}\left(s',\pi\left(s'\right)\right)\right]$$

Since the optimal policy must maximize Q at each step, the optimal value satisfies:

$$Q^{*}(s,a) = \sum_{s'} T\left(s,a,s'\right)\left[R(s') + \gamma \max_{a'} Q^{*}\left(s',a'\right)\right] \tag{5.2}$$

This is known as the Bellman equation, which provides a recursive relationship between all the action values in the MDP.

The optimal policy can be extracted directly from the optimal value function, if it is known. That is, an agent can achieve maximal expected reward by simply choosing the maximally valued action at each step: $\pi^*(s) = \mathrm{argmax}_a Q^*(s,a)$. Accordingly, we next consider two different approaches to learning to choose actions, which each work via learning to estimate $Q^*(s,a)$.

5.2.3 Algorithms for RL

5.2.3.1 Model-Free RL

The recursive nature of the state-action value motivates one approach to RL, often exemplified by temporal-difference learning (Sutton 1988). Here, an agent attempts directly to estimate the optimal value function Q^*. (A closely related variant, the actor/critic algorithm, estimates the policy π^* itself using similar methods.)

The recursion in Eq. (5.2) shows how such an estimate may be updated, by changing it so as to reduce the observed deviation between the left and right hand sides of the equation, known as the prediction error. Specifically, consider any step at which an action, a, is taken in state s, and a new state, s', and reward, r, are observed. From Eq. (5.2), it can be seen that the quantity $r + \gamma \max_{a'} Q(s', a')$ is a *sample* of the value of the preceding state and action, $Q(s, a)$, where the state s' samples the transition distribution $T(s, a, s')$, and the agent's own estimate of the new state's value, $Q(s', a')$, stands in for the true Q^*. We can then update the estimated value toward the observed value, with learning rate α:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\underbrace{\overbrace{\left(R\left(s'\right) + \gamma \max_{a'} Q\left(s', a'\right)\right)}^{Q \text{ sample}} - Q(s,a)\Big)}_{\text{prediction error, } \delta}$$

Algorithms of this sort are known as model-free approaches because they do not directly represent or make use of the underlying MDP transition or reward functions, but instead learn the relevant summary quantity directly: the state-action value function.

5.2.3.2 Model-Based RL

A second approach to RL is model-based learning. Here, representations of the transition and reward functions are themselves learned, which function as a model of the MDP, thus giving rise to the name. This is quite straightforward; for instance, the transition function may be estimated simply by counting state-action-state transitions. Given any estimate of these functions, the state-action value function may be computed directly, for example, through the iterative expansion of Eq. (5.1) to explicitly compute the expected rewards over different possible trajectories.

Such a recursive computation can be laborious, in contrast to and thus motivating model-free methods which involve minimal computation at choice time (e.g., simply comparing learned state-action values). The flip side of this trade-off is that computing these values on the basis of a full world model, rather than simply relying on a previously learned summary, offers more flexible possibilities for combining information learned at different times, and enables the agent to respond more dynamically under changing situations.

5.2.4 RL and Behavioral Neuroscience

These two frameworks for solving an MDP, one (model-free) computationally fast and reactive, and the other (model-based) involving more deliberative or proactive consideration of possibilities, are closely related to the psychological concepts of habits and goal-directed actions, respectively.

In psychology, these two sorts of instrumental behavior are envisioned as relying on different underlying representations (Balleine and Dickinson 1998). Goal-directed actions are supposed to be based on a representation of the action-outcome contingency (e.g., that pressing a lever produces a certain amount of cheese; or, in a spatial task, a 'cognitive map' of the maze), allowing deliberative choice by examining the consequences of different possible actions. Habits are instead assumed to be based on direct stimulus-response associations, which may be learned by a simple reinforcement rule (i.e., if a response in the presence of some stimulus is followed by reward, strengthen it, as proposed by Thorndike 1898) and embody a very simple, switchboard-like choice strategy.

