Addiction as a computational process gone awry

A. David Redish

Department of Neuroscience

University of Minnesota

6-145 Jackson Hall

321 Church St. SE

Minneapolis MN 55455

In press, *Science*

accepted 19 October 2004

Abstract

Addictive drugs have been hypothesized to access the same neurophysiological mechanisms as natural learning systems. These natural learning systems can be modeled through temporal-difference reinforcement learning (TDRL), which requires a reward-error signal that has been hypothesized to be carried by dopamine. TDRL learns to predict reward by driving that reward-error signal to 0. By adding a non-accommodatable drug-induced dopamine increase to a TDRL model, a computational model of addiction is constructed which overselects actions leading to drug-receipt. The model provides an explanation for important aspects of the addiction literature and provides a theoretic viewpoint with which to address other aspects.

If addiction accesses the same neurophysiological mechanisms used by normal reinforcement-learning systems (*1–3*), then it should be possible to construct a computational model based on current reinforcement-learning theories (*4–7*) that inappropriately selects an "addictive" stimulus. In this paper, I present a computational model of the behavioral consequences of one effect of drugs of abuse, that of increasing phasic dopamine levels through neuropharmacological means. Many drugs of abuse increase dopamine levels either directly (e.g. cocaine (*8*)) or indirectly (e.g. nicotine (*9*, *10*) and heroin (*11*)). A neuropharmacologically-driven increase in dopamine is not the sole effect of these drugs, nor is it likely to be the sole reason that drugs of abuse are addictive. However, this model provides an immediate explanation for several important aspects of the addiction literature, including the sensitivity of the probability of selection of drug-receipt to prior drug-experience, to the size of the contrasting non-drug reward, and the sensitivity but inelasticity of drugs of abuse to cost.

The proposed model is based on temporal-difference reinforcement models in which actions are selected so as to maximize future reward (*6*, *7*). This is done through the calculation of a value function $V(s(t))$, dependent on the state of the world $s(t)$. The value function is defined as the expected future reward, discounted by the expected time to reward:

$$V(t) = \int_t^\infty \gamma^{\tau - t} E[R(\tau)] d\tau \tag{1}$$

where $E[R(\tau)]$ is the expected reward at time τ, and γ is a discounting factor ($0 < \gamma < 1$), reducing the value of delayed rewards. Eq. 1 assumes exponential discounting in order to accommodate the learning algorithm (*6*, *7*), however, animals (including humans) show hyperbolic discounting of future rewards (*12*, *13*). This will be addressed by including multiple discounting time-scales within the model (*14*).

In TDRL, an agent (the subject) traverses a world consisting of a limited number of explicit states. The state of the world can change due to the action of the agent or as a process inherent in

the world (i.e. external to the agent). For example, a model of delay conditioning may include an inter-stimulus-interval state (indicated to the agent by the observation of an ongoing tone); after a set dwell-time within that state, the world transitions to a reward-state and delivers a reward to the agent. This is an example of changing state due to processes external to the agent. In contrast, in an model of FR1 conditioning, an agent may be in an action-available state (indicated by the observation of a lever available to the agent) and the world will remain in the action-available state until the agent takes the action (of pushing the lever) which will move the world into a reward-state. For simplicity later, an available action will be written as $S_k \xrightarrow{a_i} S_l$ which indicates that the agent can achieve state S_l if it is in state S_k and selects action a_i. Although the model in this paper is phrased in terms of the agent taking "action" a_i, addicts have very flexible methods of finding drugs. It is not necessary for the model "actions" to be simple motor actions. $S_k \xrightarrow{a_i} S_l$ indicates the availability of achieving state S_l from state S_k. The agent selects actions proportional to the expected "benefit" that would be accrued from taking the action; the expected benefit can be determined from the expected change in value and reward (*4*, *6*, *14*, *15*).

The goal of TDRL is to correctly learn the value of each state. This can be learned by calculating the difference between expected and observed changes in value (*6*). This signal, termed δ, can be used to learn sequences which maximize the amount of reward received over time (*6*). δ is not equivalent to pleasure, instead, it is an internal signal indicative of the discrepancy between expectations and observations (*5*, *7*, *15*). Essentially, if the change in value or the achieved reward was better than expected ($\delta > 0$), then one should increase the value of the state that led to it. If it was no different from expected ($\delta = 0$), than the situation is well-learned and nothing needs to be changed. Because δ transfers backwards from reward states to anticipatory states with learning, actions can be chained together to learn sequences (*6*). This is the heart of the temporal-difference reinforcement-learning (TDRL) algorithm (*4–7*).

TDRL learns the value function by calculating two equations as the agent takes each action. If the agent leaves state S_k and enters state S_l at time t, at which time it receives reward $R(S_l)$, then

$$\delta(t) = \gamma^d(R(S_l) + V(S_l)) - V(S_k) \tag{2}$$

where γ^d indicates raising the discounting factor γ by the delay d spent by the animal in state S_k (*14*). $V(S_k)$ is then updated as

$$V(S_k) \leftarrow V(S_k) + \eta_V \cdot \delta \tag{3}$$

where η_V is a learning rate parameter.

Phasic increases in dopamine are seen after unexpected natural rewards (*16*), however, with learning, these phasic increases shift from the time of reward delivery to cueing stimuli (*16*). Transient increases in dopamine are now thought to signal changes in the expected future reward (i.e. unexpected changes in value) (*4*, *16*). This can occur either with unexpected reward or with unexpected cue stimuli known to signal reward (*16*), and have been hypothesized to signal δ (*4*, *7*, *16*). Models of dopamine signalling as δ have been found to be compatible with many aspects of the data (*4*, *5*, *16*, *17*).

The results simulated below follow from the incorporation of neuropharmacologically produced dopamine into temporal difference models. The figures below were generated from a simulation using a TDRL instantiation that allows for action selection within a semi-Markov state space, enabling simulations of delay-related experiments (*14*). The model also produces hyperbolic discounting under normal conditions, consistent with experimental data (*12*, *13*) by a summation of multiple exponential discounting components (*14*), a hypothesis supported by recent fMRI data (*18*).

The effect of neuropharmacological release of dopamine on TDRL. The key to TDRL is that once the value function correctly predicts the reward, learning stops. The value function can be said to "accommodate" the reward: the change in value in taking action $S_k \xrightarrow{a_i} S_l$ counterbalances the reward achieved on entering state S_l. When this happens, $\delta = 0$. Taking transient dopamine as the δ signal (*4*, *5*, *7*), correctly predicted rewards produce no dopamine signal (*16*, *17*).

