2021 Computational Psychiatry: A Primer, edited by P. Series. MIT Press.

9 Addiction from a Computational Perspective

A. David Redish
University of Minnesota

9.1 Introduction: What Is Addiction?

Everyone knows what addiction is. We all know people whose lives have been ruined by drugs, and we all have behaviors that we wish we could stop, but don't. However, the definition of addiction remains elusive. Early definitions related to a "lack of will" and suggested addiction was a moral failing. However, this theory did not lead to reliable treatments and left many incapable of ending their addictions. Later definitions defined addiction as a disease and suggested that behavioral and chemical treatments could alleviate it. In particular, these disease-related theories suggested that many drug addictions arose from biological responses to chemical imbalances that could be treated pharmacologically. Some of these pharmacological treatments, such as methadone treatment for heroin addictions (Meyer and Mirin 1979) and the nicotine patch for smoking (Hanson et al. 2003), have been very successful, but other addictions (stimulants, alcohol) have been much more difficult to treat pharmacologically. Furthermore, pharmacological definitions do not include the possibility of nonchemical addictions, such as gambling, which is now seen as an addiction-like problem.

Current definitions of addiction are based on conceptualizations of addiction as a problem with decision-making systems (Heyman 2009; Redish 2013), often evidenced as continued use despite stated preferences (Goldstein 2001; Ainslie 2001) and as continued use despite high cost (Robinson and Berridge 2003; Koob and Le Moal 2006). The most recent models identify addiction as arising from vulnerabilities leading to failure modes in decision-making algorithms (Redish, Jensen, and Johnson 2008).

One of the most common popular descriptions of addiction lies in the addict's continued use despite making explicit statements of a desire to stop. Current theories of decision making reject the hypothesis of the unitary decision-maker—each individual is actually a multiplicity of decision-making systems (algorithms, processes) competing

for behavioral control (O'Keefe and Nadel 1978; Daw et al. 2005; Rangel, Camerer, and Montague 2008; Redish et al. 2008; Kahneman 2011; van der Meer, Kurth-Nelson, and Redish 2012; Redish 2013). While this theory provides an explanation for this conflict (Kurzban 2010), computational models of addiction have not emphasized this conflict because it is hard to study in nonlinguistic animals (i.e., nonhumans), while human rights limitations make it difficult to do controlled studies of addiction in humans. Nevertheless, the study of decision-making systems and their interaction is well established in both human and nonhuman animals and has been used computationally to guide treatment.

One of the classic descriptions of addiction is based on the observation that addicts will continue to use even in the face of high costs. This can be quantified through the economic concept of elasticity as a measure of how much one's willingness to buy something changes by its cost (Bickel, DeGrandpre, and Higgins 1993; Hursh 2005). Things that diminish slowly by cost are inelastic. Researchers have suggested that drugs are fundamentally inelastic: as costs increase, the number of rewards paid for decrease less than they should. Of course, there are many things that are inelastic that are not considered addictive—oxygen, for example (where the withdrawal symptoms are particularly traumatic), but also some behaviors continued even in the face of high costs are celebrated, such as Kerri Strug's 1996 Olympic vault performed on a sprained ankle, or Osip Mandelstam continuing to write poetry even after Stalin had thrown him in the gulag for it.

A key to the question of addiction is to separate the science of why an agent continues its behavior from the decision to treat and change that behavior. This conceptualization parallels Jerome Wakefield's conceptualization of psychiatry as depending on harmful dysfunction (Wakefield 1992). "Dysfunction" reflects a system not working as it was intended to. For example, mu-opioid activation signals pleasure in mammalian brains (Berridge and Robinson 2003). These receptors were certainly not evolved to respond to heroin, but they do. "Harmful" reflects a society's choice of what to change. For example, American society is currently transitioning from treating marijuana as so dangerous as to be illegal with severe penalties to something that can better be handled under legal regulation. Things can be harmful without being dysfunctional, such as tribal wars, which are extremely harmful, but likely reflect the natural evolution of human behavior (Turchin 2003; Diamond 2006), and dysfunctional without being harmful, such as synesthesia (Cytowic 1998).

Computational models of addiction are aimed at understanding the science of why an agent continues its behavior and the science of how one could change that behavior if one so desired. Importantly, the decision of whether to change that behavior has not

been computationally assessed. Such a decision would depend on sociological models, which are not the focus of this chapter. Instead, this chapter will focus on computational approaches to addictive behavior and its modification.

9.2 Past Approaches

Past computational approaches to addiction can be divided into three broad categories: economic models, in which drugs are seen as economic objects that have feedback properties that make them overvalued; homeostatic models, in which drugs change intrinsic biological properties and shift allostatic set-points, which subsequently require drugs to reach that set-point; and reinforcement learning models, in which drugs hijack learning algorithms to produce aberrant learning. Current views on addiction suggest that these three hypotheses are all failure modes of decision-making systems, and that there are many endophenotypes of drug addiction.

9.2.1 Economic Models

Although popular descriptions of drug use (e.g., *Reefer Madness* [Gasnier 1936], *Long Day's Journey into Night* [O'Neill 1956], *The Lost Weekend* [Wilder and Brackett 1945], *Sid and Nancy* [Cox 1986]) have suggested that drugs are overwhelming and addicts would go to any cost to achieve drug-taking, experimental studies have long suggested that drugs are economic objects and that drug use decreases with increasing costs (Bickel et al. 1993; Liu et al. 1999; Grossman and Chaloupka 1998; Hursh 2005). The first economic model of drug use is Becker and Murphy's (1988) "Rational Addiction" model, which is an economic utility model in which subjects are assumed to select the most cost-effective choice with the highest value. Drugs are assumed to have a positive feedback, so that the more one takes those drugs, the more valuable they become. Becker and Murphy show that under these assumptions, a hypothetical user could be shown to become addicted when the positive feedback overwhelms the negative consequences of the drug use.

