Chapter 6
Modeling Decision-Making Systems in Addiction

Zeb Kurth-Nelson and A. David Redish

Abstract This chapter describes addiction as a failure of decision-making systems. Existing computational theories of addiction have been based on temporal difference (TD) learning as a quantitative model for decision-making. In these theories, drugs of abuse create a non-compensable TD reward prediction error signal that causes pathological overvaluation of drug-seeking choices. However, the TD model is too simple to account for all aspects of decision-making. For example, TD requires a state-space over which to learn. The process of acquiring a state-space, which involves both situation classification and learning causal relationships between states, presents another set of vulnerabilities to addiction. For example, problem gambling may be partly caused by a misclassification of the situations that lead to wins and losses. Extending TD to include state-space learning also permits quantitative descriptions of how changing representations impacts patterns of intertemporal choice behavior, potentially reducing impulsive choices just by changing cause-effect beliefs. This approach suggests that addicts can learn healthy representations to recover from addiction. All the computational models of addiction published so far are based on learning models that do not attempt to look ahead into the future to calculate optimal decisions. A deeper understanding of how decision-making breaks down in addiction will certainly require addressing the interaction of drugs with model-based look-ahead decision mechanisms, a topic that remains unexplored.

Decision-making is a general process that applies to all the choices made in life, from which ice cream flavor you want to whether you should use your children's college savings to buy drugs. Neural systems evolved to make decisions about what actions to take to keep an organism alive, healthy and reproducing. However, the same decision-making processes can fail under particular environmental or pharmacological conditions, leading the decision-maker to make pathological choices.

Z. Kurth-Nelson · A.D. Redish (✉)
Department of Neuroscience, University of Minnesota, 6-145 Jackson Hall, 321 Church St. SE, Minneapolis, MN 55455, USA
e-mail: redish@umn.edu

Z. Kurth-Nelson
e-mail: kurt0073@umn.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_6, © Springer Science+Business Media, LLC 2012

Both substance addiction and behavioral addictions such as gambling can be viewed in this framework, as failures of decision-making.

The simplest example of a failure in decision-making is in response to situations that are engineered to be disproportionately rewarding. In the wild, sweetness is a rare and useful signal of nutritive value, but refined sugar exploits this signal, and given the opportunity, people will often select particularly sweet foods over more nutritive choices. A more dangerous failure mode can be found in drugs of abuse. These drugs appear to directly modulate elements of the decision-making machinery in the brain, such that the system becomes biased to choose drug-seeking actions.

There are three central points in this chapter. First, a mathematical language of decision-making is developed based on *temporal difference (TD)* algorithms applied to *reinforcement learning (RL)* (Sutton and Barto 1998). Within this mathematical language, we review existing quantitative theories of addiction, most of which are based on identified failure modes within that framework (Redish 2004; Gutkin et al. 2006; Dezfouli et al. 2009). However, we will also discuss evidence that the framework is incomplete and that there are decision-making components that are not easily incorporated into the TD-RL framework (Dayan and Balleine 2002; Daw et al. 2005; Balleine et al. 2008; Dayan and Seymour 2008; Redish et al. 2008). Second, an organism's understanding of the world is central to its decision-making. Two organisms that perceive the contingencies of an experiment differently will behave differently. We extend quantitative decision-making theories to account for ways that organisms identify and utilize structure in the world to make decisions (Redish et al. 2007; Courville 2006; Gershman et al. 2010), which may be altered in addiction. Third, decision-making models naturally accommodate a description of how future rewards can be compared to immediate ones (Sutton and Barto 1998; Redish and Kurth-Nelson 2010). Both drug and behavioral addicts often exhibit impulsive choice, where a small immediate reward is preferred over a large delayed reward (Madden and Bickel 2010). There is evidence that impulsivity is both cause and consequence of addiction (Madden and Bickel 2010; Rachlin 2000). In particular, a key factor in recovery from addiction seems to be the ability to take a longer view on one's decisions and the ability to construct representations that support healthy decision-making (Ainslie 2001; Heyman 2009; Kurth-Nelson and Redish 2010).

6.1 Multiple Decision-Making Systems, Multiple Vulnerabilities to Addiction

Organisms use a combination of decision-making strategies. When faced with a choice, a human or animal may employ one or more of these strategies to produce a decision. The strategies used may also change with experience. For example, a classic experiment in rodent navigation involves a plus-shaped maze with four arms. On each trial, a food reward is placed in the east arm of the maze and the animal is placed in the south arm. The animal quickly learns to turn right to

the east arm to reach the food. On a probe trial, the animal can be placed in the north arm instead of the south arm. If these probe trials are conducted early in the course of learning, the animal turns left to the east arm, indicating that the animal is following a *location-based strategy* that dynamically calculates appropriate actions based on new information. On the other hand, if probe trials are conducted after the animal has been overtrained on the original task, the animal turns right into the west arm of the maze, indicating that it is following a *response strategy* where actions are precalculated and stored (Tolman 1948; Restle 1957; Packard and McGaugh 1996).

These different decision-making systems have different neuroanatomical substrates. In the rodent navigation example, the location-based strategy requires hippocampal integrity (Barnes 1979; Packard and McGaugh 1996), while the response strategy is dependent on the integrity of lateral aspects of striatum (Packard and McGaugh 1996; Yin et al. 2004). The location-based system is more computationally intensive but is more flexible to changing environments, while the response-based system is quick to calculate but inflexible to changing environments (O'Keefe and Nadel 1978; Redish 1999).

How the results of these different decision-making systems are integrated into a final decision remains an important open question. Obviously, if the two predicted actions are incompatible (as in the example above where one system decides to turn right while the other decides to turn left) and the animal takes an action, then the results must be integrated by the time the signals reach the muscles to perform the action. For example, an oversight system could enable or disable the place and response strategies, or could decide between the suggested actions provided by the two systems. However, economic theory implies the results are integrated much sooner (Glimcher et al. 2008). In neuroeconomic theory, every possible outcome is assumed to have a *utility*. The utilities of any possible outcome can be represented in a *common currency*, allowing direct comparison of the expected utilities to select a preferred action. In between the two extremes of common currency and muscle-level integration, there is a wide range of possibilities for how different decision-making systems could interact to produce a single decision. For example, a location-based strategy and a response strategy could each select an action (e.g., "turn left" or "turn right"), and these actions could compete to be transformed into a motor pattern.

In the following sections, we will develop a theoretical description of the brain's decision-making systems and show how drugs of abuse can access specific failure modes that lead to addictive choice. Addictive drugs have a variety of pharmacological effects on the brain, ranging from blockade of dopamine transporters to agonism of μ-opioid receptors to antagonism of adenosine receptors. Fundamentally, the common effect of addictive drugs is to cause pathological over-selection of the drug-taking decision, but this may be achieved in a variety of ways by accessing vulnerabilities in the different decision-making systems. This theory suggests that addicts may use and talk about drugs differently depending on which vulnerability the drugs access, and that appropriate treatment will likely differ depending on how the decision-making system has failed (Redish et al. 2008). For example, craving and relapse are separable entities in addictive processes—overvaluation in a stimulus-response based system could lead to relapse of the

action of drug-taking even in the absence of explicit craving, while overvaluation in the value system could lead to explicit identifiable desires for drug, but may not necessarily lead to relapse (Redish and Johnson 2007; Redish et al. 2008; Redish 2009).

6.1.1 *Temporal Difference Reinforcement Learning and the Dopamine Signal*

To explain why reward learning seems to occur only when an organism is confronted with an unexpected reward, Rescorla and Wagner (1972) introduced the idea of a *reward learning prediction error*. In their model, an agent (i.e., an organism or a computational model performing decision-making) learns how much reward is predicted by each cue, and generates a prediction error if the actual reward received does not match the net prediction of the cues they experienced. The prediction error is then used to update the reward prediction. To a first approximation, the fast phasic firing of midbrain dopamine neurons matches the Rescorla-Wagner prediction error signal (Ljungberg et al. 1992; Montague et al. 1996; Schultz 2002): when an animal is presented with an unexpected reward, dopamine neurons fire in a phasic burst of activity. If the reward is preceded by a predictive cue, the phasic firing of dopamine neurons gradually diminishes over several trials. The loss of dopamine firing at reward matches the loss of Rescorla-Wager prediction error, as the reward is no longer unpredicted.

However, there are several phenomena that the Rescorla-Wagner model does not account for. First, in animal behavior, conditioned stimuli can also act as reinforcers (Domjan 1998), and this shift is also reflected in the dopamine signals (Ljungberg et al. 1992). The Rescorla-Wagner model cannot accommodate this shift in reinforcement (Niv and Montague 2008). Second, a greater latency between stimulus and reward slows learning, reduces the amount of responding at the stimulus, and reduces dopamine firing at the stimulus (Mackintosh 1974; Domjan 1998; Bayer and Glimcher 2005; Fiorillo et al. 2008). The Rescorla-Wagner model does not represent time and cannot account for any effects of timing. Third, the Rescorla-Wagner model is a model of Pavlovian prediction and does not address instrumental action-selection. A generalized version of the Rescorla-Wagner model that accounts for stimulus chaining, temporal effects and action-selection is temporal difference reinforcement learning (TDRL).