However, cocaine and other addictive drugs produce a transient increase in dopamine through

neuropharmacological mechanisms (*1*, *2*, *8*). The concept of a neuropharmacologically-produced dopamine surge can be modeled by assuming that these drugs induce a "non-accommodatable" increase in δ (*19*). In other words, the effect of addictive drugs is to produce a positive δ, independent of the change in value function, making it impossible for the agent to learn a value function which will cancel out the drug-induced increase in δ. Eq. 2 is thus replaced with

$$\delta = \max(\gamma^d(R(S_l)+V(S_l))-V(S_k)+D(S_l),D(S_l)) \tag{4}$$

where $D(S_l)$ indicates a dopamine surge occurring on entry into state S_l. Eq. 4 reduces to normal TDRL (Eq. 2) when $D(S_l)=0$, but asymptotes to a minimum δ of $D(S_l)$ when $D(S_l)>0$. This always produces a positive reward-error signal. Thus the values of states leading to a dopamine surge $D>0$ will approach infinity.

When given a choice between two actions $S_0 \xrightarrow{a_1} S_1$ and $S_0 \xrightarrow{a_2} S_2$, the agent chooses actions proportional to the values of the subsequent states, S_1 and S_2. The more valuable the state taking an action leads to, the more likely the agent is to take that action. In TDRL, the values of states leading to natural rewards asymptote at a finite value (the discounted, total expected future reward); however, in the modified model, the values of states leading to drug-receipt increase without bound. Thus the more the agent traverses the action sequence leading to drug-receipt, the larger the value of the states leading to that sequence, and the more likely the agent is to select an action leading to those states.

Simulations. In this model, drug-receipt produces a $\delta>0$ signal, which produces an increase in the value of states leading to the drug-receipt. Thus, the value of states leading to drug-receipt increase without bound. In contrast, the values of states leading to natural reward asymptote to a value approximating Eq. 1. This implies that the selection probability between actions leading to natural rewards will reach an asymptotic balance. However, the selection probability of actions leading to drug-receipt will depend on the number of experiences. Simulations bear this out (Fig. 1).

In the simulations, drug-receipt entails a normal-sized, accomodatable reward $R(s)$ and a small,

non-accomodatable dopamine signal $D(s)$ (*14*). Early use of drugs occurs because they are highly rewarding (*1*, *3*, *20*), but this use transitions to a compulsive use with time (*1*, *3*, *20*, *21*, *32*). In the model, the $R(s)$ term provides for the early rewarding component, while the gradual effect of the $D(s)$ term provides for the eventual transition to addiction. This model thus shows that a transition to addiction can occur without any explicit sensitization or tolerance to dopamine, at least in principle.

The unbounded increase in value of states leading to drug-reward does not mean that with enough experience, drugs of abuse are always selected over non-drug rewards. Instead, it predicts that the likelihood of selecting the drug over a non-drug reward will be dependent on the size of the contrasting non-drug reward relative to the current value of the states leading to drug-receipt (Fig. 1).

When animals are given a choice between food and cocaine, the probability of selecting cocaine depends on the amount of food available as an alternative and the cost of each choice (*22*, *23*). Similarly, humans given a choice between cocaine and money will decrease their cocaine selections with increased value of the alternative (*24*). This may explain the success of vouchers in treatment (*24*). This will continue to be true even in well-experienced (highly-addicted) subjects, but the sensitivity to the alternate should decrease with experience (see below). This may explain the incompleteness of the success of vouchers (*24*).

Natural rewards are sensitive to cost in that animals (including humans) will work harder for more valuable rewards. This level of sensitivity is termed elasticity in economics. Addictive drugs are also sensitive to cost in that increased prices decrease usage (*25*, *26*). However, while the use of addictive drugs does show sensitivity to cost, that sensitivity is significantly inelastic relative to similar measures applied to natural rewards (*25*, *27*). The TDRL model proposed here produces just such an effect: Both modeled drugs and natural rewards are sensitive to cost, but drug reward is less elastic than the natural rewards (Fig. 2).

In TDRL, the value of states leading to natural rewards asymptote to a stable value which depends on the time to the reward, the reward level, and the discounting factors. However, in the modified TDRL model, the value of states leading to drug rewards increase without bound, producing a ratio

of a constant cost to increasing value. This decreasing ratio predicts that the elasticity of drugs to cost should decrease with experience, while it should not for natural rewards (Fig. S4).

Predictions

Developing inelasticity. The hypothesis that values of states leading to drug-receipt increase without bound implies that the elasticity to cost should decrease with use, while the elasticity of natural rewards should not. This also suggests that increasing the reward for not choosing the drug (such as vouchers (*24*)) will be more effective early in the transition from casual drug-use to addiction.

Blocking. The hypothesis that cocaine produces a $\delta > 0$ dopamine signal on drug-receipt implies that cocaine should not show blocking. Blocking is an animal-learning phenomenon in which pairing a reinforcer with a conditioning stimulus does not show association if the reinforcer is already predicted by another stimulus (*17*, *28*, *29*). For example, if a reinforcer X is paired with cue A, animals will learn to respond to cue A. If X is subsequently paired with simultaneously-presented cues A and B, animals will not learn to associate X with B. This is thought to occur because X is completely predicted by A, and there is no error signal ($\delta = 0$) to drive the learning (*17*, *28*, *29*). If cocaine is used as the reinforcer instead of natural rewards, the dopamine signal should always be present ($\delta > 0$), even for the AB stimulus. Thus, cocaine (and other drugs of abuse) should not show blocking.

Dual dopamine signals in experienced users. The hypothesis that the release of dopamine by cocaine accesses TDRL systems implies that experienced animals will show a double dopamine signal in cued-response tasks (*14*). As with natural rewards, a transient dopamine signal should appear to a cueing signal that has been associated with reward (*16*). However, whereas natural rewards only produce dopamine release if unexpected (*16*, *17*), cocaine produces dopamine release directly (*8*). Thus, after learning, both the cue and the cocaine should produce dopamine (Fig. 3). Supporting this hypothesis, Phillips *et al.* (*30*) found using fast-scan cyclic voltammetry that in rats trained to associate an audiovisual signal with cocaine, both the audiovisual stimulus and the cocaine itself produced dramatic increases in the extracellular concentration of dopamine in the nucleus accumbens.

Discussion. Substance abuse is a complex disorder. TDRL explains some phenomena that arise in addiction, and makes testable predictions about other phenomena. The test of a theory such as this one is not whether it encompasses all phenomena associated with addiction, but whether the predictions that follow from it are confirmed.

This model has been built on assumptions about cocaine, but cocaine is far from the only substance that humans (and other animals) abuse. Many drugs of abuse indirectly produce dopamine signals, including nicotine (*10*) and heroin and other opiates (*11*). Although these drugs have other effects as well (*1*), the effects on dopamine should produce the consequences described above, leading to inelasticity and compulsion.

Historically, an important theoretical explanation of addictive behavior has been that of "rational addiction" (*31*), in which the user is assumed to maximize value or "utility" over time, but because long-term rewards for quitting are discounted more than short-term penalties, the maximized function entails remaining addicted. The TDRL theory proposed in this paper differs from that of rational addiction in that TDRL proposes that addiction is inherently "irrational" — it uses the same mechanisms as natural rewards, but the system behaves in a non-optimal way due to neuropharmacological effects on dopamine. Because the $D(s)$ component cannot be accomodated, it eventually overwhelms the $R(s)$ reward terms (from both drug and contrasting natural rewards). Eventually, the agent behaves irrationally and rejects the larger rewards in favor of the (less-rewarding) addictive stimulus. The TDRL and rational-addiction theories make testably different predictions: While rational-addiction predicts that drugs-of-abuse will show similar elasticity to cost as natural rewards do, the TDRL theory predicts that drugs-of-abuse will show increasing inelasticity with use.