These models led to quantitative analyses of drug use, asking direct questions of the economic demand curves of drug use. Demand curves are quantitative measures of elasticity. This can be measured either through effort (how many lever presses will a nonhuman animal push for reward?) or through monetary costs (how many grams of drug will you buy?). In a typical demand curve (figure 9.1), there is an inelastic portion, where increases in cost have little effect on number of rewards bought, and an elastic portion, where the number of rewards bought falls off very quickly. These are separated by an inflection point (*pMax*). Addicts can be defined as people for whom

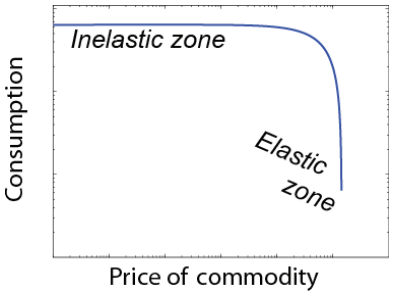

Figure 9.1
The shape of a typical demand curve. As the price of the commodity increases, the number of samples consumed decreases. There is typically an inelastic zone, where large ratio changes have little effect, and an elastic zone, where large ratio changes have a larger effect. Note that both axes are logarithmic. Compare this, for example, to real demand curves as seen in Bruner and Johnson (2013), where subjects were asked how much cocaine they would buy at a hypothetical given price.

this inflection point has shifted far to the right, but nevertheless, their demand curves do have this typical, canonical shape.

A key insight from this economic perspective on drug use is that drugs provide fast rewards and slow consequences. All animals (human and nonhuman) discount future rewards, valuing rewards more if they are delivered in a shorter time frame (Ainslie 1992; Madden and Bickel 2010). Economically, this makes sense, since immediate rewards can be invested, and consequences can prevent the use of later rewards. Importantly, as described in section 5.2, all animals (human and nonhuman) show nonexponential discounting curves (figure 9.2), which means that preferences can cross—thus, it is possible both to prefer to smoke the cigarette in your hand and to prefer to not smoke in the future. (Of course, when the future becomes now, one will want to smoke the cigarette now again.) Addicts show particularly fast discounting functions, which can exacerbate this problem (Bickel and Marsch 2001). There is some evidence that successful treatment modifies these discounting rates in subjects with particularly fast discounting functions (Bickel et al. 2014) and that these discounting rates are predictive of relapse (Sheffer et al. 2014). It is possible to modify discounting rates, guiding the subject's attention to delayed rewards by providing episodic cues about the delayed rewards to make those delayed rewards more concrete (Peters and Büchel 2010). Recent evidence has suggested that these changes can reduce drug use (Stein et al. 2018; Snider et al. 2018). However, whether these changes are due to changes in discounting rates per se or to changes in interacting multiple decision systems remains an open question.

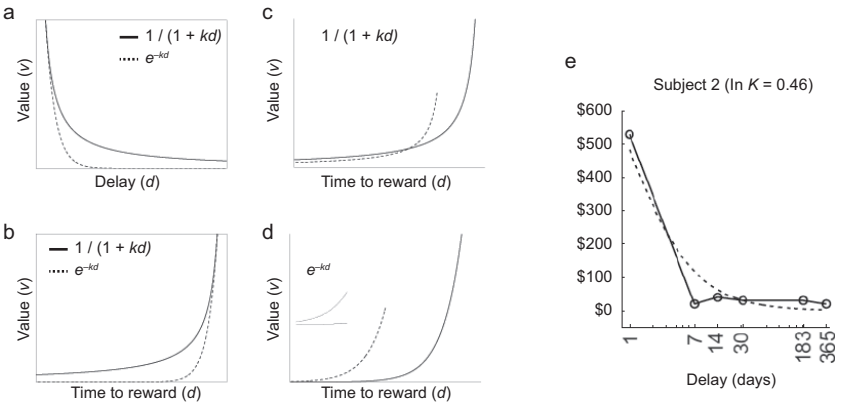

Figure 9.2
Delay discounting. (a) Delay discounting entails a loss of value as a function of delay to an event. Two discounting functions are typically used, hyperbolic $[V = r/(1+kd)]$, and exponential $[e^{-kd}]$, where d is the delay to the event and k is a parameterization factor. (b) Logically, this can be understood in terms of value of an expected event as one approaches the event in time. (c) Hyperbolic discounting functions can create a preference reversal where one prefers one option (the larger, later, solid line) to another (the smaller, sooner, dashed line) that reverses as one approaches the options in time. (d) Exponential discounting, however, does not reverse, even when both options are far away (see inset), showing an expansion of the far-left edge of the graph. (e) A real discounting curve from an individual, reprinted from Kurth-Nelson and Redish (2009).

While the basic economic story that drugs are economic objects that are discounted quickly is clearly correct, drug use is context-sensitive in ways that make these simple economic descriptions incomplete (Bernheim and Rangel 2004). We will return to the question of these economics later, when we come to the interacting multiple systems models below.