Reinforcement learning is the general problem of how to learn what actions to take in order to maximize reward. Temporal difference learning is a common theoretical approach to solving the problem of reinforcement learning (Sutton and Barto 1998). Although the agent may be faced with a complex sequence of actions and observations before receiving a reward, temporal difference learning allows the agent to assign a value to each action along the way.

In order to apply a mathematical treatment, TDRL formalizes the learning problem as a set of states and transitions that define the situation of the animal and how

that situation can change (for example, see the very simple state-space in Fig. 6.1A). This collection of states and transitions is called a *state-space*, and defines the cause-effect relationships of the world that pertain to the agent. The agent maintains an estimate, for each state, of the reward it expects to receive in the future of that state. This estimate of future reward is called *value*, or V. We will use S_t to refer to the state of the agent at time t; $V(S_t)$ is the value of this state.

When the agent receives reward, it compares this reward with the amount of reward it expected to receive at that moment. Any difference is an error signal, called δ, which represents how incorrect the prior expectation was.

$$\delta = (R_t + V(S_t)) \cdot disc(d) - V(S_{t-1}) \tag{6.1}$$

where R_t is the reward at time t, d is the time spent in state S_{t-1}, and *disc* is a monotonically decreasing temporal discounting function with a range from 0 to 1. (Note that in the *semi-Markov* formulation of temporal difference learning (Daw 2003; Si et al. 2004; Daw et al. 2006), which we use here, the world can dwell in each state for an extended period of time.) A commonly used discounting function is

$$disc(d) = \gamma^d \tag{6.2}$$

where $\gamma \in [0, 1]$ is the exponential discounting rate. δ (Eq. (6.1)) is zero if the agent correctly estimated the value of state S_{t-1}; that is, it correctly identified the discounted future reward expected to follow that state. The actual reward received immediately following S_{t-1} is R_t, and the future reward expected after S_t is $V(S_t)$. Together, $R_t + V(S_t)$ is the future reward expected following S_{t-1}. This is discounted by the delay between S_{t-1} and S_t. The difference between this and the prior expectation $V(S_{t-1})$ is the value prediction error δ.

The estimated value of state S_{t-1} is updated proportional to δ, so that the expectation is brought closer to reality.

$$V(S_{t-1}) \leftarrow V(S_{t-1}) + \delta \cdot \alpha \tag{6.3}$$

where $\alpha \in (0, 1)$ is a learning rate. With appropriate exploration parameters and unchanging state space and reward contingencies, this updating process is guaranteed to converge on the correct expectation of discounted future reward for each state (Sutton and Barto 1998). Once reward expectations are learned, the agent can choose the actions that lead to the states with highest expected reward.

6.1.2 Value Prediction Error as a Failure Mode

The psychostimulants, including cocaine and amphetamine, directly increase dopamine action at the efferent targets of dopaminergic neurons (Ritz et al. 1987; Phillips et al. 2003; Aragona et al. 2008). The transient, or *phasic*, component of dopamine neuron firing appears to carry a reward prediction error signal like δ

(Montague et al. 1996; Schultz et al. 1997; Tsai et al. 2009). Thus, the psychostimulant drugs may act by pharmacologically increasing the δ signal (di Chiara 1999; Bernheim and Rangel 2004; Redish 2004).

Redish (2004) implemented this hypothesis in a computational model. Drug delivery was simulated by adding a non-compensable component to δ,

$$\delta = \max(D_t,\ D_t + (R_t + V(S_t)) \cdot disc(d) - V(S_{t-1})) \tag{6.4}$$

This is the same as Eq. (6.1) with the addition of a D_t term representing the drug delivered at time t. The value of δ cannot be less than D_t, due to the max function. The effect of D_t is that even after $V(S_{t-1})$ has reached the correct estimation of future reward, $V(S_{t-1})$ will keep growing without bound. In other words, D_t can never be compensated for by increasing $V(S_{t-1})$, so δ is never driven to zero. If there is a choice between a state that leads to drugs and a state that does not, the state leading to drugs will eventually (after a sufficient number of trials) have a higher value and thus be preferred.

This model exhibits several features of real drug addiction. The degree of preference for drugs over natural rewards increases with drug experience. Further, drug use is less sensitive to costs (i.e., drugs are less elastic) than natural rewards, and the elasticity of drug use decreases with experience (Christensen et al. 2008). Like other neuroeconomic models of addiction (e.g., Becker and Murphy (1988)), the Redish (2004) model predicts that even highly addicted individuals will still be sensitive to drug costs, albeit less sensitive than non-addicts, and less sensitive than to natural reward costs. (Even though they are willing to pay remarkably high costs to feed their addiction, addicts remain sensitive to price changes in drugs (Becker et al. 1994; Grossman and Chaloupka 1998; Liu et al. 1999).) The Redish (2004) model achieves inelasticity due to overvaluation of drugs of abuse.

The hypotheses that phasic dopamine serves as a value prediction error signal in a Rescorla-Wagner or TDRL-type learning system and that cocaine increases that phasic dopamine signal imply that Kamin blocking should not occur when cocaine is used as a reinforcer. In Kamin blocking (Kamin 1969), a stimulus X is first paired with reward until the X$\rightarrow$reward association is learned. (The existence of a learned association is measured by testing whether the organism will respond to the stimulus.) Then stimuli X and Y are together paired with reward. In this case, no association between Y and reward is learned. The Rescorla-Wagner model explains this result by saying that because X already fully predicts reward, there is no prediction error and thus no learning when X and Y are paired with reward. Consistent with the dopamine-as-δ hypothesis, phasic dopamine signals do not appear in response to the blocked stimuli (Waelti et al. 2001). However, if the blocking experiment is performed with cocaine instead of a natural reinforcer, the hypothesis that cocaine produces a non-compensable δ signal predicts that the δ signal should still occur when training XY$\rightarrow$cocaine, so the organism should learn to respond for Y. Contrary to this prediction, Panlilio et al. (2007) recently provided evidence that blocking does occur with cocaine in rats, implying that either the phasic dopamine signal is not equivalent to the δ signal, or cocaine does not boost phasic dopamine. Recently, Jaffe et al. (2010) presented data that a subset of high-responding animals

did not show Kamin blocking when faced with nicotine rewards, suggesting that the lack of Kamin blocking may produce overselection of drug rewards in a subset of subjects. An extension to the Redish model to produce overselection of drug rewards while still accounting for blocking with cocaine is given by Dezfouli et al. (2009) (see also Chap. 8 in this book). In this model, new rewards are compared against a long-term average reward level. Drugs increase this average reward level, so the effect of drugs is compensable and the δ signal goes to zero with long-term drug exposure. If this model is used to simulate the blocking experiment with cocaine as the reinforcer, then during the X→cocaine training, the average reward level is elevated, so that when XY→cocaine occurs, there is no prediction error signal and Y does not acquire predictive value.

Other evidence also suggests that the Redish (2004) model is not a complete picture. First, the hypotheses of the model imply that continued delivery of cocaine will eventually overwhelm any reinforcer whose prediction error signal is compensable (such as a food reward). Recent data (Lenoir et al. 2007) suggest that this is not the case, implying that the Redish (2004) model is not a complete picture. Second, the Redish (2004) model is based on the assumption that addiction arises from the action of drugs on the dopamine system. Many addictive drugs do not act directly on dopamine (e.g., heroin, which acts on μ-opioid receptors (Nestler 1996)), and some drugs that boost dopamine are not addictive (e.g., bupropion (Stahl et al. 2004)). Most psychostimulant drugs also have other pharmacological effects; for example, cocaine also has an action on the norepinephrine and serotonin systems (Kuhar et al. 1988). Norepinephrine has been implicated in signaling uncertainty (Yu and Dayan 2005) and attention (Berridge et al. 1993), while serotonin has other effects on decision-making structures in the brain (Tanaka et al. 2007). All of these actions could also potentially contribute to the effects of cocaine on decision-making.

Action selection can be performed in a variety of ways. When multiple actions are available, the agent may choose the action leading to the highest valued state. Alternatively, the benefit of each action may be learned separately from state values. Separating *policy learning* (i.e., learning the benefit of each action) from value learning has the theoretical advantage of being easier to compute when there are many available actions (for example, if the action space is continuous Sutton and Barto 1998). In this case, the policy learning system is called the *actor* and the value learning system is called the *critic*. The actor and critic systems have been proposed to correspond to different brain structures (Barto 1994; O'Doherty et al. 2004; Daw and Doya 2006). The dopamine-as-δ hypothesis can provide another explanation for drug addiction if learning in the critic system is saturable. During actor learning, feedback from the critic is required to calculate how much unexpected reinforcement occurred, and thus how much the actor should learn. If drugs produce a large increase in δ that cannot be compensated for by the saturated critic, then the actor will over-learn the benefit of the action leading to this drug-delivery (see Chap. 8 in this book).