The rational addiction theory (*31*) assumes exponential discounting of future rewards, while humans and other animals consistently show hyperbolic discounting of future rewards (*12*, *13*). Ainslie (*13*) has suggested that the "cross-over" effect that occurs with hyperbolic discounting explains many aspects of addiction. The TDRL model used here also shows hyperbolic discounting (*14*), and so accesses the results noted by Ainslie (*13*). However, in the theory proposed here, hyperbolic discount-

ing is not the fundamental reason for the agent getting trapped in a non-optimal state. Rather, the TDRL theory hypothesizes that it is the neuropharmacological effect of certain drugs on dopamine signals that drives the agent into the non-optimal state.

Robinson and Berridge (*32*) have suggested that dopamine mediates the desire to achieve a goal ("wanting"), differentiating "wanting" from the hedonic desire of "liking." As noted by McClure *et al*. (*15*), the concept of incentive salience has a direct correspondence to variables in TDRL: that of value of a state reachable by an action: If an agent is in state S_0, and can achieve state S_1 via action $S_0 \xrightarrow{a_i} S_1$, then if state S_1 has a much greater value than state S_0, $S_0 \xrightarrow{a_i} S_1$ can be said to be a pathway with great incentive salience. The value function is a means of guiding decisions, and thus is more similar to "wanting" than to "liking" in the terminology of Robinson and Berridge (*15*, *32*). In TDRL, dopamine does not directly encode "wanting," but because learning an appropriate value function depends on an accurate δ signal, dopamine will be necessary for acquisition of "wanting".

Many unmodeled phenomena play important roles in the compulsive self-administration of drugs of abuse (*1*), including titration of internal levels of drug (*33*), sensitization and tolerance (*34*), withdrawal symptoms and release from them (*20*), and compensation mechanisms (*35*, *38*). Additionally, individuals show extensive inter-personal variability (*36*, *37*). While these aspects are not addressed in the model presented here, many of these can be modeled by adding parameters to the model: for example, sensitization can be included by allowing the drug-induced-delta parameter $D(s)$ to vary with experience.

TDRL forms a family of computational models with which to model addictive processes. Modifications of the model can be used to incorporate the unmodeled experimental results from the addiction literature. For example, an important question in this model is whether the values of states leading to drug-receipt truly increase without bound. I find this highly unlikely. Biological compensation mechanisms (*35*, *38*) are likely to limit the maximal effect of cocaine on neural systems including the value representation. This can be modeled in a number of ways. One of which is to include a global effectiveness-of-dopamine factor, which multiplies all $R(s)$ and $D(s)$ terms. If this

factor decreased with each drug-receipt, the values of all states would remain finite. Simulations based on an effectiveness-of-dopamine factor that decreases exponentially with each drug-receipt (factor $= 0.99^n$, where n = number of drug-receipts) showed similar properties to those reported here, but the values of all states remained finite.

Another important issue in reinforcement-learning is what happens when the reward or drug is removed. In normal TDRL, the value of states leading to reward decay back to zero when that reward is not delivered (*6*). This follows from the existence of a strongly negative δ signal in the absense of expected reward. Although firing of dopamine neurons is inhibited in the absense of expected reward (*16*), the inhibition is dramatically less than the corresponding excitation was (*7*). In general, the simple decay of value seen in TDRL (*6*, *39*) does not model extinction very well, particularly in terms of reinstantiation after extinction (*40*). Modeling extinction (even for natural rewards) is likely to require additional components not included in current TDRL models, such as state-space expansion.

Conclusion A theory of addiction that is compatible with a large literature of extant data and that makes explicitly testable predictions has been deduced from two simple hypotheses: (1) dopamine serves as a reward-error learning signal to produce temporal-difference learning in the normal brain, and (2) cocaine produces a phasic increase in dopamine directly (i.e. neuropharmacologically). A computational model was derived by adding a non-accomodatable δ signal to a TDRL model. The theory makes predictions about human behavior (developing inelasticity), animal behavior (resistance to blocking), and neurophysiology (dual dopamine signals in experienced users). Addiction is likely to be a complex process arising from transitions between learning algorithms (*3*, *20*, *32*). Bringing addiction theory into a computational realm will allow us to make these theories explicit and to directly explore these complex transitions.

References and Notes

1. J. H. Lowinson, P. Ruiz, R. B. Millman, J. G. Langrod, eds. *Substance Abuse: A comprehensive Textbook* (Williams and Wilkins, 1997), third ed.

2. M. E. Wolf, S. Mangiavacchi, X. Sun. *Annals of the NY Academy of Science* **1003**, 241 (2003).

3. B. J. Everitt, A. Dickinson, T. W. Robbins. *Brain Research Reviews* **36**, 129 (2001).

4. P. R. Montague, P. Dayan, T. J. Sejnowski. *Journal of Neuroscience* **16**, 1936 (1996).

5. W. Schultz, P. Dayan, R. Montague. *Science* **275**, 1593 (1997).

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

7. N. D. Daw. Reinforcement learning models of the dopamine system and their behavioral implications. Ph.D. thesis, Carnegie Mellon University (2003).

8. M. C. Ritz, R. J. Lamb, S. R. Goldberg, M. J. Kuhar. *Science* **237**, 1219 (1987).

9. M. R. Picciotto. *Drug and Alcohol Dependence* **51**, 165 (1998).

10. V. I. Pidoplichko, M. DeBiasi, J. T. Williams, J. A. Dani. *Nature* **390**, 401 (1997).

11. S. R. Laviolette, R. A. Gallegos, S. J. Henriksen, D. van der Kooy. *Nature Neuroscience* **7**, 160 (2004).

12. J. E. Mazur. *Psychological Review* **108**, 96 (2001).

13. G. Ainslie. *Picoeconomics* (Cambridge Univ Press, 1992).

14. See *Supplementary online material* for details.

15. S. M. McClure, N. D. Daw, P. R. Montague. *Trends in Neurosciences* **26**, 423 (2003).

16. W. Schultz. *Journal of Neurophysiology* **80**, 1 (1998).

17. P. Waelti, A. Dickinson, W. Schultz. *Nature* **412**, 43 (2001).

18. S. C. Tanaka, K. Doya, G. Okada, K. Ueda, Y. Okamoto, *et al. Nature Neuroscience* **7**, 887 (2004).

19. G. Di Chiara. *European Journal of Pharmacology* **375**, 13 (1999).

20. G. F. Koob, M. L. Moal. *Neuropsychopharmacology* **24**, 97 (2001).

21. L. J. M. J. Vanderschuren, B. J. Everitt. *Science* **305**, 1017 (2004).

22. M. E. Carroll, S. T. Lac, S. L. Nygaard. *Psychopharmacology* **97**, 23 (1989).

23. M. A. Nader, W. L. Woolverton. *Psychopharmacology* **105**, 169 (1991).

24. S. T. Higgins, S. H. Heil, J. P. Lussier. *Annual Review of Psychology* **55**, 431 (2004).

25. M. E. Carroll. *Drug and Alcohol Dependence* **33**, 201 (1993).

26. M. Grossman, F. J. Chaloupka. *Journal of Health Economics* **17**, 427 (1998).

27. W. K. Bickel, L. A. Marsch. *Addiction* **96**, 73 (2001).

28. R. A. Rescorla, A. R. Wagner. In *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokesy, eds. (Appleton Century Crofts, New York, 1972), pp. 64–99.