9.2.2 Homeostatic Models

All drugs that are reliably self-administered, either by humans or other animals, are pharmacologically similar in some way to endogenous chemicals used in neural processing (Koob and Le Moal 2006). For example, active opioids such as morphine, heroin, or oxycodone activate the mu-opiate receptor; cocaine blocks dopamine reuptake in the synapse, which increases dopamine in the synapse; amphetamine encourages release of dopaminergic vesicles; and nicotine activates acetylcholine receptors. Biological systems in general and neural systems in particular are very sensitive to levels of these endogenous chemicals and have extensive negative feedback processes (such as

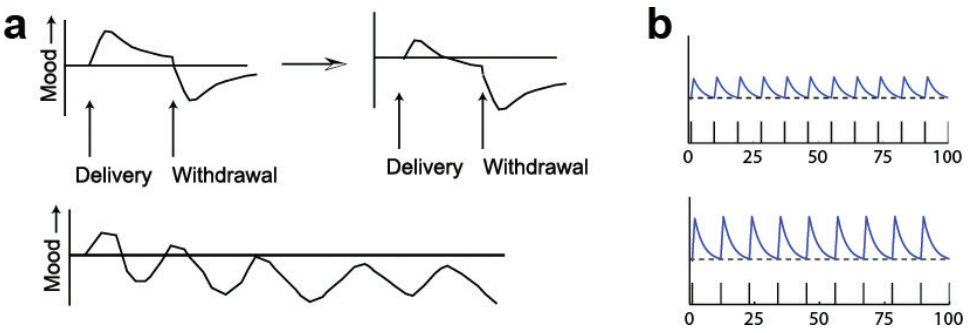

Figure 9.3
Homeostatic/allostatic processes. (a) Drug delivery produces a positive reaction state, which then adapts and collapses to a negative state when the drug is removed. Over time, the user is hypothesized to adapt to the positive state, producing a shift in the allostatic set-point toward the negative state. Redrawn after Koob (2013). (b) Tsibulsky and Norman (1999) and Keramati et al. (2017) modeled self-administration as an attempt to maintain the total level of drug at a given set-point. As the drug was processed internally and reduced beyond the set-point, the animal was hypothesized to seek the drug through lever-pressing. This model explains the different rates of lever-pressing as a function of the drug dose. Redrawn after Tsibulsky and Norman (1999) and Keramati et al. (2017).

trafficking of receptors in and out of the synaptic membrane) that keep the sensitivity balanced. In situations where receptors are flooded, they will normalize their levels, requiring more activation to produce the same effects.

For example, many self-administration experiments (in which animals are trained to press a lever for drug reward) can be described quantitatively in terms of maintenance of pharmacological levels of the drug (Tsibulsky and Norman 1999, Keramati et al. 2017). Negative feedback processes driving maintenance interact with the positive feedback processes of drug utility (as suggested by Becker and Murphy) to produce dramatic differences in valuation between drugs and nondrug rewards (with drugs being valued much higher than nondrugs, leading to overtaking of drug rewards.)

Three quantitative models based on issues of homeostatic balance are the Tsibulsky and Norman (1999), the Keramati et al. (2017), and the Dezfouli et al. (2009) models (figure 9.3). The Tsibulsky and Norman model explicitly hypothesizes that animals are attempting to maintain a specific level of cocaine, which explains quantitatively the observed shifts in response to changes in the dosages given with each lever press. Keramati et al. notes, however, that there are short-term dynamics when the changes actually occur that are not explicable by a simple set-point hypothesis, particularly

in the transition that occurs with increased access to drug. They therefore add in a learning component based on the reinforcement learning models detailed below. The Dezfouli et al. model is based on a homeostatic expansion of the Redish (2004) model (see below), particularly looking at the effect of homeostatic set-points driving pharmacological effects of dopamine on learning. While the Redish model is based on the temporal-difference, reinforcement learning dopamine-as-delta model of Montague et al. 1996, and is thus a hijacked-learning model, the Dezfouli et al. model is based on the average reward dopamine hypothesis of Daw and Touretzky (2000), and becomes a homeostatic model.

9.2.3 Opponent Process Theory

One of the earliest models of drug use is the opponent process theory of Solomon and Corbit (1974; see Koob and Le Moal 2006 for extensive discussion of this model), in which drugs are assumed to produce a strong positive reward followed by a strong negative recovery. Homeostatic processes are assumed to normalize the excess drug to decrease the positive factors, and increase the negative factors, which leads to increased need for drugs to return the homeostatic process to baseline. These models have been supported by evidence that chronic drug use leads to enhancement of positive-valuation neuron activity in the nucleus accumbens (Kourrich et al. 2007; Volman et al. 2013) and evidence that the emotional crash after drug use is an important factor in driving self-administration (Rothwell, Gewirtz, and Thomas 2010).

While the Solomon and Corbit (1974) and the Koob and Le Moal (2006) models are not quantitative, Gutkin, Dehaene, and Changeux (2006) proposed an opponent process model in which there is habituation of response processes to a continuous delivery of nicotine—a phasic increase at the start and a phasic decrease at the end, and a decrease in the overall tonic dopamine levels. The normalization caused by the assumed decrease in dopamine levels leads to a decrease in ability to learn non–drug-related cues, which leads to an increase in attention to and learning of drug-related cues. Thus, Gutkin et al. show how an opponent process model can hijack learning process by disrupting the difference between learning on and off drug.

9.2.4 Reinforcement Models

The third family of computational models is based on the concept that learning depends on physical processes, and those physical processes can be modulated by external chemicals and other processes. In animal learning theory, the concept of reinforcement is separate from the concept of reward. Reinforcement is any mechanism that makes an agent more likely to return to an action. An external chemical that

increases reinforcement would increase drug-seeking and drug-taking (di Chiara 1999, Redish 2004).

In the 1950s, it was discovered that electrical stimulation of specific neural sites was reinforcing, in that both human and nonhuman animals would activate the stimulation (Olds and Milner 1954), even to the extent of avoiding many other rewards. Interestingly, in humans (who could rate "pleasure" linguistically), these studies found that the most reinforcing stimulations were not always the most pleasant (Heath 1963).