However, since the stimulus-response association lacks any representation of the specific outcome (e.g., cheese) that originally reinforced it, a choice mechanism of this sort predicts odd inflexibilities and insensitivities to certain shifts in circumstances. For instance, it predicts that a rat who is trained to lever-press for food while hungry, but then fed to satiety, will continue to work for food given the opportunity, at least until given enough experience to unlearn or relearn the association. In contrast, since choosing a goal-directed action involves examining the action-outcome association, this approach can adjust behavioral preferences instantly to comply with new situations such as changes in outcome values. Another important capability of a goal-directed approach is the ability to plan novel actions to obtain new goals or react to new information. For instance, in a maze, an animal might use a cognitive map to plan a route not previously followed, such as a shortcut between two locations (Tolman 1948). Such flexibility is not possible using only stimulus-response associations (since such a route will not have previously been reinforced).

All this motivates standard experimental procedures, such as outcome devaluation, for distinguishing these two sorts of behaviors. The results of such tests (specifically, whether actions are or are not sensitive to devaluation under different circumstances) indicate that the brain uses both approaches (Dickinson and Balleine 2002).

The two sorts of RL algorithms directly mimic these psychological theories in key respects (Daw et al. 2005; Balleine et al. 2008). Like habits, model-free approaches support easy choices by relying on a summary representation of an immediately relevant decision variable: the value function or policy. For the same reason, these representations lack information about outcome identity and are insensitive to changes; they can be updated only following additional experience with the consequences of a state and action, and often through its repetition. Conversely, model-based algorithms formalize the idea of an associative or cognitive search in which possible outcomes are explicitly considered in relation to their likelihood of achieving some goal (i.e., reward). These forms of reasoning depend on representations of

outcomes and state transitions (analogous to a cognitive map or action-outcome association), and, like their psychological counterparts, can adjust rapidly to changes in the worth or availability of outcomes and can combine previously experienced sequences of actions in novel ways to reach goals.

RL approaches have also been associated with specific neural systems. Model-free algorithms in particular have been a valuable tool for explaining the function of the dopamine system, as the firing rates of midbrain dopamine neurons closely match the error signals predicted by these algorithms (Houk et al. 1994; Schultz et al. 1997). There is also evidence for representation of state-action values in other areas of the brain, including prefrontal cortex, striatum, and parietal regions (Delgado et al. 2000; Arkadir et al. 2004; Tanaka et al. 2004; Samejima et al. 2005; Plassmann et al. 2007; Tom et al. 2007; Kable and Glimcher 2007; Hare et al. 2008; Kim et al. 2009; Wunderlich et al. 2009; Chib et al. 2009).

Less is known about the neural substrate for model-based or goal-directed actions, though there are now a number of reports of potentially model-related activity throughout the brain (Hampton et al. 2006, 2008; Pan et al. 2007; Bromberg-Martin et al. 2010). In general, these actions are not envisioned to involve dopamine, since model-based approaches rely on quite different learning mechanisms with error signals that do not match the dopaminergic response (Gläscher et al. 2010) and because lesions of the dopaminergic system appear to spare goal-directed action while affecting habits (Faure et al. 2005). More generally, the use of the reward devaluation procedure together with numerous brain lesions has allowed the demonstration of an anatomical double-dissociation, wherein different areas of striatum (and associated parts of cortex and thalamus) support each learning strategy even under circumstances when the other would be observed in intact animals (Killcross and Coutureau 2003; Yin et al. 2004, 2005; Balleine et al. 2007). These findings have suggested that the brain implements both model-based and model-free approaches as parallel and, to some extent, independent systems.

5.2.5 RL and Drugs of Abuse

In this light it seems natural to interpret the strongly habitual behaviors associated with drug taking as an effect of drug abuse specifically on model-free valuations (Redish 2004; Redish et al. 2008; Schultz 2011). In particular, compulsive behaviors have been attributed to overly strong habitual responses (or state-action values), whereby learned responses persist despite contrary evidence of their value available to a contemplative, model-based system (Everitt and Robbins 2005). A candidate mechanism for such uncontrolled reinforcement is effects of drugs on the dopaminergic signal carrying the reward prediction error supposed to train model-free values or policies (Redish 2004; Panlilio et al. 2007; Redish et al. 2008). This interpretation is consistent with the fact that most if not all drugs of abuse share effects on dopamine as a common mechanism of their reinforcing action.