The models we have discussed so far use the assumption that decision-making is based on learning, for each state, an expectation of future value that can be expressed in a common currency. There are many experiments that show

that not all decisions are explicable in this way (Balleine and Dickinson 1998; Dayan 2002; Daw et al. 2005; Dayan and Seymour 2008; Redish et al. 2008; van der Meer and Redish 2010). The limitations of the temporal difference models can be addressed by incorporating additional learning and decision-making algorithms (Pavlovian systems, deliberative systems) and by addressing the representations of the world over which these systems work.

6.1.3 Pavlovian Systems

Unconditioned stimuli can provoke an approach or avoidance response that does not depend on the instrumental contingencies of the experiment (Mackintosh 1974; Dayan and Seymour 2008). These Pavlovian systems can produce non-optimal decisions in some animals under certain conditions (Breland and Breland 1961; Balleine 2001, 2004; Dayan et al. 2006; Uslaner et al. 2006; Flagel et al. 2008; Ostlund and Balleine 2008). For example, in a classic experiment, birds were placed on a linear track, near a cup of food that was mechanically designed to move in the same direction as the bird, at twice the bird's speed. The optimal strategy for the bird was to move away from the food until the food reached the bird, but in the experiment, birds never learned to move away; instead always chasing the food to a greater distance (Hershberger 1986). Theories of Pavlovian influence on decision-making suggest that the food-related cues provoked an approach response (Breland and Breland 1961; Dayan et al. 2006). Similarly, if animals are trained that a cue predicts a particular reward in a Pavlovian conditioning task, later presenting that cue during an instrumental task in which one of the choices leads to that reward will increase preference for that choice (Pavlovian-instrumental transfer (Estes 1943; Kruse et al. 1983; Lovibond 1983; Talmi et al. 2008)). Although models of Pavlovian systems exist (Balleine 2001, 2004; Dayan et al. 2006) as do suggestions that Pavlovian failures underlie aspects of addiction (Robinson and Berridge 1993, 2001, 2004; Berridge 2007), computational models of addiction taking into account interactions between Pavlovian effects and temporal difference learning are still lacking.

6.1.4 Deliberation, Forward Search and Executive Function

During a decision, the brain may explicitly consider alternatives in order to predict outcomes (Tolman 1939; van der Meer and Redish 2010). This process allows evaluation of those outcomes in the light of current goals, expectations, and values (Niv et al. 2006). Therefore part of the decision-making process plausibly involves predicting the future situation that will arise from taking a choice and accessing the reinforcement associations that are present in that future situation. This stands in contrast to decision-making strategies that use only the value associations present in the current situation.

When rats running in a maze come to an important choice-point where they could go right or left and possibly receive reward, they will sometimes pause and turn their head from side to side as if to sample the options. This is known as vicarious trial and error (VTE) (Muenzinger 1938; Tolman 1938, 1939, 1948). VTE behavior is correlated to hippocampal activity and is reduced by hippocampal lesions (Hu and Amsel 1995; Hu et al. 2006). During most behavior, cells in the hippocampus encode the animal's location in space (O'Keefe and Dostrovsky 1971; O'Keefe and Nadel 1978; Redish 1999). But during VTE, this representation sometimes projects forward in one direction and then the other (Johnson and Redish 2007). Johnson and Redish (2007) proposed that this "look-ahead" that occurs during deliberation may be part of the decision making process. By imagining the future, the animal may be attempting to determine whether each choice is rewarded (Tolman 1939, 1948). Downstream of the hippocampus, reward-related cells in the ventral striatum also show additional activity during this deliberative process (van der Meer and Redish 2009), which may be evidence for prediction and calculation of expectancies (Daw et al. 2005; Redish and Johnson 2007; van der Meer and Redish 2010).

Considering forward search as part of the decision making process permits a computational explanation for the phenomena of craving and obsession in drug addicts (Redish and Johnson 2007). Craving is the recognition of a high-value outcome, and obsession entails constraint of searches to a single high-value outcome. Current theories suggest that endogenous opioids signal the hedonic value of received rewards (Robinson and Berridge 1993). If these endogenous opioids also signal imagined rewards, then opioids may be a key to craving (Redish and Johnson 2007). This fits data that opioid antagonists reduce craving (Arbisi et al. 1999; Levine and Billington 2004). Under this theory, an opioidergic signal at the time of reward or drug delivery may cause neural plasticity in such a way that the dynamics of the forward search system become biased to search toward the outcome linked to the opioid signal. Activation of opioid receptors is known to modulate synaptic plasticity in structures such as the hippocampus (Liao et al. 2005), suggesting a possible physiological basis for altering forward search in the hippocampus.

6.2 Temporal Difference Learning in a Non-stationary Environment

Temporal difference learning models describe how to learn an expectation of future reward over a known state-space. In the real world, the state-space itself is not known a priori. It must be learned and may even change over time. This is illustrated by the problem of extinction and reinstatement. After a cue-reinforcer association is learned, it can be extinguished by presenting the cue alone (Domjan 1998). Over time, animals will learn to stop responding for the cue. If extinction is done in a different environment from the original learning, placing the animal back in the original environment causes responding to start again immediately (Bouton and Swartzentruber 1989). Similarly, even if acquisition and extinction occur in

the same environment, a single presentation of the reinforcer following extinction can cause responding to start again (Pavlov 1927; McFarland and Kalivas 2001; Bouton 2002). This implies that the original association was not unlearned during extinction. A similar phenomenon occurs in abstaining human drug addicts, where drug-related cues can trigger relapse to full resumption of drug-seeking behavior much faster than the original development of addiction (Jaffe et al. 1989; Childress et al. 1992). In extinction paradigms, the world is non-stationary: a cue that used to lead to a reward or drug-presentation now no longer does. Thus, a decision-making system trying to accurately predict the world requires a mechanism to construct state-spaces flexibly from the observed dynamics of the world. This mechanism does not exist in standard TDRL models.

To explain the phenomenon of renewal of responding after extinction, a recent model extended temporal difference learning by adding state-classification (Redish et al. 2007). In this model, the total information provided from the world to the agent at each moment was represented as an n-dimensional sensory cue. The model classified cue vectors into the same state if they were similar, or into different states if they were sufficiently dissimilar. During acquisition of a cue-reinforcer association, the model grouped these similar observations (many trials with the same cue)into a state representing "cue predicts reward". The model learned to associate the value of the reward with instrumental responding in this "cue predicts reward" state. This learning occurred at the learning rate of the model. During extinction, as the model accumulated evidence that a cue did not predict reward in a new context, these observations were classified into a new state representing "cue does not predict reward", from which actions had no value. When returned to the original context, the model switched back to classifying cue observations into the "cue predicts reward" state. Because instrumental responding in the "cue predicts reward" state had already been associated with reward during acquisition, no additional learning was needed, and responding immediately resumed at the pre-extinction rate.

This situation-classification component may be vulnerable to its own class of failures in decision-making. Based on vulnerabilities in situation-classification, Redish et al. (2007) were also able to simulate behavioral addiction to gambling. These errors followed both from over-separation of states, in which two states that were not actually different were identified as different due to unexpected consistencies in noise, and from over-generalization of states, in which two states that were different were not identified as different due to the similarities between them. The first process is similar to that of "the illusion of control" in which subjects misperceive that they have control of random situations, producing superstition (Langer and Roth 1975; Custer 1984; Wagenaar 1988; Elster 1999). The illusion of control can be created by having too many available cues, particularly when combined with the identification of near-misses (Cote et al. 2003; Parke and Griffiths 2004). The phenomenon of "chasing", in which subjects continue to place deeper and deeper losing bets, may arise because gamblers over-generalize a situation in which they received a large win, to form a belief that gambling generally leads to reward (Custer 1984; Wagenaar 1988;

Elster 1999). We suggest this is a problem of state-classification: the gamblers classify the generic gambling situation as leading to reward.

In the Redish et al. (2007) model, states were classified from sensory and reinforcement experience, but the transition structure of the world was not learned. Smith et al. (2006) took the converse approach. Here the algorithm started with a known set of states, each with equal temporal extent, and learned the transition probability matrix based on observed transitions. A "surprise" factor measured the extent to which a reinforcer was unpredicted by previous cues, also allowing the model to reproduce the Kamin blocking effect (Kamin 1969) and the reduction of latent inhibition by amphetamine (Weiner et al. 1988).