An important breakthrough in the understanding of reinforcement came when Berridge and Robinson directly measured reinforcement and pleasure in nonhuman animals and discovered that they were separable. It was well-known that many drugs of abuse affected dopaminergic functioning and that the stimulation drove dopamine release, and it was thought that dopamine would drive pleasure signals. However, when Berridge and Robinson (2003) directly tested this hypothesis, it was discovered that this was wrong—dopamine and pleasure were dissociable. In their elegant studies, they measured facial expressions of pleasure and disgust in rats under manipulations of dopamine and opiate signals. Dopamine manipulations affected reinforcement but did not affect facial expressions of pleasure. In contrast, manipulations of opiate signals (e.g., mu-opiate and kappa-opiate agonists and antagonists) affected pleasure responses. This led them to hypothesize that drugs that affected dopamine increased the "incentive salience" or "value" of a reward, which drove seeking, independently of the pleasure experienced by that reward.

Around this time, a major breakthrough occurred in the understanding of dopamine function in animal learning—Wolfram Schultz and his team discovered that dopamine cells burst when provided a surprising reward but did not fire when the reward was predicted by a cue (Ljungberg, Apicella, and Schultz 1992). Read Montague and colleagues (1996) realized that this signal was the value-prediction error (VPE)[1] signal δ (delta) that underlay a theory of robotic learning called temporal-difference reinforcement learning (TDRL) that had become very successful in the field of computer science[2] (Sutton and Barto 1998; see also section 2.3).

As described in section 2.3, the TDRL algorithm defines value as the total reward one can expect to achieve given a policy of actions to be taken in given situations. TDRL maintains a representation of the currently believed value for each situation, and then calculates the difference between that remembered value and the observed value. This difference is the value-prediction error, or VPE. Positive VPE occurs anytime a value is better than expected and drives an increased willingness to take an action, while negative VPE occurs anytime a value is worse than expected and drives a decreased willingness to take an action. The concept of VPE is best understood through an example.

Imagine a soda machine. If you put your money in the soda machine and get two sodas out, then you will be more willing to put money in that soda machine next time. (You have positive VPE.) If you put your money in the soda machine and get nothing out, then you will be less willing to put money in that soda machine next time. (You have negative VPE.) And, most importantly, if you put the correct amount of money in the soda machine, get your expected soda out, then you understand how that machine works and you don't need to learn anything about it. (You have zero VPE.) Notice that you still get the pleasure (such as it is) of drinking the soda, but you don't need to change your willingness to put money in that machine. VPE is about learning the value of actions. Computer simulations had shown that VPE would allow an agent to learn to behave in simulated environments (Sutton and Barto 1998). These processes can be expressed in the following equations (see also section 2.3):

$$\begin{aligned} V(S_k) &= \int_t^{\infty} \gamma^{\tau - t} E\left[R(\tau)\right] d\tau \\ \delta(t) &= \gamma^d \left[R(S_l) + V(S_l)\right] - V(S_k) \\ V(S_k) &\leftarrow V(S_k) + \eta\delta \end{aligned} \tag{9.1}$$

Where $V(S_k)$ is the value of state S_k, γ^d is a discounting parameter,[3] reflecting expected value decreases over observed delay d; $R(S_l) + V(S_l)$ is the value achieved on entering state S_l; and $\delta(t)$ is the value-prediction error (the difference between the observed and expected value). By changing the value of state S_k toward the observed value (with learning rate η), $V(S_k)$ will approach the observed value. Theories hypothesized that dopamine signaled the value-prediction error $\delta(t)$.

Redish (2004) proposed that if drugs were providing a dopamine signal pharmacologically, then taking drugs would lead to positive VPE, even if the neural calculation of VPE should have been 0 (figure 9.4). Effectively, Redish's model predicted that the dopamine signal at reward contained two components, one from the calculation of $\delta(t)$, and the other from the pharmacological action of the drug. This meant that even with experience, there would always be a noncompensable VPE signal at the reward, which would increase the predicted value of the reward, driving that value to infinity. (Or with normalization, normalizing all other values to zero.)

$$\delta = \max\left\{\gamma^d \left[R(S_l) + V(S_l)\right] - V(S_k) + D(S_l), D(S_l)\right\} \tag{9.2}$$

where $D(S_l)$ reflects the effect of the pharmacological dopamine from the drug.

In his 2004 paper, Redish used computer simulations to show that this model would lead to developing inelasticity (as in the Becker and Murphy hypothesis) and made several untested predictions. The first prediction was that there would be a double

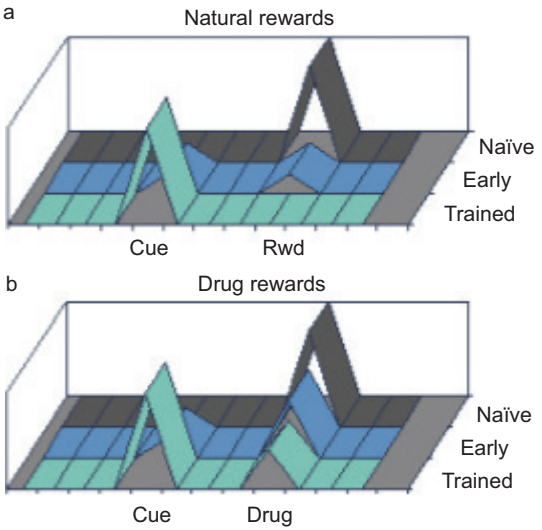

Figure 9.4
The delta signal—dopamine and delta. Diagram of delta (vertical axis) by time (horizontal axis) over three conditions: naïve (untrained), early (with limited training), and trained. (a) With normal rewards, the delta signal shifts from appearing at the unexpected reward to the unexpected cue-that-predicts-reward. (b) In the Redish (2004) model, there are two components in the delta signal, a reward-related component that shifts and a pharmacological component that remains at the reward time. Compare the classic data from Schultz (1998). When the expected reward is not delivered, dopamine cells pause their firing. Aragona et al. (2009) tested the double-bump hypothesis and found that the cue-related signal occurred in accumbens core, while the pharmacological component occurred in shell.