However, it has also been pointed out that such an account is necessarily incomplete, and in particular that drug abusers demonstrate highly elaborate and often novel drug-seeking behaviors (Tiffany 1990; Olmstead et al. 2001; Kalivas and Volkow 2005; Root et al. 2009). Just as with short cuts in mazes, such flexible planning cannot be explained by the model-free repetition of previously reinforced actions. Therefore, the remainder of this chapter considers algorithmic possibilities for ways a model-based system could be affected by drug abuse. These considerations have consequences for theories of appetitively motivated behavior more generally, since they strongly suggest some sort of integration or cooperation between the systems in commonly valuing drug outcomes.

5.3 Drugs and Model-Based RL

The problem facing us is that, under a standard theory (e.g., Daw et al. 2005), drugs of abuse affect valuations only in the model-free system, via effects on a dopaminergic prediction error. Valuations in a model-based system have been presumed to be entirely separate and independent, and in particular, to be unaffected by manipulations of dopamine. However, if the effects of drugs are isolated to a model-free system (and drugs are not, by comparison, disproportionately valued in a model-based system) then actions motivated by drugs should exclusively constitute simple repetitions of previously reinforced actions. Such a system has no mechanism for planning novel drug-seeking actions.

In this section, we consider a number of potential solutions to this issue, focusing on effects either via inflation of values per se, or biasing them via changes in the search process by which they are computed.

5.3.1 Drugs and Model-Based Reward

A typical application of model-free theories to drug abuse depends on drugs affecting the learned value or policy function, for example, by inflating the state-action values leading to drug rewards. Is there some simple analogy in a model-based system for such inflation? While model-based systems typically construct a value function on demand, rather than maintaining a representation of one, they do maintain a representation of rewards in some other form, often as an approximation to the state reward function. This reward function could theoretically be learned through prediction errors just as state-action values are, and similarly be inflated as an effect of drug abuse (Redish et al. 2008; Schultz 2011). In this case, an increased reward associated with the attainment state would flexibly elicit a wide range of goal-directed behaviors, as any actions likely to eventually reach that state would themselves have a higher computed action value, even along novel paths. However, this explanation raises a problematic question: what is the process by which drugs of abuse could inflate the reward function?

By analogy with the TD account, the natural answer would seem to be that the inflation happens in much the same way as model-free value inflation is supposed to occur: via effects on dopaminergic responses effectively exaggerating the prediction error used to learn these representations. However, as previously mentioned, the representations learned in a model-based system (notably, the reward function, R) require different sorts of prediction errors (Gläscher et al. 2010). On available evidence, the responses of dopaminergic neurons appear consistent with a prediction error appropriate for training future (discounted) value (Q), not immediate reward (R). In particular, the signature phenomenon whereby dopamine responses transfer with training to cues predicting upcoming reward is inconsistent with a prediction error for the one-step reward R: there are no immediate rewards and no errors in their predictions tied to this event (Schultz et al. 1997). Moreover, although reward values for the model-based system are likely represented in a dissociable location in the brain from model-free values, it is unlikely that this learning is driven by some atypical dopaminergic signal, since reports suggest at least anecdotally that dopamine neurons are consistent in this respect, regardless of where they project (Schultz 1998).

If dopamine controls these secondary incentives or motivational values and not representations of one-step rewards, then the latter are unlikely to be a mechanism by which drugs of abuse impact model-based valuations.

5.3.2 Drugs and Model-Based Value

In order to solve this problem, we return to the Bellman equation (5.2) which connects model-free and model-based approaches by defining the state-action value that they both compute in different ways. A key claim of the model-free approaches is that the brain maintains internal ("cached" or stored) estimates of the state-action values, which are updated in place by prediction error and are putatively inflated by drugs of abuse via their effects on this prediction error signaling. The model-based approach is assumed instead to compute the state-action values anew at decision time by evaluating the Bellman equation, deriving them from more elemental information (the reward and transition functions).