29. A. Dickinson. *Contemporary animal learning theory* (Cambridge University Press, New York, 1980).

30. P. E. M. Phillips, G. D. Stuber, M. L. A. V. Heien, R. M. Wightman, R. M. Carelli. *Nature* **422**, 614 (2003).

31. G. S. Becker, K. M. Murphy. *Journal of Political Economy* **96**, 675 (1988).

32. T. E. Robinson, K. C. Berridge. *Annual Reviews of Psychology* **54**, 10.1 (2003).

33. M. Woodward, H. Tunstall-Pedoe. *Addiction* **88**, 821 (1993). .

34. C. W. Bradberry. *The Neuroscientist* **8**, 315 (2002).

35. S. R. Letchworth, M. A. Nader, H. R. Smith, D. P. Friedman, L. J. Porrino. *Journal of Neuroscience* **21**, 2799 (2001).

36. N. Volkow, J. Fowler, G.-J. Wang. *Behavioral Pharmacology* **13**, 355 (2002).

37. V. Deroche-Gamonet, D. Belin, P. V. Piazza. *Science* **305**, 1014 (2004).

38. F. J. White, P. W. Kalivas. *Drug and Alcohol Dependence* **51**, 141 (1998).

39. R. E. Suri, W. Schultz. *Neuroscience* **91**, 871 (1999).

40. M. E. Bouton. *Biological Psychiatry* **52**, 976 (2002).

41. This work was supported by an Alfred P. Sloan Fellowship. I thank Mark Thomas, Erin Larson, Marilyn Carroll, Dorothy Hatsukami, Jadin Jackson, Adam Johnson, Neil Schmitzer-Torbert and Zeb Kurth-Nelson for comments on the manuscript and helpful discussions.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

List of Figures

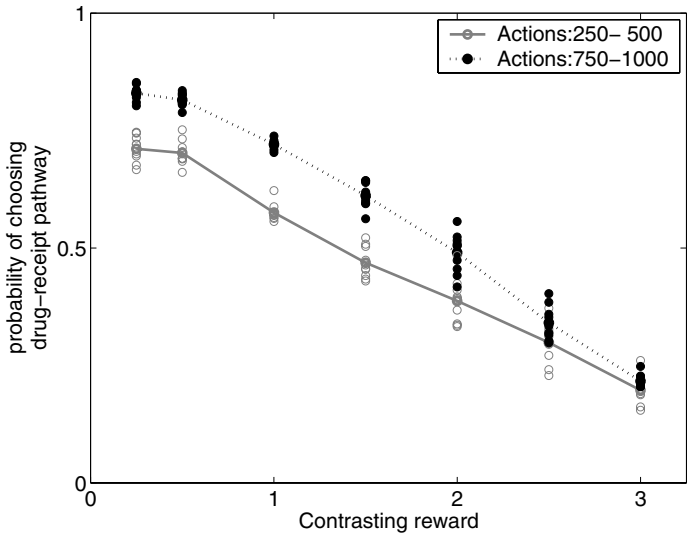

FIGURE 1: Probability of selecting a drug-receipt pathway depends on an interaction between drug-level, experience, and contrasting reward. [Simulation.] Each line shows the average probabilty of selecting the drug-receipt pathway, $S_0 \xrightarrow{a_2} S_2$, over the contrasting reward pathway, $S_0 \xrightarrow{a_1} S_1$, as a function of the size of the contrasting reward $R(S_3)$. (State-space is shown in Fig. S1.). Drug-receipt on entering state S_4 was $R(S_4) = 1.0$, $D(S_4) = 0.025$. Individual simulations are shown by dots. Additional details provided in *Supplemental Material* (*14*).

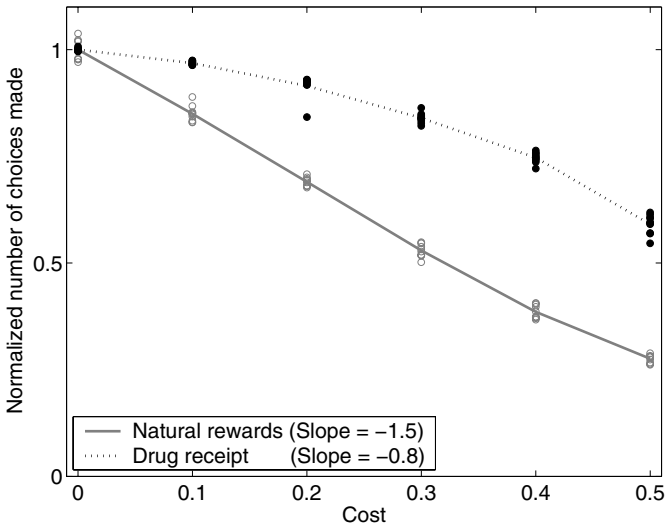

FIGURE 2: Elasticity of drug-receipt and natural rewards. [Simulation.] Both drug-receipt and natural rewards are sensitive to costs, but natural rewards are more elastic. Each dot indicates the number of choices made within a session. Sessions were limited by simulated time. The curves have been normalized to the mean number of choices made at 0 cost.