surge of dopamine in drug experiments. In the TDRL theory, $\delta(t)$ first appeared at the time of reward (as it was initially unexpected), and then it shifted to earlier cues that reliably predicted the reward (because the reward was now expected—thus $\delta = 0$, but the cues indicated an unexpected increase in value—thus $\delta > 0$). Similarly, Schultz and colleagues (see Schultz 2002) had found that dopamine shifted from the reward (when unexpected) to the cue (once the animal learned that the cue predicted the reward). In Redish's model, the extra pharmacological component would always appear, even as the dopamine signal appeared at the cue. Since then, this double surge of dopamine has been observed, but as with any theory, reality is more complex than the model, and each component of the double surge occurs separately, with the reward-related surge appearing in accumbens shell and the cue-related surge appearing in accumbens core (Aragona et al. 2009).

The two other key predictions of the Redish 2004 model were (1) that additional drug use would always lead to increased valuation of the drug and (2) that drugs would not show Kamin blocking. As detailed below, these predictions have since been tested and provide insight into the mechanisms of drug addiction.

In the Redish (2004) model, the excess dopamine provides additional value, no matter what. Marks et al. (2010) directly tested this hypothesis in an elegant experiment, where rats were trained to press two levers for a certain dose of cocaine (both levers being equal). One lever was then removed and the other provided smaller doses of cocaine. The Redish (2004) theory predicts that the second lever should gain value, while expectation or homeostatic theories like those discussed earlier would predict that the second lever should lose value (because animals would learn the second lever was providing smaller doses). The Marks et al. data was not consistent with the Redish excess-delta model. However, as noted above, a key factor in drug addiction is that not everyone who takes drugs loses control over their drug use and becomes an addict. Studies of drug use in both human and nonhuman animals suggest that most animals in self-administration experiments continue to show elasticity in drug-taking, stopping in response to high cost, but that a small proportion (interestingly similar to the proportion of humans who become addicted to drugs) become inelastic to drug-taking, being willing to pay excessive costs for their drugs (Anthony, Warner, and Kessler 1994; Hart 2013). One possibility is that the homeostatic models (like that of Tsibulsky and Norman 1999) are a good description of nonaddicted animals, which have a goal of maintaining a satiety level, but that addiction is different.

The Redish (2004) model also predicted that drugs would not show Kamin blocking. Kamin blocking is a phenomenon where animals don't learn that a second cue predicts reward if a first cue already predicts it (Kamin 1969). This phenomenon is well-described by value-prediction error (VPE)—once the animal learns that the first cue predicts the reward, there is no more VPE (because it's predicted!) and the animal does not learn about the second cue (Rescorla and Wagner 1972). Redish noted that because drugs provided dopamine, and dopamine was hypothesized to be that VPE delta signal, then when drugs were the "reward," there was always VPE. Thus, drug outcomes should not show Kamin blocking. The first tests of this, like the Marks et al. study, did not conform to the prediction—animals showed Kamin blocking, even with drug outcomes (Panlilio, Thorndike, and Schindler 2007). However, Jaffe et al. (2014) wondered whether this was related to the subset problem—that only some animals were actually overvaluing the drug. Jaffe et al. tested rats in Kamin blocking for food and nicotine. All rats showed normal Kamin blocking for food. Most rats showed normal Kamin blocking for nicotine. But the subset of rats that were high responders to nicotine did not

show Kamin blocking to nicotine, even though they did to food, exactly as predicted by the Redish model.

9.3 Interacting Multisystem Theories

Studies of decision making in both human and nonhuman animals have, for a long time, found that there are multiple decision-making processes that can drive behavior (O'Keefe and Nadel 1978; Daw et al. 2005; Rangel et al. 2008; Redish et al. 2008; Kahneman 2011; van der Meer et al. 2012; see Redish 2013 for review). These processes are sometimes referred to as different algorithms because they process information differently. They are accessed at different times and in different situations; they depend on different neural systems. How an animal is trained and how a question is asked can change which system drives behavior. Damage to one neural structure or another can shift which system drives behavior.

The key to these different systems lies in how they process information. Decision making can be understood as a consequence of three different kinds of information—what has happened in similar situations in the past (memory), the current situation (perception), and the needs/desires/goals (teleology). How information about each of these aspects is stored can change the selected action—for example, what defines "similar situations" in the past? What parameters of the current situation matter? Are the goals explicitly represented or not? Each system answers these questions differently.

Almost all current decision-making taxonomies differentiate between planning (deliberative) systems and procedural (habit) systems. Planning systems include information about consequences—if I take this action, then I expect to receive that outcome, which can then be evaluated in the context of explicitly encoded needs. Planning systems are slow but flexible. Procedural systems cache those actions—in this situation, this is the best action to take, which is fast but inflexible. As described earlier (sections 2.3 and 5.2), many current computational models refer to planning systems as model-based (because they depend on a model of the consequences in the world), while procedural systems are model-free (which is an unfortunate term because procedural systems still depend on an ability to categorize the current situation, which depends on a model of the world; Redish et al. 2007; Gershman, Blei, and Niv 2010). Some taxonomies also include reflex systems, in which the past experience, the parameters of the current situation that matter, and the action to be taken are all hard-wired within a given organism and are learned genetically over generations. Most taxonomies also include a fourth decision system, variously termed Pavlovian, emotional, affective, or instinctual, in which a species-important action (e.g., salivating, running

away, approaching food) is released as a consequence of a learned perception (context or cue).