If indeed both systems operate in the brain and aim to compute equivalently defined state-action values, then the Bellman equation suggests an obvious possibility for their interaction: a model-based system could make use of the cached state-action values maintained by the model-free system. In particular, because of the recursive form of the Bellman equation, at any point in its iterative, tree-structured expansion, it is possible to substitute a cached (e.g., model-free) estimate of the right-hand value, Q, to terminate the expansion. One motivation for this "partial evaluation" is that the full reevaluation of the Bellman equation at each decision step is computationally laborious; moreover, repeating this computation each step may have diminishing returns if, for instance, the learned estimates of transition

and reward functions change little between each evaluation (Moore and Atkeson 1993).

If a model-based search immediately terminated with cached action values (i.e., on the first step) it would simply revert to a model-free system, while each additional step of evaluation using the model's transition and reward functions would provide a view of value which is model-based out to a horizon extended one step further into the future, at the cost of additional computation. Thus, if a model-based system engaged in such partial evaluation by terminating its search at states with model-free state-action values inflated by the theorized dopamine mechanisms (such as states associated with drug attainment), the model-based system would be similarly compromised, with this exaggeration carried back to other computed action values that may reach such a state. The combination of the two sorts of evaluation would allow the model-based system to plan novel action trajectories aimed at attaining states with high (potentially drug-inflated) value in the model-free system's estimates. In this sense, the model-free estimates can serve as secondary incentives for guiding the model-based system's preferences, an idea reminiscent of "incentive salience" accounts of drug motivation (Robinson and Berridge 2008).

The foregoing considerations suggest a new perspective on the joint contribution of model-based and model-free evaluations to behavior. Whereas previous work (Daw et al. 2005) envisioned that the brain must select between separate model-based and model-free values, the partial evaluation approach suggests that the key question is instead where to integrate the values: at each step, whether to further evaluate a decision branch or to truncate the trajectory using cached values. With this extension, the traditional story of a shift from goal-directed to automatic processing can make a broader range of behavioral predictions as a shift towards more limited searches under model-based evaluation (Nordquist et al. 2007).

Also, interacting architectures of this broad sort may help to explain numerous indications from the neuroscientific literature that model-free and model-based evaluation may be more interacting than separate. For instance, goal-directed learning appears to involve a subregion of striatum, dorsomedial, which is adjacent to the part apparently responsible for habits, and which also receives heavy dopaminergic innervation (Yin et al. 2005). Moreover, indications of model-based computations (such as devaluation sensitivity) have been observed throughout areas of the brain traditionally thought to be part of the model-free system including ventral striatum (Daw et al. 2011; Simon and Daw 2011; van der Meer et al. 2010), downstream ventral pallidum (Tindell et al. 2009), and even dopaminergic neurons (Bromberg-Martin et al. 2010).

In the drug context, this view also raises a new set of questions, surrounding how drugs might affect search termination. For instance, if drug-inflated estimates of state-action values serve as secondary incentives for model-based search, why would the model-based system terminate with them, rather than planning *past* the contaminated states?

5.3.3 Drugs and Model-Based Search

If a model-based search process were biased at search time to adopt exaggerated cached values rather than pursuing further evaluation, the resulting behavior would show strong preferences for actions (even novel ones) that tend to lead to such outcomes. The question is why such a bias would arise. That is, the concept of partial evaluation explains how inflated values in the model-free system could affect the model-based system, but may not adequately account for the particular fixations drugs of abuse engender, whereby goal-directed behaviors may operate to fulfill the craving to the exclusion of other goals.

To begin to address this question, we consider how search progress and search termination might be affected by drugs of abuse. A more general and flexible framework for reasoning about these issues is Sutton's Dyna architecture (Sutton 1990), which provides a framework by which model-based and model-free RL can coexist and dynamically trade-off their contributions to learning. This architecture has also been employed in theories of model-based learning in the brain (Johnson and Redish 2005). The Dyna-Q algorithm envisions that an agent will maintain a single set of cached state-action values, but that these can be updated by both model-based and model-free updates in any mixture. As with standard model-free learning, state-action values may be updated directly by prediction errors according to actual experience. A learned world model can also be used to produce simulated experience (i.e., state, action and reward trajectories sampled from the modeled transition and reward functions), which can train the cached state-action values in the same way as real experience. Full model-based value updates (i.e., averaging rather than sampling over possible successor states for an action using the Bellman equation) can also be applied in place.