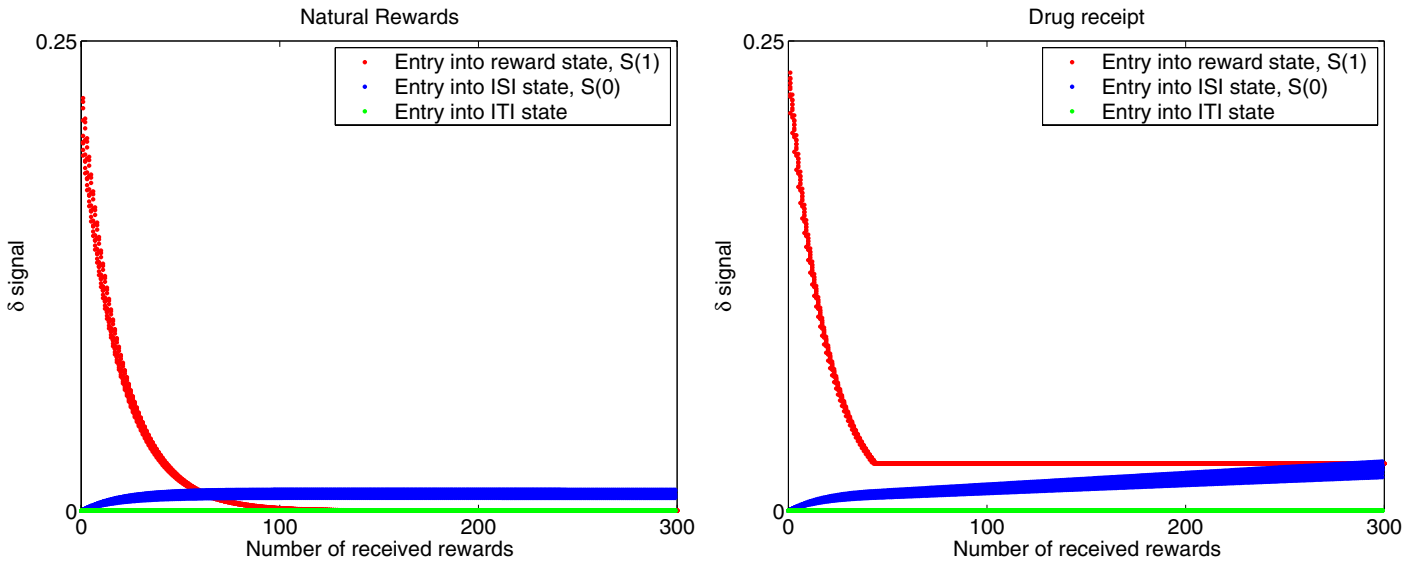

FIGURE 3: Dopamine signals. [Simulation.] **Left** With natural rewards, dopamine initially occurs primarily at reward receipt (on entry into reward state S_1), and shifts to the conditioned stimulus (on entry into ISI state S_0) with experience. (State space is shown in Fig. S7.) **Right** With drugs that produce a dopamine signal neuropharmacologically, dopamine continues to occur at the drug-receipt (on entry into reward state S_1), even after experience, as well as shifting to the conditioned stimulus (on entry into ISI state S_0), thus producing a double dopamine signal.

Addiction as a computational process gone awry
SUPPLEMENTAL MATERIAL

A. David Redish*

19 October 2004

Contents

List of Figures

List of Tables

*Department of Neuroscience, University of Minnesota, 6-145 Jackson Hall, 321 Church St. SE, Minneapolis MN 55455. email: redish@ahc.umn.edu

The μAgent temporal difference reinforcement learning model

As noted in the paper, the goal of temporal-difference reinforcment learning (TDRL) is to learn to select actions so as to maximize future reward. This is done by learning a value function $V(s)$ dependent on the state of the world s. See Equation 1 of the main paper.

The qualitative results and predictions in the main paper derive from the explicit hypotheses, specifically that (1) the normal brain uses a temporal-difference reinforcement learning algorithm for normal learning of action selection, (2) dopamine serves as the reward-error signal within this TDRL algorithm, and (3) drugs of abuse produce a phasic increase in dopamine directly (i.e. neuropharmacologically). The qualitative results and predictions do not critically depend on the specific instantiation of TDRL used. A number of TDRL variants exist, each with subtle differences (*S1–S15*). Sufficient data are not yet available to enable a decision between these detailed instantiations of TDRL. However, in order to show simulation results, we must commit to an instantiation. We will commit to the μAgents model of Kurth-Nelson and Redish (*S13*, full paper in preparation). This TDRL instantiation lives within a partially-observable semi-Markov process model, enabling time-dependent experiments, including discounting. In addition, it is the only current model to show true hyperbolic discounting which is an important aspect of the extant data (*S16–S18*). But, again, I stress that neither the compatibility of the model with extant data, nor the predictions arising from the underlying hypotheses are dependent on the specifics of this model.

History. The importance of reward-error as a learning signal traces back to Rescorla and Wagner (*S19*). Temporal difference reinforcement learning came into being in the early 1980's following from earlier work by Bellman (*S20*) and other dynamic programming algorithms (see Sutton and Barto (*S11*) for review). The identification of the TD delta signal with dopamine can be traced to three papers in the seminal book *Models of information processing in the basal ganglia* (*S21*), particularly articles by Barto (*S22*), Houk *et al.* (*S23*), and Schultz *et al.* (*S24*), with the first explicit connections between dopamine and TDRL made by Montague, Dayan, and Sejnowski (*S25*) and Schultz, Dayan, and Montague (*S26*), based in large part on the work by Schultz and colleages (*S27–S30*). Since then, a number of TDRL models have been developed (*S1–S15*), each with subtle differences. Due to space limitations, I will not review all of them here.

Three keys to our ability to model the data described in the main paper are semi-Markov state-spaces, the ability to perform action-selection within those state-spaces, and hyperbolic discounting. Semi-Markov state-spaces were first used in TDRL models of natural reward systems by Daw (*S8*), but the complex representation of the agent's believed state as implemented by Daw precluded action-selection (*S8*). The μAgent model used here allows each μAgent to commit to a single believed state, thus allowing action-selection within a semi-Markov state-space. The action-selection procedure, itself, is similar to that used by Montague *et al.* (*S25*), with the modification that each μAgent proposes an action (based on Montague *et al.*'s action-selection procedure), and the overall agent acts based on a

weighted poll of the actions preferred by each μAgent. Weighting is done by the current fitness factor f_i of each μAgent i. (See below.) The hyperbolic discounting derives from each μAgent having a separate discounting function γ_i. This allows the simulation of each μAgent to use the exponential TDRL equations (*S8*, *S20*), while the overall agent shows hyperbolic discounting, consistent with the experimental literature (*S16*, *S18*). Recent fMRI data suggest a gradient of discounting factors across the striatal ventromedial-dorsolateral axis, with faster discounting factors occuring in the ventromedial portion and slower discounting factors occuring in the dorsolateral portion (*S31*).

Model justifaction. The model described in this paper is an abstract model of temporal-difference learning. While more concrete models of basal ganglia exist (*S5*, *S7*, *S23*, *S32*), the actual relationship of TDRL to the basal ganglia is still hotly debated (see, for example, Refs. *S2*, *S33*). I have therefore chosen to use a more abstract model of TDRL, so as to more directly address the hypotheses of the modified TDRL theory.

The world model. In all of the simulations below, the agent lived within a discrete set of possible states, consisting of a semi-Markov process model. Each state entailed a dwell time distribution $T(s)$, and an observation $O(s)$. $O(s)$ was not required to be unique to state s, thus making the process model partially observable. On entering a state, the agent received a (possibly 0) reward $R(s)$, and a (possibly 0) drug-receipt $D(s)$. Transitions between states could occur due to actions selected by the agent or probabilistically according to the dwell time distribution. For most of the simulations below, the dwell time distribution was a single value — that is the agent remains in the state for $T(s) = T_0$ time steps. But the model does not require this.