The importance of these systems is threefold. (1) How a question is asked can change which system controls behavior; (2) damage to one system can drive behavior to be controlled by another (intact) system; and (3) there are multiple failure modes of each of these systems and their interaction. We will address each of these in turn.

9.3.1 How a Question Is Asked Can Change Which System Controls Behavior

One way to measure how much rats value a reward such as cocaine is to test them in a progressive ratio self-administration experiment (Hodos 1961). In this experiment, the first hit of cocaine costs one lever press, but the second costs two, the third costs four, the fourth eight, and so on. Eventually a rat has to press the lever a thousand times for its hit of cocaine. Measuring when the rat stops pressing the lever indicates the willingness-to-pay and the value of the cocaine to the rat. Not surprisingly, many experiments have found that rats will pay more for cocaine than for other rewards such as saccharine, indicating that cocaine was more valuable than saccharine. However, Serge Ahmed's laboratory found that if those same rats were offered a choice between two levers, one of which provided saccharine while the other provided cocaine, the rats would reliably choose the saccharine lever over the cocaine lever, indicating that saccharine was more valuable than cocaine (Lenoir et al. 2007, see Ahmed 2010). The most logical explanation for this contradiction is that the progressive ratio accesses one decision system (probably procedural) while the choice accesses another (probably deliberative), and that the two systems value cocaine differently. Interestingly, Perry, Westenbroek, and Becker (2013) found that a subset of rats will choose the cocaine, even in the two-option paradigm. These are the same subset of rats that overvalue cocaine in other contexts, such as being willing to cross a shock to get to the cocaine (Deroche-Gamonet 2004). Whether they are also the high responders or whether they no longer show Kamin blocking remains unknown.

9.3.2 Damage to One System Can Drive Behavior to Another

Imagine an animal pressing a lever for an outcome (say, cheese). If the animal is using a planning system to make its decisions, then it is effectively saying, "If I push this lever, I get cheese. Cheese is good. Let's press the lever!" If the animal is using the procedural system, then it is effectively saying, "Pressing the lever is a good thing. Let's press the lever!"—and cheese never enters into the calculation. What this means is that if we make cheese bad (by devaluing it, which we can do by pairing cheese with a nauseating agent like lithium chloride), then rats using planning systems won't press the lever

anymore ("If I push this lever, I get cheese. Yuck!"), but rats using procedural systems will ("Pressing the lever is a good thing. Let's press the lever!"). (See, for example, Niv, Joel, and Dayan 2006 for a model of this dichotomy.) Many experiments have determined that with limited experience, animals are sensitive to devaluation (i.e., they are using a planning system), while with extended experience they are not (i.e., they are using a procedural system), and that lesions to various neural systems can shift this behavior (Killcross and Coutureau 2003; Schoenbaum, Roesch, and Stalnaker 2006). A number of studies have suggested that many drugs (cocaine, amphetamine, alcohol) drive behavior to procedural devaluation-insensitive systems, which has led some theoreticians to argue that drug addiction entails a switch from planning to habit modes (Everitt and Robbins 2005).

Building on the anatomical data known to drive the typical shift from planning to procedural decision systems, Piray et al. (2010) proposed a computational model in which drugs disrupted the planning-valuation systems and accelerated learning in the procedural-valuation systems. This model suggested that known changes in dopaminergic function in the nucleus accumbens as a consequence of chronic drug use could lead to overly fast learning of habit behaviors in the dorsal striatum and would produce a shift from planning to habit systems due to changes in valuation between the two systems.

9.3.3 There Are Multiple Failure Modes of Each of These Systems and Their Interaction

However, rats and humans will take drugs even when they plan. A drug addict who robs a convenience store to get money to buy drugs is not using a well-practiced procedural learning system. A teenager who starts smoking because he (incorrectly) thinks it will make him look cool and make him attractive to girls is making a mistake about outcomes and taking drugs because of an error in the planning system (the error is in his understanding of the structure of the world.)

Some researchers have argued that craving depends on the ability to plan, because craving is transitive (one always craves *something*), and thus it must be depend on expectations and a model-based process (Tiffany 1999; Redish and Johnson 2007). In fact, there are many ways that these different decision systems could drive drug-seeking and drug-taking (Redish et al. 2008). Some of those processes would depend on expectations (i.e., would be model-based, and depend on planning) and explicit representations of outcomes, and could involve craving, while other processes would not (i.e., would be model-free, depending, for example, on habit systems). (An important consequence of this is the observation that seems to get rediscovered every decade or so

that craving and relapse are dissociable—you can crave without relapsing and you can relapse without craving.)

In 2008, Redish and colleagues surveyed the theories of addiction and found that all theories of addiction could be restated in terms of different failure modes of this multi-algorithm decision-making system. An agent that succumbed to overproduction of dopamine signals (Redish 2004) from drug delivery would overvalue drugs and would make economic mistakes to take those drugs. An agent that switched decision systems to habit faster under drugs (Everitt and Robbins 2005; Piray et al. 2010) would become inflexible in response to drug offerings and take drugs even while knowing better. An agent with incorrect expectations ("smoking makes you cool," "I won't get cancer") would make planning mistakes and take drugs in incorrect situations. An agent who discounted the future ("I don't care what happens tomorrow, I want my pleasure today") would be more likely to take drugs than an agent who included future consequences in its plans (Bickel and Marsch 2001). All of these are different examples of vulnerabilities within the decision-making algorithms. Redish et al. (2008) proposed that drug addiction was a symptom, not a disease—that there were many potential causes that could drive an agent to return to drug use, and that efficacious treatment would depend on which causes were active within any given individual.