As opposed to the traditional view of a tree-structured search, Dyna-Q has the freedom to apply these model-based updates in arbitrary orders. All these updates may be interleaved during behavior, at decision time, or off-line. Given sufficient updates, the values learned will approach the same model-based values a fully expanded search would. Because of the possibility of learning from simulated sample trajectories, the theory also exposes the connection between model-based valuation and simulation. Intuitively this idea comports well with ideas that search may be implemented by cognitive simulation (Buckner and Carroll 2007; Buckner 2010) as well as evidence for various sorts of on- and off-line replay or preplay over spatial trajectories in hippocampal place cells (Johnson and Redish 2005; Foster and Wilson 2006; Hasselmo 2008; Koene and Hasselmo 2008; Davidson et al. 2009; Lansink et al. 2009; Derdikman and Moser 2010; Carr et al. 2011; Dragoi and Tonegawa 2011).

The question of drug abuse now can be further refined to which trajectories are simulated, as well as where these trajectories are terminated. One principled approach to this question is the prioritized sweeping algorithm (Moore and Atkeson 1993). In its original form, it is fully model-based (i.e., no direct TD updates from experience are used) but the same principle is equally applicable within Dyna. The

general idea is that if new experience or computation changes the value (or transition and reward functions) at a state, then these changes will have the most extreme effects on the state-action values for actions leading up to those states, and so those predecessors should have the highest priority for simulated updates. For example, if a novel reward is experienced following an action, with the standard TD algorithm, this reward will not have an effect on other actions that may lead to the reward state until those actions are taken, while a model-based system will be able to update other action values accordingly, but only with extensive computation. Under a Dyna algorithm, however, this reward value could be propagated to other cached, model-free action values through simulated sampling of actions. By sampling states in reverse order along trajectories leading to the reward state, for instance, 'backing up' the values to more distant states, this can happen quite efficiently without requiring any additional real experience (Foster and Wilson 2006).

A neural system that implements such an algorithm suggests a mechanism for exploitation by drugs of abuse, whereby values inflated by distorted prediction errors could preferentially be selected for backing up. In particular, the principle that model-based updates are prioritized toward areas of the state space with new learning will be directly compromised by inflated prediction errors, since these will drive new learning and thereby attract more priority for model-based updates. Thus, the standard dopamine-mediated drug abuse story, whereby effective prediction errors are enhanced by drug experiences even when no new reward information is available, now cleanly predicts such prioritized model-based value updates as well. The action values associated with drug-taking would continue to increase in such a scenario, and thus always be given high priority for backups. As a result, these inflated values would propagate throughout the model, even to actions not previously resulting in drug attainment that have some probability of leading to other inflated states, to the exclusion of other potential goals or even negative experiences that may occur subsequent to fulfillment. This may constitute a computational description of phenomena associated with drug abuse, such as salience-driven sensitization or motivational magnets (Di Ciano 2008; Robinson and Berridge 2008), and can also explain suggestions that even goal-directed drug-seeking actions are insensitive to devaluation (Root et al. 2009). Here, the high priority given to such continually changing values is analogous to high salience for drug-associated stimuli.

Finally, a related phenomenon observed in drug abuse that might be similarly explained in this framework is cue-specific craving, in which stimuli associated with drug-taking result in increased drug-seeking motivation (Meil and See 1996; Garavan et al. 2000; Bonson et al. 2002; See 2005; Volkow et al. 2008). A potentially related effect in psychology is known as outcome-specific Pavlovian-instrumental transfer (PIT), in which presentation of cues associated with a particular reward increase the preference for instrumental actions associated with the same reward (Lovibond 1983; Rescorla 1994). A pure model-free learning system has no way to explain these effects, as action values abstract specific outcomes, and so while cues could generally enhance motivation, they cannot do so in an outcome-specific way. Further, it is unclear why cues in themselves should change an agent's action preferences or valuations, since the cues do not in fact carry information relevant to action

valuation. A model-based system, however, stores specific outcomes as part of the reward function. These, in the Dyna framework, may be used to drive simulation priorities for value updates. Through a priority mechanism, and since this approach allows on-the-fly updating of model-free values based on model-driven updates, it could theoretically drive updates preferentially toward a cued goal, ignoring other rewards to effect an updated value map more biased toward that outcome. Similarly, a drug-associated cue could simply trigger further updates back from objective states, pushing the values for related actions higher.