μAgents. The agent itself consised of a constantly changing set of μAgents, each of which was specified by a four-tuple $\langle s_i, t_i, f_i, \gamma_i \rangle$, which identified the μAgent's believed-state s_i, believed dwell-time within that state t_i, the fitness of the μAgent f_i, and the μAgent's internal discounting parameter γ_i. A μAgent thus represented a hypothesis about the current state of the world, but carried no history with it. Thus the μAgent was essentially Markov and the standard TDRL equations could be used. The fitness of the μAgent $0 \geq f_i \geq 1$ was recalculated on each time-step, reflecting the likelihood of the μAgent's hypothesis, given the observation (or lack of observation) received in the time-step, and given the time spent by μAgent i in it's current state. At each time-step, μAgents were selected for "survival" with a probability equal to their fitness f_i. Rejected μAgents were then replaced by a copy of a "surviving" μAgent selected at random from the remaining population, again with probabilty equal to fitness (so that fitter μAgents were more likely to be chosen to replace rejected μAgents). When "copying", only the s_i and t_i parameters of the μAgent were replaced; γ_i was not changed. The set of μAgents thus provided an instantiation of the belief distribution of the agent across the multiple states of the process model.

Action-selection. Action-selection proceeded as a three-step process. First, the agent calculated the expected "benefit"[1] of each action. Then, the agent selected an action proportional to the benefit of each. Finally, the agent decided whether to take the action or not.

Overall benefit expected from action a was calculated as a weighted average of the expected benefits as calculated by the μAgents. First, each μAgent calculated the μBenefit as:

$$B_i(a) = \begin{cases} V(S_l) + E[R(S_l)] - V(s_i) & \text{if } a \text{ available from } s_i \\ 0 & \text{otherwise} \end{cases} \tag{S1}$$

where $V(s_i)$ was the value of the state the μAgent believed itself to be in, and $V(S_l)$ the value of and $E[R(S_l)]$ the expected reward of the state reached by taking action $S_i \xrightarrow{a} S_l$. The overall expected benefit of each action was then defined as the average over all μAgents:

$$B(a) = \sum_i B_i(a) \tag{S2}$$

Based on these benefits, actions were selected proportionally. Thus, the probability of selecting an action a was proportional to the benefit $B(a)$:

$$P(\text{select } a) = \frac{B(a)}{\sum_a B(a)} \tag{S3}$$

Finally, once the agent selected action a, it decided whether to take action a using a soft-max mechanism:

$$P(\text{take selected action}) = \frac{1}{1 + e^{(-m(B_a - 1))}} \tag{S4}$$

This action-selection process captures the three keys to action-selection: the identification of useful actions, the selection of action based on the change in value expected upon taking the action, and a process that decides whether to act or not, presumably depenendent on the benefit of acting. Other action-selection mechanisms which capture these three key processes (such as that proposed by McClure *et al.* (*S9*)) also produced qualitatively similar results to those shown in the main paper.

Overview This μAgent model, although more complex than some TDRL models, is simple to implement, replicates the extant data on dopamine and cued- and uncued-reward (*S13*, full paper in preparation), allows us to model the important results of the addiction literature (main paper), and shows hyperbolic discounting. Hyperbolic discounting in this model arises because the agent includes multiple exponential discounting parameters (distributed across the μAgents).

[1] "Benefit" as defined here is very similar to "advantage" (*S12*), but because the formulation is not identical, I will term it differently.

Discounting parameter	uniform distribution, $[0.001 < \gamma < 0.999]$
Number of μAgents	1000
Learning rate (η)	0.05
Softmax selection parameter (m)	4

TABLE S1: Parameters used in all simulations.

Simulation details: Selection of drug-reward over non-drug reward

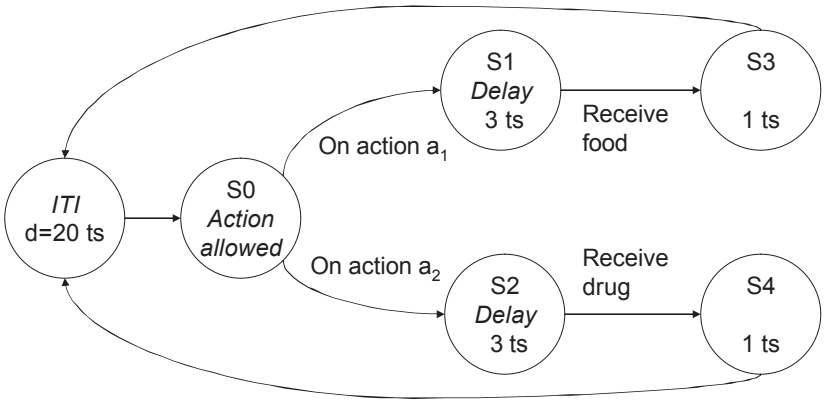

FIGURE S1: State space for selection simulations.

Simulations were based on the 6-state world-model (Figure S1). The five main states S_0, S_1, S_2, S_3, S_4 were fully observable (providing unique observations O_0, O_1, O_2, O_3, O_4 respectively); the *ITI* state was implemented as 1000 identical states, each providing observation O_5. At the beginning of each simulation, the agent began in state S_0. The agent remained in state S_0 until it took an action. On taking action a_1, the world changed to state S_1, where it remained for 3 time-steps, after which it provided a reward $R(S_3)$ to the agent. On taking action a_2, the world changed to state S_2, where it remained for 3 time-steps, after which it provided drug $R(S_4), D(S_4)$ to the agent. After 1 time-step in either state S_3 or S_4 (as appropriate) the world entered the *ITI* state. Actually, the world entered one of the 1000 possible *ITI* states, but the agent distributed it's belief across those states. After 20 time-steps, the world transitioned to state S_0.

This world-model simulates a standard two-lever choice paradigm in which an agent must push one lever to receive food reward and one lever to receive drug, each of which is delivered as appropriate after a short delay. The *ITI* state models the agents lack of knowledge about inter-trial intervals and provides for more realistic simulations in the μAgent model (*S13*).

All non-reward related parameters were held constant. Figure 1 in the *main paper* shows how the probabilty of selecting the drug-reward depended on number of times the agent reached the drug-receipt state (S_4) and on the size of the contrasting reward $R(S_3)$. The selection probability also depended on the size of the drug reward $R(S_4), D(S_4)$. For the figure in the *main paper*, $R(S_4) = 1.0, D(S_4) =$

$0.025, R(S_3) = \{0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. As shown in Figure S2, below, increasing $D(S_4)$ increased the likelihood of selecting the $S_0 \xrightarrow{a_2} S_2$ pathway, but it also changed the shape of the response to counter-food reward.