9.4 Implications

9.4.1 Drug Use and Addiction Are Different Things

At this point, the evidence that a subset of subjects have runaway valuations in response to drugs is overwhelming (Anthony et al. 1994; Deroche-Gamonet 2004, Koob and Le Moal 2006; Hart 2013; Perry et al. 2013; Jaffe et al. 2014). This is true both of animal models of drug addiction and humans self-administering drugs. This suggests a very important point, which is that drug use and addiction are different things. If we, as a society, want to address the health and sociological harm that drugs cause, then we may want to tackle drug use rather than addiction, which would require sociological changes (Hart 2013). As noted above, these sociological models are beyond the scope of this chapter, which is addressing computational models of addiction.

9.4.2 Failure Modes

This chapter has discussed three families of models. The first family was *economic models*, which simply define addiction as inelasticity, particularly due to misvaluations. However, these models do not identify what would cause that misvaluation. The second family was *pharmacological models*, which define addiction as a shift in a

pharmacological set-point that drives value in an attempt to return the pharmacological levels back to that set-point. The third family was *learning and memory models*, which suggest that addiction derives from vulnerabilities in the neural implementations of these algorithms, which drives errors in action-selection.

The multiple-failure-modes model suggests that all three families provide important insights into addiction. It suggests that there are multiple potential vulnerabilities that could drive drug use (which could lie in pharmacological changes in set-points or in many potential failure modes of these learning systems). The multiple vulnerabilities model suggests that addiction is a symptom, not a disease. Many failure modes can create addiction. Importantly, identifying which failure modes occur within any given individual would require specially designed probe tests; this model suggests that it would not be enough to merely identify extended drug use. In fact, these failure modes are likely to depend on specific interactions between the drug and the individual and the specific decision processes driving the drug-seeking/drug-taking behavior.

9.4.3 Behavioral Addictions

If addictions are due to failure modes within neural implementations of decision-making algorithms, then addiction does not require pharmacological effects (even if pharmacological effects can cause addictions), and it becomes possible to define behavioral problems as addictions. For example, problem gambling is now considered an addiction, and other behaviors (such as internet gaming, porn, or even shopping) are now being considered as possible addictions. As noted at the beginning of the chapter, the definition of addiction is difficult. Nevertheless, computational models of addiction have provided insight into problem gambling and behavioral change in general, whether we call those behaviors addictive or not.

Classic computational models of problem gambling have been based on the certainty and uncertainty of reward delivery, but these models have been unable to explain observed properties of gamblers, such as that gamblers tend to have had a large win in their past (Custer 1984; Wagenaar 1988), that they are notoriously superstitious about their gambling (Griffiths 1994), or that they often show hindsight bias (in which they "explain away losses"; Parke and Griffiths 2004), or the illusion of control (in which they believe they can control random effects; Langer 1975).

Redish and colleagues (2007) noted that most models of decision making were based on learning value functions over worlds in which the potential states were already defined. Furthermore, they noted that most animal learning experiments took place in cue-poor environments, where the question the animal faced was "*What is the*

consequence of this cue?" However, most lives (both human and nonhuman) are lived in cue-rich environments, in which the repeated structure of the world is not given to the subject. Instead, subjects have to identify which cues are critical to the definition of the situation the subjects finds themselves in. Redish and colleagues (2007) noted that this becomes a categorization problem and had been well studied in computational models of perception. Attaching a perceptual categorization process based on competitive-learning models (Hertz et al. 1991) to a reinforcement learning algorithm, Redish and colleagues built a model in which the tonic levels of dopamine [i.e., longer-term averages of $\delta(t)$] controlled the stability of the situation-categorization process. This identified two important vulnerabilities in the system depending on over- and under-categorization, particularly in different responses to wins and losses. In their model, wins produced learning of value, while losses produced recategorizations of situations. Their simulated agents were particularly susceptible to near-misses and surprising wins, leading to models of hindsight bias and the illusion of control.

In general, these multi-system models suggest that addiction is a question of harmful dysfunction—dysfunction (vulnerabilities leading to active failure modes) within a system that causes sufficient harm to suggest we need to treat it. They permit both behavioral and pharmacological drivers of addiction.

9.4.4 Using the Multisystem Model to Treat Patients

The suggestion that different decision-making systems can drive behavior provides a very interesting treatment possibility, which is that one could potentially use one decision-system to correct for errors in another. Three computational analyses of this have been done—changing discounting rates with episodic future thinking (Peters and Büchel 2010; Snider et al. 2018; Stein et al. 2018), analyses of contingency management (Petry 2012; Regier and Redish 2015), and analyses of precommitment (Kurth-Nelson and Redish 2009).

Episodic future thinking is a process in which one imagines a future world (Atance and O'Neill 2001), which is the key to planning and model-based decision making, in which one simulates (imagines) an outcome, and then makes one's decision based on that imagined future world (Niv et al. 2006; Redish 2013, 2016). Models of planning suggest that discounting rates may depend in part on the ability to imagine those concrete futures. Part of the discounting may arise from the intangibility of that future (Rick and Loewenstein 2008; Trope and Liberman 2010; Kurth-Nelson, Bickel, and Redish 2012), which may explain why making future outcomes more concrete reduces discounting rates (Peters and Büchel 2010). Other models have suggested that these discounting rate decreases occur through changes in the balance between impulsive

and more cognitive decision systems (McClure and Bickel 2014). Nevertheless, recent work has found that treatments in which subjects are provided concrete episodic future outcomes to guide episodic future thinking can decrease discounting rates (providing a more future-oriented attitude) and decrease drug use (Snider et al. 2018; Stein et al. 2018). Whether this effect comes from the changes in discounting rates per se or whether those changes are reflective of other processes (such as an increased ability to use planning and deliberative systems) is currently unknown.