5.4 Conclusion

Drug abuse is a disorder of decision making, and as such its phenomena are relevant to and can be informed by the established computational theories of the domain. Building on two-system theories of learned decision making (Dickinson 1985; Balleine and Dickinson 1998; Poldrack et al. 2001; Daw et al. 2005; Wood and Neal 2007) and on the broad taxonomy of their potential vulnerabilities to drugs of abuse by Redish et al. (2008), we have considered the implications of drug-seeking behavior for algorithms and architectures hypothesized to comprise such a system. Drugs of abuse are commonly thought to target a habit learning system, specifically via their effects on dopamine and resultant amplification of model-free prediction errors. That they appear to serve as incentives for goal-directed behavior as well strongly suggests that the two decision systems interchange information rather than operating independently. We suggest this interchange might be captured within a modified architecture, such as Dyna or tree search with partial evaluation, allowing model-free and model-based influences to converge within a single representation. Importantly, such a mechanism, coupled with a scheduling principle for model-based searches like prioritized-sweeping, allows the single, ubiquitous, model-free mechanism of drug action to account for the range of behavioral phenomena.

The implications of these hypothesized mechanisms for decision making theories more generally remain to be developed. In particular, previous work has addressed a range of data on how animals' behaviors are differentially sensitive to devaluation in different circumstances by assuming two separate RL algorithms whose preferences were arbitrated according to relative uncertainty (Daw et al. 2005). It remains to be seen whether the same phenomena can be understood in the more integrated architectures suggested here, either in terms of prioritized sweeping heuristics or, alternatively, by developing the uncertainty explanation in this setting. That said, indications are accumulating rapidly, beyond the context of drugs of abuse, that the systems are more interactive than was assumed in previous theories (Root et al. 2009; Bromberg-Martin et al. 2010; van der Meer et al. 2010; Daw et al. 2011; Simon and Daw 2011). This accumulation of evidence strongly motivates the investigation of hybrid algorithms and interacting architectures of the type discussed here to expand our understanding of the range of strategies by which humans make decisions.

Acknowledgements The authors are supported by a Scholar Award from the McKnight Foundation, a NARSAD Young Investigator Award, Human Frontiers Science Program Grant RGP0036/2009-C, and NIMH grant 1R01MH087882-01, part of the CRCNS program.

References

Ainslie G (2001) Breakdown of will. Cambridge University Press, Cambridge

Arkadir D, Morris G, Vaadia E, Bergman H (2004) Independent coding of movement direction and reward prediction by single pallidal neurons. J Neurosci 24(45):10047–10056

Balleine BW, Daw ND, O'Doherty JP (2008) Multiple forms of value learning and the function of dopamine. In: Glimcher PW, Camerer CF, Fehr E, Poldrack RA (eds) Neuroeconomics: decision making and the brain. Academic Press, London, pp 367–387

Balleine BW, Delgado MR, Hikosaka O (2007) The role of the dorsal striatum in reward and decision-making. J Neurosci 27(31):8161–8165

Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37(4–5):407–419

Bechara A (2005) Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. Nat Neurosci 8(11):1458–1463

Berns GS, McClure SM, Pagnoni G, Montague PR (2001) Predictability modulates human brain response to reward. J Neurosci 21(8):2793–2798

Blodgett HC, McCutchan K (1947) Place versus response learning in the simple T-maze. J Exp Psychol 37(5):412–422

Bonson KR, Grant SJ, Contoreggi CS, Links JM, Metcalfe J, Weyl HL et al (2002) Neural systems and cue-induced cocaine craving. Neuropsychopharmacology 26(3):376–386

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. J Neurophysiol 104(2):1068–1076

Buckner RL (2010) The role of the hippocampus in prediction and imagination. Annu Rev Psychol 61:27–48, C1-8

Buckner RL, Carroll DC (2007) Self-projection and the brain. Trends Cogn Sci 11(2):49–57

Carr MF, Jadhav SP, Frank LM (2011) Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nat Neurosci 14(2):147–153