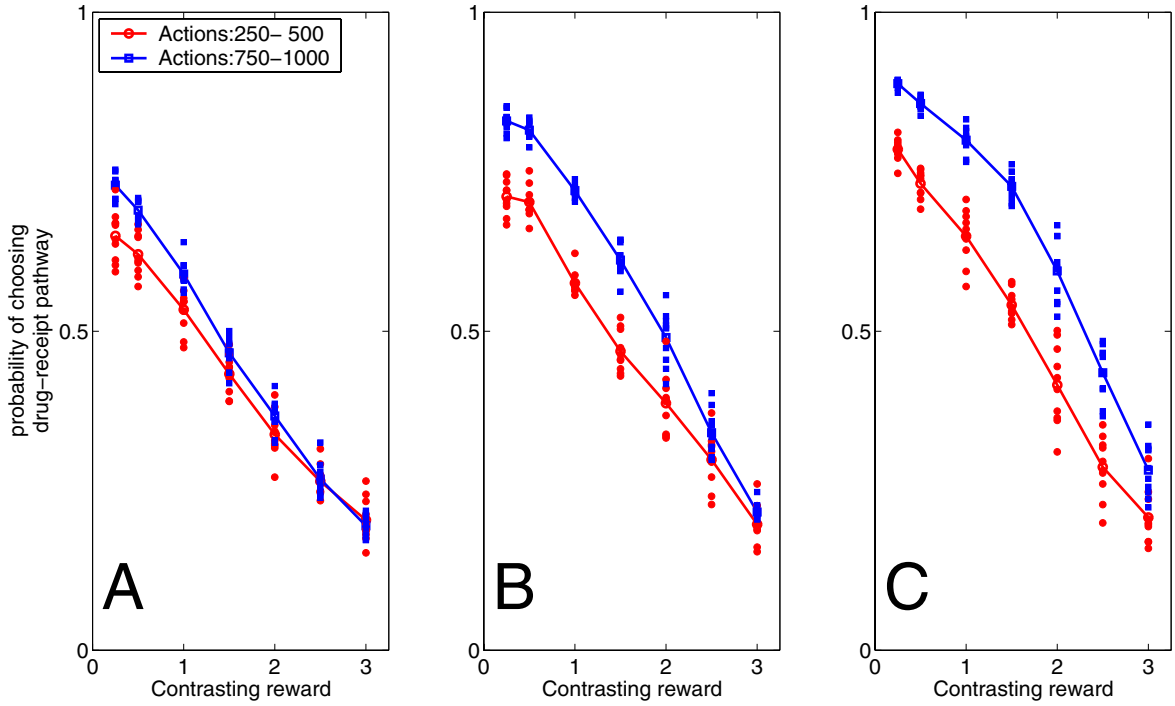

FIGURE S2: Sensitivity of selection to number of drug experiences, size of contrasting food reward, and size of drug-receipt forced-dopamine signal (i.e. strength/dose of the drug). (**A**) $R(S_4) = 1.0, D(S_4) = 0.010$; (**B**) $R(S_4) = 1.0, D(S_4) = 0.025$; (**C**) $R(S_4) = 1.0, D(S_4) = 0.040$.

Simulation details: Sensitivity but inelasticity of drugs of abuse to cost

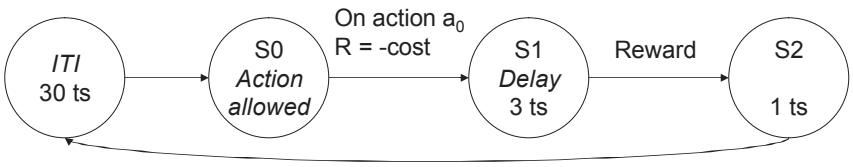

FIGURE S3: State space for elasticity simulation.

The simulations for elasticity were based on a 4-state world-model (Figure S3). Simulations always started in the S_0 *Action-available* state. The world remained in that state unless the agent took action $S_0 \xrightarrow{a_0} S_1$. On taking action a_0, the agent was assesed a cost ($R(S_1) < 0$). The world then remained within state S_1 for 3 time-steps, at which time the world transitioned to state S_2 and the agent received reward. For the simulation of natural rewards, reward was provided as $R(S_2) = 1.0$. For the simulation

of drug rewards, reward was provided as $R(S_2) = 1.0, D(S_2) = 0.025$. States S_0, S_1, S_2 were fully observable, providing observations O_0, O_1, O_2. The *ITI* state was implemented as before (1000 identical states, each providing observation O_3).

Simulations were run for 10^5 time-steps, and the total number of actions taken was measured. In order to determine the elasticity, the number of actions taken when faced with cost C was normalized to the total number of actions taken with no cost ($C = 0$). See Figure 2 in the *main paper*. In order to measure developing inelasticity, the first 500 actions (and thus the first 500 rewards) were measured. See Figure S4.

As noted in the main paper, elasticity changes for drug-receipt, but not for natural rewards. This occurs because the values of states leading to natural rewards asymptote to a bound (approximating Equation 1 in the main paper), while states leading to drug-receipt increase without bound.

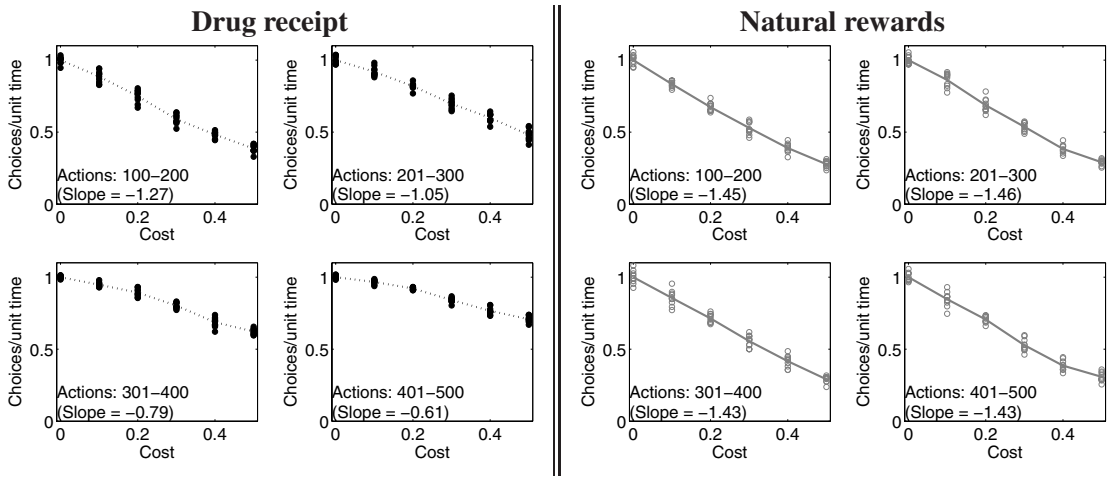

FIGURE S4: Elasticity decreases for drug-receipt but not reward-receipt.

Simulation details: Discounting

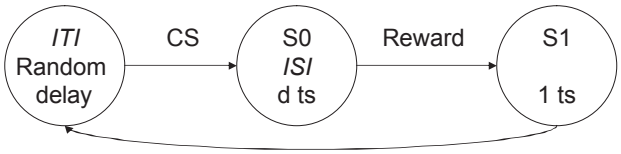

FIGURE S5: State space for discounting simulation.