Both the Redish et al. (2007) and Smith et al. (2006) models are special cases of the more general *latent cause theory*, in which the agent attempts to identify hidden causes underlying sets of observations (Courville 2006; Gershman et al. 2010). In these models, agents apply an approximation of Bayesian statistical inference to all observations to infer hidden causes that could underlie correlated observations. Because latent cause models take into account any change in stimulus–stimulus or stimulus–outcome contingencies, these models are able to accommodate any non-stationary environment.

The ability of the brain to dynamically construct interpretations of the causal structure of the world is likely seated in frontal cortex and hippocampus. Hippocampus is involved in accommodating cue-reward contingency changes (Hirsh 1974; Isaacson 1974; Hirsh et al. 1978; Nadel and Willner 1980; Corbit and Balleine 2000; Fuhs and Touretzky 2007). Returning to a previously reinforced context no longer triggers renewal of extinguished responding if hippocampus is lesioned (Bouton et al. 2006). Medial prefrontal cortex appears to be required for learning the relevance of new external cues that signal altered reinforcement contingencies (Lebron et al. 2004; Milad et al. 2004; Quirk et al. 2006; Sotres-Bayon et al. 2006). Classification and causality representations in hippocampus and frontal cortex may form a cognitive input to the basal ganglia structures that perform reinforcement learning. Drugs of abuse that negatively impact the function of hippocampal or cortical structures could inhibit the formation of healthy state-spaces, contributing to addiction. Alcohol, for example, has been hypothesized to preferentially impair both hippocampal and prefrontal function (Hunt 1998; Oscar-Berman and Marinkovic 2003; White 2003).

In general, if the brain constructs state-spaces that do not accurately reflect the world but instead overemphasize the value of the addictive choice, this constitutes an addiction vulnerability. Behavioral addiction to gambling may arise from a failure of state classification as described above. Addiction to drugs could result from state-spaces that represent only the immediate choice and not the long-range consequences. This would suggest that training new state-space constructions, and mechanisms designed to prevent falling back into old state-spaces, may improve relapse outcomes in addicts.

6.3 Discounting and Impulsivity

In this section we will discuss the phenomenon of intertemporal choice (how the delay to a reward influences decisions), and show how changes in the agent's state-space can change the intertemporal decisions made by an organism.

If offered a choice between $10 right now and $11 tomorrow, many people will feel it is not worth waiting one day for that extra dollar, and choose the $10 now. When offered a choice between a small immediate reward and a large delayed reward, *impulsivity* is the extent to which the agent prefers the small immediate reward, being unwilling to wait for the future reward. This is sometimes viewed as a special case of temporal discounting, which is the general problem of how the value of rewards diminishes as they recede into the future.[1] As discussed above, a discounting function $disc(d)$ maps a delay d to a number in $[0, 1]$ specifying how much a reward's value is attenuated due to being postponed by time d. The impulsive decision to take a smaller-sooner reward rather than a larger-later one can be studied in the context of temporal difference learning.

Addicts tend to be more impulsive than non-addicts. It is easy to see why impulsivity could lead to addiction: the benefit of drug-taking tends to be more immediate than the benefits of abstaining. It is also possible that drugs increase impulsivity. Smokers discount faster than those who have never smoked, but ex-smokers discount at a rate similar to those who have never smoked (Bickel et al. 1999). In the Dezfouli et al. (2009) model, simulations show that choice for non-drug rewards becomes more impulsive following repeated exposure to drugs. Although the causal relationship between drug-taking and impulsivity is difficult to study in humans, animal data show that chronic drug-taking increases impulsivity (Paine et al. 2003; Simon et al. 2007).

If offered a choice between $10 right now and $11 tomorrow, many people will choose $10; however, if offered a choice between $10 in a year and $11 in a year and a day, the same people often prefer the $11 (Ainslie 2001). This is an example of *preference reversal*. Economically, the two decisions are equivalent and, under simple assumptions of stability, it should not matter if the outcomes are each postponed by a year. But in practice, many experiments have found that the preferred option changes as the time of the present changes relative to the outcomes (Madden and Bickel 2010).

In principle, any monotonically decreasing function with a range from 0 to 1 could make a reasonable discounting function. Exponential discounting (as in Eq. (6.2)) is often used in theoretical models because it is easy to calculate and matches economic assumptions of behavior. However, preference reversal does not occur in exponential discounting, but does occur with any non-exponential

[1] There are multiple decision factors often referred to as "impulsivity", including the inability to inhibit a pre-potent response, the inability to inhibit an over-learned response, and an over-emphasis on immediate versus delayed rewards (which we are referring to here). These multiple factors seem to be independent (Reynolds et al. 2006) and to depend on different brain structures (Isoda and Hikosaka 2008) and we will not discuss the other factors here.

discounting function (Frederick et al. 2002). Discounting data in humans and animals generally does show preference reversal (Chung and Herrnstein 1967; Baum and Rachlin 1969; Mazur 1987; Kirby and Herrnstein 1995), indicating that organisms are not performing exponential discounting. Human and animal discounting data are often best fit by a hyperbolic discount function (Ainslie 2001):

$$disc(d) = \frac{1}{1 + kd} \tag{6.5}$$

where $k \in [0, \infty)$ is the discount rate. It is therefore important to consider how hyperbolic discounting can fit into reinforcement learning models.

Hyperbolic discounting is empirically a good fit to human and animal discounting data, but it also has a theoretical basis in uncertain hazard rates. Agents are assumed to discount future rewards because there is some risk that the reward will never be received, and this risk grows with temporal distance (but see Henly et al. 2008). Events that would prevent reward receipt, such as death of the organism, are called *interruptions*. If interruptions are believed to occur randomly at some rate (i.e., the hazard rate), then the economically optimal policy is exponential discounting at that rate. However, if the hazard rate is not known a priori, it could be taken to be a uniform distribution over the possible rates (ranging from 1 where interruptions never occur to 0 where interruptions occur infinitely fast). Under this assumption, the economically optimal policy is hyperbolic discounting (Sozou 1998). Using the data from a large survey, it was found that factoring out an individual's expectation and tolerance of risk leaves individuals with a discounting factor well-fit by an exponential discounting function (Andersen et al. 2008). This function was correlated with the current interest rate, suggesting that humans may be changing their discounting rates to fit the expected hazard functions. Studies in which subjects could maximize reward by discounting exponentially at particular rates have found that humans can match their discounting to those exponential functions (Schweighofer et al. 2006). However, neurological studies have found that risk and discounted rewards may be utilizing different brain structures (Preuschoff et al. 2006).

Semi-Markov temporal difference models, such as those described above, can represent varying time intervals within a single state, permitting any discount function to be calculated across a single state-transition. However, the value of a state is still calculated recursively using the discounted value of the next state (rather than looking ahead all the way to the reward). Thus, across multiple state-transitions, the discounting of semi-Markov models depends on the way that the total temporal interval between now and reward is divided between states. With exponential discounting, the same percent reduction in value occurs for a given delay, regardless of the absolute distance in the future. Because of this, exponential discounting processes convolve appropriately; that is, the discounted value of a reward R is independent of whether the transition is modeled as one state with delay d or two states with delay $d/2$. In contrast, hyperbolic discounting functions do not convolve to produce hyperbolic discounting across a sequence of multiple states, and the discounted value of a reward R depends on the number of state transitions encompassing the delay.

As a potential explanation for how hyperbolic discounting could be calculated in a way that is not dependent on the division of time into states, Kurth-Nelson and Redish (2009) noted that a hyperbolic discount function is mathematically equivalent to the sum of exponential discounting functions with a range of exponential discount factors.

$$\int_0^1 \gamma^x d\gamma = \frac{1}{1+x} \tag{6.6}$$

Kurth-Nelson and Redish extended TDRL using a population of "micro-agents", each of which independently performed temporal difference learning using exponential discounting. Each micro-agent used a different discount rate. Actions were selected in the model by a simple voting process among the micro-agents. The overall model exhibited hyperbolic discounting that did not depend on the division of time into states (Fig 6.1).

There is evidence that a range of discounting factors are calculated in the striatum, with a gradient from faster discount rates represented in ventral striatum to slower rates in dorsal striatum (Tanaka et al. 2004). Doya (2000) proposed that serotonin levels regulate which of these discounting rates are active. Tanaka et al. (2007) and Schweighofer et al. (2007) showed that changing serotonin levels (by loading/unloading the serotonin precursor tryptophan) produced changes in which components of striatum were active in a given task. Drugs of abuse could pharmacologically modulate different aspects of striatum (Porrino et al. 2004). Kurth-Nelson and Redish (2009) predicted that drugs of abuse may change the distribution of discount factors and thus speed discounting. The multiple-discount hypothesis predicts that if the distribution of discount rates is altered by drugs, the shape of the discounting curve will be altered as well.

6.3.1 Seeing Across the Intertrial Interval

Discounting is often operationally measured by offering the animal a choice between a smaller reward available sooner or a larger reward available later (Mazur 1987). In the mathematical language used in this chapter, this experiment can be modeled as a reinforcement learning state-space (Fig. 6.2). The discount rate determines whether the smaller-sooner or larger-later reward will be preferred by a temporal difference model.