Contingency management is a treatment to create behavioral change (such as stopping use of drugs) through the direct payment of rewards for achieving that behavioral change—effectively paying people to stop taking drugs (Petry 2012). Contingency management was originally conceived of economically: if drugs have some elasticity (which they do; see figure 9.1), then paying people not to take drugs increases the cost of taking drugs by creating lost opportunity costs. In psychology, this would be called an alternate reinforcer.

However, Regier and Redish (2015) noted that the rewards that produced success in contingency management did not match the inelasticity seen in either animal models of addiction nor in real world measures of inelasticity due to changes of drug costs in the street. Building on the idea that choosing to take a drug or not (a go/no-go task, asking one's willingness-to-pay) accesses different decision-making algorithms than choosing between two options (take the drug or get the alternate reward), Regier and Redish suggested that contingency management had effectively nudged the subject to use their deliberative decision-making systems. They then suggested that this could provide improvements to standard contingency-management methods, including testing for prefrontal-hippocampal integrity (critical to deliberative systems) and providing concrete alternatives with reminders (making it easier to imagine those potential futures). Whether these suggestions actually improve contingency management has not yet been tested.

The fact that addicts show fast discounting functions with preferences that change over time suggests two interesting related treatments: bundling and precommitment. Bundling is a process whereby multiple rewards are grouped together so as to calculate the value of the full set rather than each individually (Ainslie 2001). For example, an alcoholic may want to go to the bar to drink one beer, but recognizing that going to the bar will entail lots of drinking may reduce the value of going to the bar relative to staying home. This can shift the person's preferences from going to the bar to staying home.

A similar process is that of precommitment, where a subject who knows in advance that if given a later option, the subject will take the poor choice, prevents the

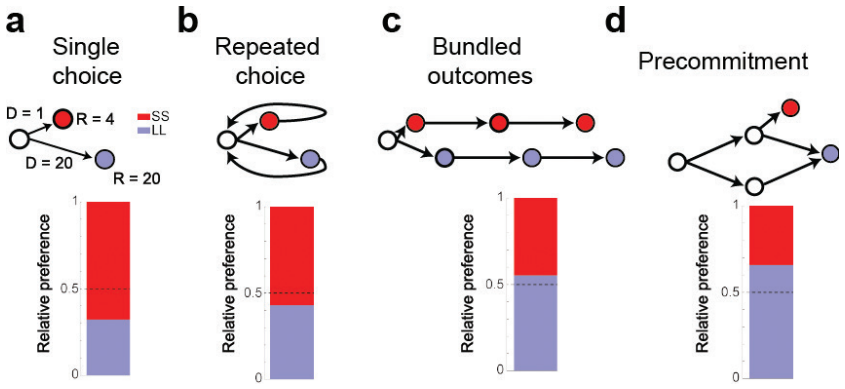

Figure 9.5
Changing state spaces. (a) Imagine a single choice between a smaller reward (R = 4) delivered sooner (after 1 second), compared to a larger reward (R = 20) delivered later (after 20 seconds). A typical agent might prefer the smaller-sooner over the larger-later reward. (b) If the agent realizes that this is going to be a repeated choice, then it is possible to drive the relative preference to 50/50 with a long look ahead, but it is impossible to change the actual preference. An agent that prefers the smaller-sooner option in (a) will still prefer it in (b). (c) Bundling creates new options such that there are consequences to one's decision. An agent can switch preferences by bundling. (d) Precommitment adds a new option to skip the choice. An agent making a decision at the earlier option can prefer the larger-later and learn to skip the choice in the right conditions. After models in Kurth-Nelson and Redish (2012).

opportunity in the first place. The classic example is that a person who knows they will drink too much at the bar decides not to go to the bar in the first place. Economically, precommitment depends on the hyperbolic discounting factors that lead to preference reversals (Ainslie 2001). Preference reversals imply that the earlier person wants one option (to not drink) while the later person wants a different one (to drink). Although many experiments have found that the average subject shows hyperbolic discounting (Madden and Bickel 2010), individuals can show large deviations from good hyperbolic fits. Computationally, an individual's willingness to precommit should depend on the specific shape of their discounting function (Kurth-Nelson and Redish 2010).

Furthermore, Kurth-Nelson and Redish (2010) proved that, neurophysiologically, precommitment depends on having a multifaceted value function—that is, the neural implementation of valuation has to be able to represent multiple values simultaneously. One obvious possibility is that the multiple decision-making systems each value options differently, and conflict between these options can be used to drive precommitment to prevent being offered the addictive option in the first place.

9.5 Chapter Summary

Because addiction is fundamentally a problem with decision making, computational models of decision making (whether economic, motivational [pharmacological], or neurosystem) have been important to our definitions and understanding of addiction. These theories have led to new treatments and new modifications that could improve those treatments.

9.6 Further Study

Koob and Le Moal (2006) provide a thorough description of the known neurobiology of addiction.

Bickel et al. (1993) is a seminal article showing that behavioral economics provides a conceptual framework that has utility for the study of drug dependence.

Redish (2004) was the first explicitly computational model of drug addiction and set the stage for considering addiction as computational dysfunction in decision systems.

Redish et al. (2008) provides evidence that addiction is a symptom rather than a fundamental disease and proposed that the concept of vulnerabilities in decision processes offers a unified framework for thinking about addiction.