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29(39):12315–12320

Davidson TJ, Kloosterman F, Wilson MA (2009) Hippocampal replay of extended experience. Neuron 63(4):497–507

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16(2):199–204

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

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan R (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69(6):1204–1215

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. J Neurophysiol 84(6):3072–3077

Derdikman D, Moser M-B (2010) A dual role for hippocampal replay. Neuron 65(5):582–584

Di Chiara G (1999) Drug addiction as dopamine-dependent associative learning disorder. Eur J Pharmacol 375(1–3):13–30

Di Ciano P (2008) Facilitated acquisition but not persistence of responding for a cocaine-paired conditioned reinforcer following sensitization with cocaine. Neuropsychopharmacology 33(6):1426–1431

Dickinson A (1985) Actions and habits: The development of behavioural autonomy. Philos Trans R Soc Lond B, Biol Sci 308:67–78

Dickinson A, Balleine B (2002) The role of learning in the operation of motivational systems. In: Stevens' handbook of experimental psychology. Wiley, New York

Doya K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw 12(7–8):961–974

Dragoi G, Tonegawa S (2011) Preplay of future place cell sequences by hippocampal cellular assemblies. Nature 469(7330):397–401

Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 8(11):1481–1489

Everitt BJ, Dickinson A, Robbins TW (2001) The neuropsychological basis of addictive behaviour. Brains Res Rev 36(2–3):129–138

Faure A, Haberland U, Condé F, Massioui NE (2005) Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. J Neurosci 25(11):2771–2780

Foster DJ, Wilson MA (2006) Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440(7084):680–683

Garavan H, Pankiewicz J, Bloom A, Cho JK, Sperry L, Ross TJ et al (2000) Cue-induced cocaine craving: neuroanatomical specificity for drug users and drug stimuli. Am J Psychiatry 157(11):1789–1798

Gläscher J, Daw ND, Dayan P, O'Doherty JP (2010) States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66(4):585–595

Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26(32):8360–8367

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci 105(18):6741–6746

Hare TA, O'Doherty JP, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci 28(22):5623–5630

Hasselmo ME (2008) Temporally structured replay of neural activity in a model of entorhinal cortex, hippocampus and postsubiculum. Eur J Neurosci 28(7):1301–1315

Houk JC, Adams JL, Barto AG (1994) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk JC, Davis JL, Beiser DG (eds) Models of information processing in the basal ganglia. MIT Press, Cambridge, pp 249–270

Johnson A, Redish AD (2005) Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. Neural Netw 18(9):1163–1171

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10(12):1625–1633

Kahneman D, Frederick S (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: Gilovich T, Griffin DW, Kahneman D (eds) Heuristics and biases: the psychology of intuitive judgement. Cambridge University Press, New York, pp 49–81

Kalivas PW, Volkow ND (2005) The neural basis of addiction: a pathology of motivation and choice. Am J Psychiatry 162(8):1403–1413

Killcross S, Coutureau E (2003) Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb Cortex 13(4):400–408

Kim H, Sul JH, Huh N, Lee D, Jung MW (2009) Role of striatum in updating values of chosen actions. J Neurosci 29(47):14701–14712

Koene RA, Hasselmo ME (2008) Reversed and forward buffering of behavioral spike sequences enables retrospective and prospective retrieval in hippocampal regions CA3 and CA1. Neural Netw 21(2–3):276–288

Lansink CS, Goltstein PM, Lankelma JV, McNaughton BL, Pennartz CMA (2009) Hippocampus leads ventral striatum in replay of place-reward information. PLoS Biol 7(8):e1000173

Loewenstein G, O'Donoghue T (2004) Animal spirits: Affective and deliberative processes in economic behavior (Working Papers Nos. 04–14). Cornell University, Center for Analytic Economics