Rather than running a single trial, the animal is usually required to perform multiple trials in sequence. In these experiments the total trial length is generally held constant (i.e. the intertrial interval following the smaller-sooner choice is longer than the intertrial interval following the larger-later choice) so that smaller-sooner does not become the superior choice simply by hastening the start of the next trial. This creates a theoretical paradox. On any individual trial, the animal may prefer the smaller-sooner option because of its discount rate. But consistently choosing smaller-sooner over larger-later only changes the phase of reward delivery and decreases the overall reward magnitude.

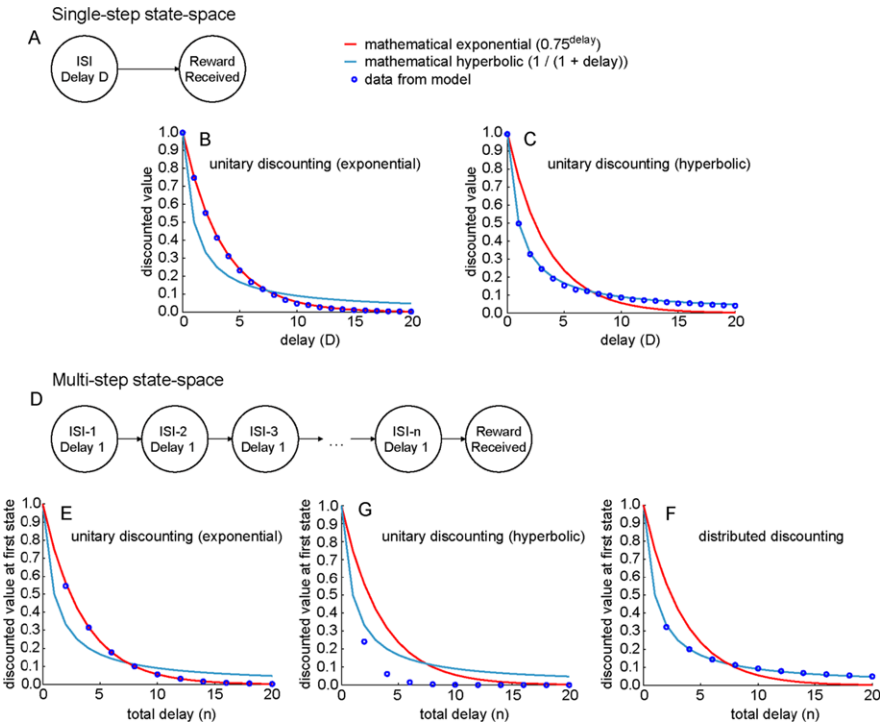

Fig. 6.1 Distributed discounting permits hyperbolic discounting across multiple state transitions. **A**, All delay between stimulus and reward is represented in a single state, permitting any discount function to be calculated over this delay, including exponential (**B**) or hyperbolic (**C**). (**D**) The delay between stimulus and reward is divided into multiple states. Exponential discounting (**E**) can still be calculated recursively across the entire delay (because $\gamma^a \gamma^b = \gamma^{a+b}$), but if hyperbolic discounting is calculated at each state transition, the net discounting at the stimulus is not hyperbolic (**G**). However, if exponential discounting is performed in parallel at many different rates, the average discounting across the entire time interval is hyperbolic (**F**). [From Kurth-Nelson and Redish (2009).]

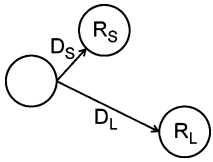

Fig. 6.2 A state-space representing intertemporal choice. From the initial state, a choice is available between a smaller reward (of magnitude R_S) available after a shorter delay (of duration D_S), or a larger reward (R_L) after a longer delay (D_L)

This suggests that there are two different potential state-space representations to describe this experiment. In one description, each trial is seen independently (Fig. 6.3, top); this is the standard approach in TDRL. In the other description,

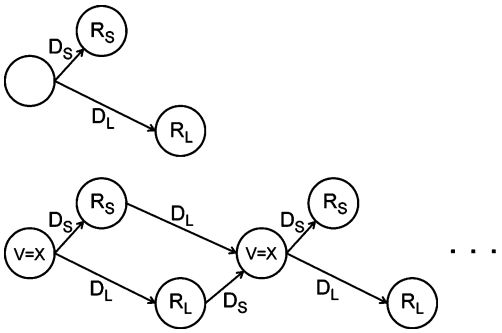

Fig. 6.3 Allowing the agent to see across the inter-trial interval changes the state-space representation of the task. *Top*, A state-space in which each trial is independent from the next. *Bottom*, A state-space in which the end of one trial has a transition to the beginning of the next trial, allowing the value estimates to include expectation of reward from future trials. The delays following the rewards are set to keep the total trial length constant. Note that the states are duplicated for illustrative purposes; an equivalent diagram would have only three states, with arrows wrapping back from R_S and R_L states to the initial choice state

the end of the last trial has a transition to the beginning of the next trial (Fig. 6.3, bottom). By adding this transition (which we will call a *wrap-around* transition), the algorithm can integrate expectation of future reward across all future trials. The total expectation is still convergent because future trials are discounted increasingly with temporal distance.

Adding a wrap-around transition to the state-space has the effect of slowing the apparent rate of discounting. Without wrap-around, the value of the smaller-sooner option is $R_S \cdot disc(D_S)$, and the value of the larger-later option is $R_L \cdot disc(D_L)$. With wrap-around, the smaller-sooner option becomes $R_S \cdot disc(D_S) + X$, and the larger-later option becomes $R_L \cdot disc(D_L) + X$, where X is the value of the initial state in which the choices are available. In other words, wrap-around adds the same constant to the reward expectation for each choice. Thus, if the smaller-sooner option was preferred without wrap-around, with wrap-around it is still preferred but to a lesser degree. Because additional delay devalues the future reward less (proportional to its total value), the apparent rate of discounting is reduced. Note that adding a wrap-around transition does not change the underlying discount function $disc(d)$, but the agent's behavior changes as if it were discounting more slowly. Also, because X is a constant added to both choices, X can change the degree to which the smaller-sooner option is preferred to the larger-later, but it cannot reverse the preference order. Thus, if the agent prefers the smaller-sooner option without a wrap-around state transition, adding wrap-around cannot cause the agent to switch to prefer the larger-later option.

If addicts could be influenced to change their state-space to see across the inter-trial interval, they should exhibit slower discounting. Heyman (2009) observes that recovered addicts have often made the time-course at which they view their lives more global. An interesting question is whether this reflects a change in state-space in the individuals.

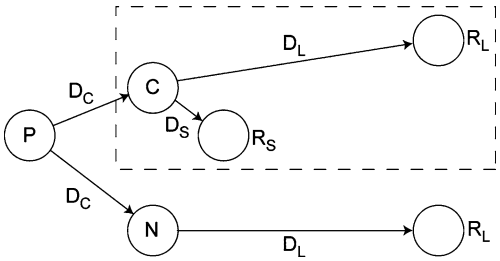

Fig. 6.4 A state-space in which the agent can make a precommitment to avoid having access to a smaller-sooner reward option. The portion of the state-space inside the *dashed box* is the smaller-sooner versus larger-later choice state-space shown in Fig. 6.2. Now a prechoice is available to enter the smaller-sooner versus larger-later choice, or to enter a situation from which only larger-later is available. Following the prechoice is a delay D_C

6.3.2 Precommitment and Bundling

The phenomenon of preference reversal suggests that an agent who can predict their own impulsivity may prefer to remove the future impulsive choice if given an opportunity (Strotz 1956; Ainslie 2001; Gul and Pesendorfer 2001; Heyman 2009; Kurth-Nelson and Redish 2010). For example, an addict may decline to visit somewhere drugs are available. When the drug-taking choice is viewed from a temporal distance, he prefers not to take drugs. But he knows that if faced with drug-taking as an immediate option, he will take it, so he does not wish to have the choice. Precommitment to larger-later choices by eliminating future smaller-sooner choices is a common behavioral strategy seen in successful recovery from addiction (Rachlin 2000; Ainslie 2001; Dickerson and O'Connor 2006; Heyman 2009).