The simulations for discounting were based on a 3-state world model, shown in Figure S5. The world started in the *ITI* state, after a random delay, the world delivered a conditioning stimulus (CS), and entered state S_0, where it remained for a set time (the delay, d timesteps, the independent variable in the discounting simulation). After that delay, a reward was delivered to the agent and the world

entered state S_1. As before, states S_0 and S_1 were fully observable, providing observations O_0 and O_1, respectively; the *ITI* state was implemented as 1000 states providing identical observations O_2. This models a standard conditioned-stimulus Pavlovian task. No action is required.

Proportional value of a reward was measured as the value of state S_0 after the delivery of 300 rewards. Natural rewards were modeled as $R(S_1) = 1.0, D(S_1) = 0.0$ Figure S6, below, shows that the μAgents model showed hyperbolic discounting with natural rewards.

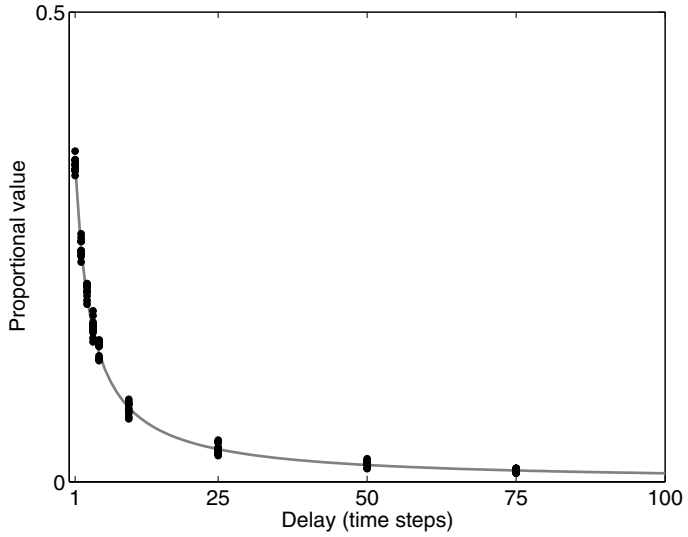

FIGURE S6: Discounting with natural rewards.

Simulation details: Dual dopaminergic signals in experienced users

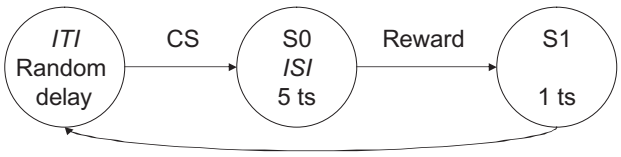

FIGURE S7: State space for dopamine simulation.

The simulations of the dual dopaminergic signal used the same Pavlovian state space as the discounting simulations (Figure S7). The inter-stimulus interval delay (state S_0) was set to a constant 5 steps. Natural rewards were modeled as $R(S_1) = 1.0, D(S_1) = 0.0$; drug-receipt was modeled as $R(S_1) = 1.0, D(S_1) = 0.025$.

References and Notes

S1. N. D. Daw, S. Kakade, P. Dayan. *Neural Networks* 15, 603 (2002).

S2. P. Dayan, B. W. Balleine. *Neuron* 36, 285 (2002).

S3. P. Dayan. In *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, Z. Ghahramani, eds. (MIT Press, Cambridge, MA, 2002).

S4. S. Kakade, P. Dayan. *Neural Networks* 15, 549 (2002).

S5. R. E. Suri. *Neural Networks* 15, 523 (2002).

S6. R. E. Suri, J. Bargas, M. A. Arbib. *Neuroscience* 103, 65 (2001).

S7. R. E. Suri, W. Schultz. *Neuroscience* 91, 871 (1999).

S8. N. D. Daw. Reinforcement learning models of the dopamine system and their behavioral implications. Ph.D. thesis, Carnegie Mellon University (2003).

S9. S. M. McClure, N. D. Daw, P. R. Montague. *Trends in Neurosciences* 26, 423 (2003).

S10. N. D. Daw, D. S. Touretzky. *Neural Computation* 14, 2567 (2002).

S11. R. S. Sutton, A. G. Barto. *Reinforcement Learning: An introduction* (MIT Press, Cambridge MA, 1998).

S12. L. C. Baird. *Advantage Updating*. Tech. Rep. WL-TR-93-1146, Wright-Patterson Air Force Base Ohio: Wright Laboratory (1993). Available from the Defense Technical Information Center, Cameron Station, Alexandria, VA 22304-6145.

S13. Z. Kurth-Nelson, A. D. Redish. *Society for Neuroscience Abstracts* (2004).

S14. K. Doya, K. Samejima, K.-I. Katagiri, M. Kawato. *Neural Computation* 14, 1347 (2002).

S15. H. Miyamoto, J. Morimoto, K. Doya, M. Kawato. *Neural Networks* 17, 299 (2004).

S16. G. Ainslie. *Picoeconomics* (Cambridge Univ Press, 1992).

S17. J. Mazur. *Animal Learning and Behavior* 25, 131 (1997).

S18. J. E. Mazur. *Psychological Review* 108, 96 (2001).

S19. R. A. Rescorla, A. R. Wagner. In *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokesy, eds. (Appleton Century Crofts, New York, 1972), pp. 64–99.

S20. R. Bellman. *Quarterly Journal of Applied Mathematics* 16, 87 (1958).

S21. J. C. Houk, J. L. Davis, D. G. Beiser, eds. *Models of Information Processing in the Basal Ganglia* (MIT Press, Cambridge MA, 1995).

S22. A. G. Barto. In *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, eds. (MIT Press, Cambridge MA, 1995), pp. 215–232.

S23. J. C. Houk, J. L. Adams, A. G. Barto. In *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, eds. (MIT Press, Cambridge MA, 1995), pp. 249–270.

S24. W. Schultz, R. Romo, T. Ljungberg, J. Mirenowicz, J. R. hollerman, *et al.* In *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, eds. (MIT Press, Cambridge MA, 1995), pp. 233–248.

S25. P. R. Montague, P. Dayan, T. J. Sejnowski. *Journal of Neuroscience* 16, 1936 (1996).

S26. W. Schultz, P. Dayan, R. Montague. *Science* 275, 1593 (1997).

S27. T. Ljungberg, P. Apicella, W. Schultz. *Journal of Neurophysiology* 67, 145 (1992).

S28. W. Schultz. *Journal of Neurophysiology* 80, 1 (1998).

S29. W. Schultz. *Neuron* 36, 241 (2002).

S30. W. Schultz. *Current Opinion in Neurobiology* 14, 139 (2004).

S31. S. C. Tanaka, K. Doya, G. Okada, K. Ueda, Y. Okamoto, *et al. Nature Neuroscience* 7, 887 (2004).

S32. M. A. Arbib, P. F. Dominey. In *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, eds. (MIT Press, Cambridge MA, 1995), pp. 149–162.

S33. D. Joel, Y. Niv, E. Ruppin. *Neural Networks* 15, 535 (2002).