Lovibond PF (1983) Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. J Exp Psychol, Anim Behav Processes 9(3):225–247
McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38(2):339–346
van der Meer MAA, Johnson A, Schmitzer-Torbert NC, Redish AD (2010) Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. Neuron 67(1):25–32
Meil W, See R (1996) Conditioned cued recovery of responding following prolonged withdrawal from self-administered cocaine in rats: an animal model of relapse. Behav Pharmacol 7(8):754–763
Moore AW, Atkeson CG (1993) Prioritized sweeping: Reinforcement learning with less data and less time. Mach Learn 13:103–130. (10.1007/BF00993104)
Nordquist RE, Voorn P, de Mooij-van Malsen JG, Joosten RNJMA, Pennartz CMA, Vanderschuren LJMJ (2007) Augmented reinforcer value and accelerated habit formation after repeated amphetamine treatment. Eur Neuropsychopharmacol 17(8):532–540
O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38(2):329–337
Olmstead MC, Lafond MV, Everitt BJ, Dickinson A (2001) Cocaine seeking by rats is a goal-directed action. Behav Neurosci 115(2):394–402
Pan X, Sawa K, Sakagami M (2007) Model-based reward prediction in the primate prefrontal cortex. Neurosci Res 58(Suppl 1):229
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(4):774–777
Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27(37):9984–9988
Poldrack RA, Clark J, Paré-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C et al (2001) Interactive memory systems in the human brain. Nature 414(6863):546–550
Rangel A, Camerer C, Montague P (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev, Neurosci 9(7):545–556
Redish AD (2004) Addiction as a computational process gone awry. Science 306(5703):1944–1947
Redish AD, Johnson A (2007) A computational model of craving and obsession. Ann NY Acad Sci 1104(1):324–339
Redish AD, Jensen S, Johnson A (2008) Addiction as vulnerabilities in the decision process. Behav Brain Sci 31(04):461–487
Rescorla RA (1994) Control of instrumental performance by Pavlovian and instrumental stimuli. J Exp Psychol, Anim Behav Processes 20(1):44–50
Robinson TE, Berridge KC (2008) The incentive sensitization theory of addiction: some current issues. Philos Trans R Soc Lond B, Biol Sci 363(1507):3137–3146
Root DH, Fabbricatore AT, Barker DJ, Ma S, Pawlak AP, West MO (2009) Evidence for habitual and goal-directed behavior following devaluation of cocaine: a multifaceted interpretation of relapse. PLoS ONE 4(9):e7170
Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310(5752):1337–1340
Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80(1):1–27
Schultz W (2011) Potential vulnerabilities of neuronal reward, risk, and decision mechanisms to addictive drugs. Neuron 69(4):603–617
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275(5306):1593–1599
See RE (2005) Neural substrates of cocaine-cue associations that trigger relapse. Eur J Pharmacol 526(1–3):140–146
Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. J Neurosci 31(14):5526–5539

Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3(1):9–44

Sutton RS (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proceedings of the seventh International Conference on Machine Learning. Morgan Kaufmann, San Mateo, pp 216–224

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

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7(8):887–893

Tanaka SC, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S et al (2006) Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. Neural Netw 19(8):1233–1241

Thorndike EL (1898) Animal intelligence: An experimental study of the associative processes in animals. Psychol Rev Monogr Suppl 2(4):1–8

Tiffany ST (1990) A cognitive model of drug urges and drug-use behavior: Role of automatic and nonautomatic processes. Psychol Rev 97(2):147–168

Tindell AJ, Smith KS, Berridge KC, Aldridge JW (2009) Dynamic computation of incentive salience: "wanting" what was never "liked". J Neurosci 29(39):12220–12228

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55:189–208

Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. Science 315(5811):515–518

Vanderschuren LJMJ, Everitt BJ (2004) Drug seeking becomes compulsive after prolonged cocaine self-administration. Science 305(5686):1017–1019

Verplanken B, Aarts H, van Knippenberg AD, Moonen A (1998) Habit versus planned behaviour: a field experiment. Br J Soc Psychol 37(1):111–128

Volkow ND, Wang G-J, Telang F, Fowler JS, Logan J, Childress A-R et al (2008) Dopamine increases in striatum do not elicit craving in cocaine abusers unless they are coupled with cocaine cues. NeuroImage 39(3):1266–1273

Wood W, Neal DT (2007) A new look at habits and the habit-goal interface. Psychol Rev 114(4):843–863

Wunderlich K, Rangel A, O'Doherty JP (2009) Neural computations underlying action-based decision making in the human brain. Proc Natl Acad Sci 106(40):17199–17204

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(1):181–189

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22(2):513–523