Kurth-Nelson and Redish (2010) showed that precommitment behavior can be modeled with reinforcement learning. The reinforcement learning state-space for precommitment is represented in Fig. 6.4. The agent is given a choice to either enter a smaller-sooner versus larger-later choice, or to enter a situation where only the larger-later option is available. Because the agent discounts hyperbolically, the agent can prefer the smaller-sooner option when making the choice at C, but also prefer the larger-later option when making the earlier choice at P. Mathematically, when the agent is in state C, it is faced with a choice between two options with values $R_S \cdot disc(D_S)$ and $R_L \cdot disc(D_L)$. But when the agent is in state P, the choice is between two options with values $R_L \cdot disc(D_C + D_L)$ and $R_S \cdot disc(D_C + D_S)$. In hyperbolic discounting, the rate of discounting slows as rewards recede into the future, so $\frac{disc(D_S)}{disc(D_L)} > \frac{disc(D_C+D_S)}{disc(D_C+D_L)}$, meaning that the extra delay D_C makes the smaller-sooner choice relatively less valuable. This experiment has been performed in pigeons, and some pigeons consistently elected to take away a future impulsive choice from themselves, despite preferring that choice when it was available (Rachlin and Green 1972; Ainslie 1974). However, to our knowledge this experiment has not yet been run in humans or other species.

In order for a reinforcement learning agent to exhibit precommitment in the state-space in Fig. 6.4, it must behave in state P as if it were discounting R_S across the entire time interval $D_C + D_S$, and discounting R_L across the entire interval $D_C + D_L$. As noted earlier (cf. Fig. 6.1), hyperbolic discounting across multiple states cannot be done with a standard hyperbolic discounting model (Kurth-Nelson and Redish 2010). It requires a model such as the distributed discounting model (Kurth-Nelson and Redish 2009) described above. In this model, each μAgent has a different exponential discounting rate and has a different value estimate for each state. This model performs hyperbolic discounting across multi-step state-spaces (cf. Fig. 6.1) by not collapsing future reward expectation to a single value for each state. Thus, if the distributed discounting model is trained over the state-space of Fig. 6.4, it prefers the smaller-sooner option from state C, but from state P prefers to go to state N (Kurth-Nelson and Redish 2010).

Another way for an impulsive agent to regulate its future choices is with bundling (Ainslie 2001). In bundling, an agent reduces a sequence of future decisions to a single decision. For example, an alcoholic may recognize that having one drink is not a choice that can be made in isolation, because it will lead to repeated impulsive choice. Therefore the choice is between being an alcoholic or never drinking.

Consider the state-spaces in Fig. 6.5. If each choice is treated as independent, the value of the smaller-sooner choice is $R_S \cdot disc(D_S)$ and the value of the larger-later choice is $R_L \cdot disc(D_L)$. However, if making one choice is believed to also determine the outcome of the subsequent trial, then the value of smaller-sooner is $R_S \cdot disc(D_S) + R_S \cdot disc(D_S + D_L + D_S)$ and the value of larger-later is $R_L \cdot disc(D_L) + R_L \cdot disc(D_L + D_S + D_L)$. In an agent performing hyperbolic discounting, the attenuation of value produced by the extra $D_S + D_L$ delay is less if this delay comes later relative to the present. Thus bundling can change the agent's preferences so that the larger-later choice is preferred from the initial state. Like precommitment, bundling can be modeled with reinforcement learning, but only if the model correctly performs hyperbolic discounting across multiple state transitions (Kurth-Nelson and Redish 2010).

It is interesting to note that the agent can represent a given choice in a number of ways: existing in isolation (Fig. 6.3, top), leading to subsequent choices (Fig. 6.3, bottom), viewed in advance (Fig. 6.4), or viewed as a categorical choice (Fig. 6.5, bottom). These four different state-spaces are each reasonable representations of the same underlying choice, but produce very different behavior in reinforcement learning models. This highlights the importance of constructing a state-space for reinforcement learning. If state-space construction is a cognitive operation, it is possible that it can be influenced by semantic inputs. For example, perhaps by verbally suggesting to someone that the decision to have one drink cannot be made in isolation, they are led to create a state-space that reflects this idea.

Throughout these examples in which state-space construction has influenced the apparent discount rate, the *underlying* discount rate (the function $disc(d)$) is unaffected. The difference is in the agent's choice behavior, from which discounting is inferred. Since state-space construction in temporal difference models affects apparent discount rates, it may be that discounting in the brain is modulated by the capacity of the organism to construct state-spaces. This suggests that a potential treatment

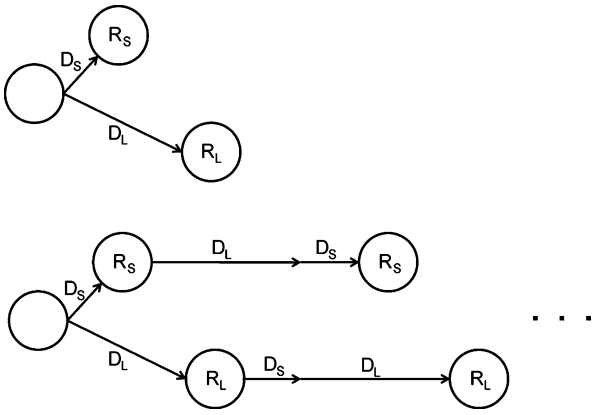

Fig. 6.5 Bundling two choices. *Top*, Each choice is made independently. *Bottom*, One choice commits the agent to make the same choice on the next trial

for addiction may lie in the creation of better state-spaces. Gershman et al. (2010) proposed that a limited ability to infer causal relations in the world explains the fact that young animals exhibit less context-dependence in reinforcement learning. This matches the data that people with higher cognitive skills exhibit slower discounting (Burks et al. 2009). It is also consistent with the emphasis of addiction treatment programs (such as 12-step programs) on cognitive strategies that alter the perceived contingencies of the world.

However, it is not clear that the learning systems for habitual or automatic behaviors always produce impulsive choice, or that the executive systems always produce non-impulsive choice. For example, smokers engage in complex planning to find the cheapest cigarettes, in line with the economic view that addicts should be sensitive to cost (Becker and Murphy 1988; Redish 2004). Addicts can perform very complex planning in order to get their drugs (Goldman et al. 1987; Goldstein 2000; Jones et al. 2001; Robinson and Berridge 2003). Thus it does not appear that the problem of addiction is simply a case of the habitual system pharmacologically programmed to carry out drug-seeking behaviors (as arises from the Redish (2004), Gutkin et al. (2006), or Dezfouli et al. (2009) models discussed above; see also Chap. 8 in this book). Rather, addictive drugs seem to have the potential to access vulnerabilities in multiple decision-making systems, including cognitive or executive systems. These different vulnerabilities are likely accessed by different drugs and have differentiable phenotypes (Redish et al. 2008).

6.4 Decision-Making Theories and Addiction

We have seen examples of how decision-making models exhibit vulnerabilities to addictive choice. Another important question is how people actually made decisions in the real-world. There is a key aspect of addiction that does not fit easily into current theories of addiction: the high rate of remission. Current theories of addiction generally account for the development and escalation of addiction by supposing that

drugs have a pharmacological action that cumulatively biases the decision-making system of the brain toward drug-choice. These models do not account for cases of spontaneous (untreated) remission, such as a long-term daily drug user who suddenly realizes that she would rather support her children than use drugs, and stops her drug use (Heyman 2009).

Approaches like the 12-step programs (originally Alcoholics Anonymous) have a high success rate in achieving lasting abstinence (Moos and Moos 2004, 2006a, 2006b). These programs use a variety of strategies to encourage people to give up their addictive behavior. These strategies may be amenable to description in the framework of decision-making modeling. For example, one effective strategy is to offer addicts movie rental vouchers in exchange for one week of abstinence (McCaul and Petry 2003; Higgins et al. 2004). If an addict is consistently making decisions that prefer having a gram of cocaine over having $60, why would the addict prefer a movie rental worth $3 over a week of drug taking? This is, as yet, an unanswered question which may require models that include changes in state-space representation, more complex forward-modeling, and more complex evaluation mechanisms than those currently included in computational models of addiction.

References

Ainslie G (1974) Impulse control in pigeons. J Exp Anal Behav 21:485

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

Andersen S, Harrison GW, Lau MI, Rutström EE (2008) Eliciting risk and time preferences. Econometrica 76:583

Aragona BJ, Cleaveland NA, Stuber GD, Day JJ, Carelli RM, Wightman RM (2008) Preferential enhancement of dopamine transmission within the nucleus accumbens shell by cocaine is attributable to a direct increase in phasic dopamine release events. J Neurosci 28:8821

Arbisi PA, Billington CJ, Levine AS (1999) The effect of naltrexone on taste detection and recognition threshold. Appetite 32:241

Balleine BW (2001) Incentive processes in instrumental conditioning. In: Handbook of contemporary Learning Theories, p 307

Balleine BW (2004) Incentive behavior. In: The behavior of the laboratory rat: a handbook with tests, p 436

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

Balleine BW, Daw ND, O'Doherty JP (2008) Multiple forms of value learning and the function of dopamine. In: Neuroeconomics: decision making and the brain, p 367

Barnes CA (1979) Memory deficits associated with senscence: A neurophysiological and behavioral study in the rat. J Comp Physiol Psychol 93:74

Barto AG (1994) Adaptive critics and the basal ganglia. In: Models of information processing in the basal ganglia, p 215

Baum W, Rachlin H (1969) Choice as time allocation. J Exp Anal Behav 12:861

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

Becker GS, Murphy KM (1988) A theory of rational addiction. J Polit Econ 96:675

Becker GS, Grossman M, Murphy KM (1994) An empirical analysis of cigarette addiction. Am Econ Rev 84:396

Bernheim BD, Rangel A (2004) Addiction and cue-triggered decision processes. Am Econ Rev 94:1558

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191:391
Berridge CW, Arnsten AF, Foote SL (1993) Noradrenergic modulation of cognitive function: clinical implications of anatomical, electrophysiological and behavioural studies in animal models. Psychol Med 23:557
Bickel WK, Odum AL, Madden GJ (1999) Impulsivity and cigarette smoking: delay discounting in current, never, and ex-smokers. Psychopharmacology (Berlin) 146:447
Bouton ME (2002) Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. Biol Psychiatry 52:976
Bouton ME, Swartzentruber D (1989) Slow reacquisition following extinction: context, encoding, and retrieval mechanisms. J Exp Psychol, Anim Behav Processes 15:43
Bouton ME, Westbrook RF, Corcoran KA, Maren S (2006) Contextual and temporal modulation of extinction: behavioral and biological mechanisms. Biol Psychiatry 60:352
Breland K, Breland M (1961) The misbehavior of organisms. Am Psychol 16:682
Burks SV, Carpenter JP, Goette L, Rustichini A (2009) Cognitive skills affect economic preferences, strategic behavior, and job attachment. Proc Natl Acad Sci 106:7745
Childress AR, Ehrman R, Rohsenow DJ, Robbins SJ, O'Brien CP (1992) Classically conditioned factors in drug dependence. In: Substance abuse: a comprehensive textbook, p 56
Christensen CJ, Silberberg A, Hursh SR, Roma PG, Riley AL (2008) Demand for cocaine and food over time. Pharmacol Biochem Behav 91:209
Chung SH, Herrnstein RJ (1967) Choice and delay of reinforcement. J Exp Anal Behav 10:67
Corbit LH, Balleine BW (2000) The role of the hippocampus in instrumental conditioning. J Neurosci 20:4233
Cote D, Caron A, Aubert J, Desrochers V, Ladouceur R (2003) Near wins prolong gambling on a video lottery terminal. J Gambl Stud 19:433
Courville AC (2006) A latent cause theory of classical conditioning. Doctoral dissertation, Carnegie Mellon University
Custer RL (1984) Profile of the pathological gambler. J Clin Psychiatry 45:35
Daw ND (2003) Reinforcement learning models of the dopamine system and their behavioral implications. Doctoral dissertation, Carnegie Mellon University
Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199
Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. Neural Netw 15:603
Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704
Daw ND, Courville AC, Touretzky DS (2006) Representation and timing in theories of the dopamine system. Neural Comput 18:1637
Dayan P (2002) Motivated reinforcement learning. Advances in neural information processing systems: proceedings of the 2002 conference
Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285
Dayan P, Seymour B (2008) Values and actions in aversion. In: Neuroeconomics: decision making and the brain, p 175
Dayan P, Niv Y, Seymour B, Daw ND (2006) The misbehavior of value and the discipline of the will. Neural Netw 19:1153
Dezfouli A, Piray P, Keramati MM, Ekhtiari H, Lucas C, Mokri A (2009) A neurocomputational model for cocaine addiction. Neural Comput 21:2869
di Chiara G (1999) Drug addiction as dopamine-dependent associative learning disorder. Eur J Pharmacol 375:13
Dickerson M, O'Connor J (2006) Gambling as an addictive behavior. Cambridge University Press, Cambridge
Domjan M (1998) The principles of learning and behavior. Brooks/Cole
Doya K (2000) Metalearning, neuromodulation, and emotion. In: Affective minds, p 101
Elster J (1999) Gambling and addiction. In: Getting hooked: rationality and addiction, p 208

967 Estes WK (1943) Discriminative conditioning. I. A discriminative property of conditioned anticipation. J Exp Psychol 32:150
968
969 Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. Nat Neurosci 11:966
970
971 Flagel SB, Watson SJ, Akil H, Robinson TE (2008) Individual differences in the attribution of incentive salience to a reward-related cue: Influence on cocaine sensitization. Behav Brain Res 186:48
972
973 Frederick S, Loewenstein G, O'Donoghue T (2002) Time Discounting and time preference: A critical review. J Econ Lit 40:351
974
975 Fuhs MC, Touretzky DS (2007) Context learning in the rodent hippocampus. Neural Comput 19:3172
976 Gershman SJ, Blei DM, Niv Y (2010) Context, learning, and extinction. Psychol Rev 117:197
977 Glimcher PW, Camerer C, Fehr E, Poldrack RA (2008) Neuroeconomics: decision making and the brain. Elsevier/Academic Press, London
978
979 Goldman MS, Brown SA, Christiansen BA (1987) Expectancy theory: thinking about drinking. In: Psychological theories of drinking and alcoholism, p 181
980 Goldstein A (2000) Addiction: from biology to drug policy. Oxford University Press, Oxford
981 Grossman M, Chaloupka FJ (1998) The demand for cocaine by young adults: a rational addiction approach. J Health Econ 17:427
982
983 Gul F, Pesendorfer W (2001) Temptation and self-control. Econometrica 69:1403
984 Gutkin BS, Dehaene S, Changeux JP (2006) A neurocomputational hypothesis for nicotine addiction. Proc Natl Acad Sci USA 103:1106
985
986 Henly SE, Ostdiek A, Blackwell E, Knutie S, Dunlap AS, Stephens DW (2008) The discounting-by-interruptions hypothesis: model and experiment. Behav Ecol 19:154
987 Hershberger WA (1986) An approach through the looking-glass. Anim Learn Behav 14:443
988 Heyman GM (2009) Addiction: a disorder of choice. Harvard University Press, Cambridge
989 Higgins ST, Heil SH, Lussier JP (2004) Clinical implications of reinforcement as a determinant of substance use disorders. Annu Rev Psychol 55:431
990
991 Hirsh R (1974) The hippocampus and contextual retrieval of information from memory: A theory. Behav Biol 12:421
992 Hirsh R, Leber B, Gillman K (1978) Fornix fibers and motivational states as controllers of behavior: A study stimulated by the contextual retrieval theory. Behav Biol 22:463
993
994 Hu D, Amsel A (1995) A Simple Test of the Vicarious Trial-and-Error Hypothesis of Hippocampal Function. Proc Natl Acad Sci USA 92:5506
995
996 Hu D, Xu X, Gonzalez-Lima F (2006) Vicarious trial-and-error behavior and hippocampal cytochrome oxidase activity during Y-maze discrimination learning in the rat. Int J Neurosci 116:265
997
998 Hunt WA (1998) Pharmacology of alcohol. In: Tarter RE, Ammerman RT, Ott PJ (eds) Handbook of substance abuse: Neurobehavioral pharmacology. Plenum, New York, pp 7–22
999
1000 Isaacson RL (1974) The limbic system. Plenum, New York
1001 Isoda M, Hikosaka O (2008) Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. J Neurosci 28:7209
1002 Jaffe JH, Cascella NG, Kumor KM, Sherer MA (1989) Cocaine-induced cocaine craving. Psychopharmacology (Berlin) 97:59
1003
1004 Jaffe A, Gitisetan S, Tarash I, Pham AZ, Jentsch JD (2010) Are nicotine-related cues susceptible to the blocking effect? Society for Neuroscience Abstracts, Program Number 268.4
1005
1006 Johnson A, Redish AD (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J Neurosci 27:12176
1007 Jones BT, Corbin W, Fromme K (2001) A review of expectancy theory and alcohol consumption. Addiction 96:57
1008
1009 Kamin LJ (1969) Predictability, surprise, attention, and conditioning. In: Learning in animals, p 279
1010
1011 Kirby KN, Herrnstein RJ (1995) Preference reversals due to myopic discounting of delayed reward. Psychol Sci 6:83
1012

Kruse JM, Overmier JB, Konz WA, Rokke E (1983) Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. Learn Motiv 14:165
Kuhar MJ, Ritz MC, Sharkey J (1988) Cocaine receptors on dopamine transporters mediate cocaine-reinforced behavior. In: Mechanisms of cocaine abuse and toxicity, p 14
Kurth-Nelson Z, Redish AD (2009) Temporal-difference reinforcement learning with distributed representations. PLoS ONE 4:e7362
Kurth-Nelson Z, Redish AD (2010) A reinforcement learning model of precommitment in decision making. Frontiers Behav Neurosci 4:184
Langer EJ, Roth J (1975) Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. J Pers Soc Psychol 32:951
Lebron K, Milad MR, Quirk GJ (2004) Delayed recall of fear extinction in rats with lesions of ventral medial prefrontal cortex. Learn Mem 11:544
Lenoir M, Serre F, Cantin L, Ahmed SH (2007) Intense sweetness surpasses cocaine reward. PLoS ONE 2:e698
Levine AS, Billington CJ (2004) Opioids as agents of reward-related feeding: a consideration of the evidence. Physiol Behav 82:57
Liao D, Lin H, Law PY, Loh HH (2005) Mu-opioid receptors modulate the stability of dendritic spines. Proc Natl Acad Sci USA 102:1725
Liu J-, Liu J-, Hammit JK, Chou S- (1999) The price elasticity of opium in Taiwan, 1914–1942. J Health Econ 18:795
Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol 67:145
Lovibond PF (1983) Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. J Exp Psychol Anim Behav Process 9:225
Mackintosh NJ (1974) The psychology of animal learning. Academic Press, San Diego
Madden GJ, Bickel WK (2010) Impulsivity: the behavioral and neurological science of discounting. American Psychological Association, Washington, DC
Mazur J (1987) An adjusting procedure for studying delayed reinforcement. In: Quantitative analyses of behavior, p 55
McCaul ME, Petry NM (2003) The role of psychosocial treatments in pharmacotherapy for alcoholism. Am J Addict 12:S41
McFarland K, Kalivas PW (2001) The circuitry mediating cocaine-induced reinstatement of drug-seeking behavior. J Neurosci 21:8655
Milad MR, Vidal-Gonzalez I, Quirk GJ (2004) Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. Behav Neurosci 118:389
Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936
Moos RH, Moos BS (2004) Long-term influence of duration and frequency of participation in alcoholics anonymous on individuals with alcohol use disorders. J Consult Clin Psychol 72:81
Moos RH, Moos BS (2006a) Participation in treatment and Alcoholics Anonymous: a 16-year follow-up of initially untreated individuals. J Clin Psychol 62:735
Moos RH, Moos BS (2006b) Rates and predictors of relapse after natural and treated remission from alcohol use disorders. Addiction 101:212
Muenzinger KF (1938) Vicarious trial and error at a point of choice. I. A general survey of its relation to learning efficiency. J Genet Psychol 53:75
Nadel L, Willner J (1980) Context and conditioning: A place for space. Physiol Psychol 8:218
Nestler EJ (1996) Under siege: The brain on opiates. Neuron 16:897
Niv Y, Montague PR (2008) Theoretical and empirical studies of learning. In: Neuroeconomics: decision making and the brain, p 331
Niv Y, Daw ND, Dayan P (2006) Choice values. Nat Neurosci 9:987
O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452
O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. Brain Res 34:171

O'Keefe J, Nadel L (1978) The hippocampus as a cognitive map. Clarendon, Oxford
Oscar-Berman M, Marinkovic K (2003) Alcoholism and the brain: an overview. Alcohol Res Health 27(2):125–134
Ostlund SB, Balleine BW (2008) The disunity of Pavlovian and instrumental values. Behav Brain Sci 31:456
Packard MG, McGaugh JL (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiol Learn Mem 65:65
Paine TA, Dringenberg HC, Olmstead MC (2003) Effects of chronic cocaine on impulsivity: relation to cortical serotonin mechanisms. Behav Brain Res 147:135
Panlilio LV, Thorndike EB, Schindler CW (2007) Blocking of conditioning to a cocaine-paired stimulus: Testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. Pharmacol Biochem Behav 86:774
Parke J, Griffiths M (2004) Gambling addiction and the evolution of the near miss. Addict Res Theory 12:407
Pavlov I (1927) Conditioned reflexes. Oxford Univ Press, Oxford
Phillips PEM, Stuber GD, Heien MLAV, Wightman RM, Carelli RM (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422:614
Porrino LJ, Lyons D, Smith HR, Daunais JB, Nader MA (2004) Cocaine self-administration produces a progressive involvement of limbic, association, and sensorimotor striatal domains. J Neurosci 24:3554
Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51:381
Quirk GJ, Garcia R, González-Lima F (2006) Prefrontal mechanisms in extinction of conditioned fear. Biol Psychiatry 60:337
Rachlin H (2000) The science of self-control. Harvard University Press, Cambridge
Rachlin H, Green L (1972) Commitment, choice, and self-control. J Exp Anal Behav 17:15
Redish AD (1999) Beyond the cognitive map: from place cells to episodic memory. MIT Press, Cambridge
Redish AD (2004) Addiction as a computational process gone awry. Science 306:1944
Redish AD (2009) Implications of the multiple-vulnerabilities theory of addiction for craving and relapse. Addiction 104:1940
Redish AD, Johnson A (2007) A computational model of craving and obsession. Ann NY Acad Sci 1104:324
Redish AD, Kurth-Nelson Z (2010) Neural models of temporal discounting. In: Impulsivity: the behavioral and neurological science of discounting, p 123
Redish AD, Jensen S, Johnson A, Kurth-Nelson Z (2007) Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol Rev 114:784
Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31:415
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II, p 64
Restle F (1957) Discrimination of cues in mazes: A resolution of the 'place-vs-response' question. Psychol Rev 64:217
Reynolds B, Ortengren A, Richards JB, de Wit H (2006) Dimensions of impulsive behavior: personality and behavioral measures. Pers Individ Differ 40:305
Ritz MC, Lamb RJ, Goldberg SR, Kuhar MJ (1987) Cocaine receptors on dopamine transporters are related to self-administration of cocaine. Science 237:1219
Robinson TE, Berridge KC (1993) The neural basis of drug craving: An incentive-sensitization theory of addiction. Brains Res Rev 18:247
Robinson TE, Berridge KC (2001) Mechanisms of action of addictive stimuli: Incentive-sensitization and addiction. Addiction 96:103
Robinson TE, Berridge KC (2003) Addiction. Annu Rev Psychol 54:25
Robinson TE, Berridge KC (2004) Incentive-sensitization and drug 'wanting'. Psychopharmacology 171:352

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241
Schultz W, Dayan P, Montague R (1997) A neural substrate of prediction and reward. Science 275:1593
Schweighofer N, Shishida K, Han CE, Yamawaki S, Doya K (2006) Humans can adopt optimal discounting strategy under real-time constraints. PLoS Comput Biol 2:e152
Schweighofer N, Tanaka SC, Doya K (2007) Serotonin and the evaluation of future rewards. Theory, experiments, and possible neural mechanisms. Ann NY Acad Sci 1104:289
Si J, Barto AG, Powell WB, Wunsch D (2004) Handbook of learning and approximate dynamic programming. Wiley/IEEE Press, New York
Simon NW, Mendez IA, Setlow B (2007) Cocaine exposure causes long-term increases in impulsive choice. Behav Neurosci 121:543
Smith A, Li M, Becker S, Kapur S (2006) Dopamine, prediction error and associative learning: a model-based account. Network: Comput Neural Syst 17:61
Sotres-Bayon F, Cain CK, LeDoux JE (2006) Brain mechanisms of fear extinction: historical perspectives on the contribution of prefrontal cortex. Biol Psychiatry 60:329
Sozou PD (1998) On hyperbolic discounting and uncertain hazard rates. R Soc Lond B 265:2015
Stahl SM, Pradko JF, Haight BR, Modell JG, Rockett CB, Learned-Coughlin S (2004) A review of the neuropharmacology of bupropion, a dual norepinephrine and dopamine reuptake inhibitor. Prim Care Companion J Clin Psychiat 6:159
Strotz RH (1956) Myopia and inconsistency in dynamic utility maximization. Rev Econ Stud 23:165
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
Talmi D, Seymour B, Dayan P, Dolan RJ (2008) Human Pavlovian instrumental transfer. J Neurosci 28:360
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 7:887
Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, Doya K (2007) Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. PLoS ONE 2:e1333
Tolman EC (1938) The determiners of behavior at a choice point. Psychol Rev 45:1
Tolman EC (1939) Prediction of vicarious trial and error by means of the schematic sowbug. Psychol Rev 46:318
Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55:189
Tsai HC, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K (2009) Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science 324:1080
Uslaner JM, Acerbo MJ, Jones SA, Robinson TE (2006) The attribution of incentive salience to a stimulus that signals an intravenous injection of cocaine. Behav Brain Res 169:320
van der Meer MA, Redish AD (2009) Covert expectation-of-reward in rat ventral striatum at decision points. Frontiers Integr Neurosci 3:1
van der Meer MA, Redish AD (2010) Expectancies in decision making, reinforcement learning, and ventral striatum. Front Neurosci 4:29
Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43
Wagenaar WA (1988) Paradoxes of gambling behavior. Erlbaum, London
Weiner I, Lubow RE, Feldon J (1988) Disruption of latent inhibition by acute administration of low doses of amphetamine. Pharmacol Biochem Behav 30:871
White AM (2003) What happened? Alcohol, memory blackouts, and the brain. Alcohol Res Health 27(2):186–196
Yin HH, Knowlton B, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181
Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681