Reinforcement learning models of the dopamine system and their behavioral implications

Nathaniel D. Daw

August 2003
CMU-CS-03-177

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Thesis Committee:
David S. Touretzky, Chair
James L. McClelland
Andrew W. Moore
William E. Skaggs, University of Arizona
Peter Dayan, University College London

This research was sponsored by a National Science Foundation (NSF) Graduate Research Fellowship and other NSF grants: IRI-9720350, IIS-9978403, and DGE-9987588. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the sponsor or any other entity.

Keywords: computational neuroscience, dopamine, reinforcement learning

Abstract

This thesis aims to improve theories of how the brain functions and to provide a framework to guide future neuroscientific experiments by making use of theoretical and algorithmic ideas from computer science. The work centers around the detailed understanding of the dopamine system, an important and phylogenetically venerable brain system that is implicated in such general functions as motivation, decision-making and motor control, and whose dysfunction is associated with disorders such as schizophrenia, addiction, and Parkinson's disease. A series of influential models have proposed that the responses of dopamine neurons recorded from behaving monkeys can be identified with the error signal from temporal difference (TD) learning, a reinforcement learning algorithm for learning to predict rewards in order to guide decision-making.

Here I propose extensions to these theories that improve them along a number of dimensions simultaneously. The new models that result eliminate several unrealistic simplifying assumptions from the original accounts; explain many sorts of dopamine responses that had previously seemed anomalous; flesh out nascent suggestions that these neurophysiological mechanisms can also explain animal behavior in conditioning experiments; and extend the theories' reach to incorporate proposals about the computational function of several other brain systems that interact with the dopamine neurons.

Chapter 3 relaxes the assumption from previous models that the system tracks only short-term predictions about rewards expected within a single experimental trial. It introduces a new model based on average-reward TD learning that suggests that long-run reward predictions affect the slow-timescale, tonic behavior of dopamine neurons. This account resolves a seemingly paradoxical finding that the dopamine system is excited by aversive events such as electric shock, which had fueled several published attacks on the TD theories. These investigations also provide a basis for proposals about the functional role of interactions between the dopamine and serotonin systems, and about behavioral data on animal decision-making.

Chapter 4 further revises the theory to account for animals' uncertainty about the timing of events and about the moment-to-moment state of an experimental task. These issues are handled in the context of a TD algorithm incorporating partial observability and semi-Markov dynamics; a number of other new or extant models are shown to follow from this one in various limits. The new theory is able to explain a number of previously puzzling results about dopamine responses to events whose timing is variable, and provides an appropriate framework for investigating behavioral results concerning variability in animals' temporal judgments and timescale invariance properties in animal learning.

Chapter 5 departs from the thesis' primary methodology of computational modeling to present a complementary attempt to address the same issues empirically. The chapter reports the results of an experiment that record from the striatum of behaving rats (a brain area that is one of the major inputs and outputs of the dopamine system), during a task designed to probe the functional organization of decision-making in the brain. The results broadly support the contention of most versions of the TD models that the functions of action selection and reward prediction are segregated in the brain, as in "actor/critic" reinforcement learning systems.

Contents

1 Introduction **1**
- 1.1 Introduction and motivation . . . 1
- 1.2 On modeling . . . 2
- 1.3 Organization of the thesis . . . 2

2 Related work **5**
- 2.1 Related Work: Reinforcement Learning . . . 5
  - 2.1.1 Reinforcement learning: background and notation . . . 5
  - 2.1.2 Reinforcement learning: policy evaluation/reward prediction . . . 6
  - 2.1.3 Reinforcement learning: policy improvement/action selection . . . 8
  - 2.1.4 Reinforcement learning: returns and temporal horizons . . . 9
  - 2.1.5 Reinforcement learning: timescales and temporal variability . . . 12
  - 2.1.6 Reinforcement learning: partial observability and non-Markovian contingencies . . . 13
- 2.2 Related work: neurophysiology and modeling . . . 15
  - 2.2.1 Neurophysiology and modeling of the dopamine system . . . 15
  - 2.2.2 Neurophysiology and modeling of the striatum . . . 21
  - 2.2.3 Neurophysiology of the serotonin system . . . 25
- 2.3 Related work: Behavioral experiment and theory . . . 26
  - 2.3.1 Classical conditioning . . . 27
  - 2.3.2 Opponency . . . 30
  - 2.3.3 Instrumental conditioning: free operant tasks . . . 32
  - 2.3.4 Instrumental conditioning: discrete choice tasks . . . 34
  - 2.3.5 Timing and timescale . . . 37

3 Returns and temporal horizons **41**
- 3.1 Background . . . 41
- 3.2 Previous models . . . 43
- 3.3 Dopamine and timescales of response . . . 44
  - 3.3.1 An average-reward TD model of the dopamine signal . . . 44
  - 3.3.2 The behavior of the model under an infinite temporal horizon . . . 46
  - 3.3.3 Discussion: Experimental evidence related to this model . . . 49
- 3.4 Opponency and serotonin . . . 53
  - 3.4.1 Opponency between timescales in the model of Solomon and Corbit (1974) . . . 54
  - 3.4.2 Negative error and opponency between appetitive and aversive systems . . . 55
  - 3.4.3 Model of serotonin as a dopaminergic opponent . . . 56
  - 3.4.4 Results of simulations of the model . . . 57
  - 3.4.5 Discussion: Experimental evidence on serotonergic firing . . . 60
  - 3.4.6 Discussion: The hypothesized dopamine response and rectification of negative error . . . 60
  - 3.4.7 Discussion: Conditioned inhibition and opponent value representation . . . 62
- 3.5 Behavioral choice experiments . . . 65
  - 3.5.1 Theoretical approaches to temporal discounting . . . 65
  - 3.5.2 A model of choice . . . 66

3.5.3 Results and discussion: discounting behavior . . . 67
3.5.4 Results and discussion: Risk sensitivity . . . 68
3.6 Discussion and open issues . . . 70

4 Uncertainty, timescale, and state representation 73
4.1 Relevant data . . . 74
4.1.1 Physiological data . . . 74
4.1.2 Behavioral data . . . 78
4.2 Previous models . . . 78
4.2.1 The model of Houk et al. (1995) . . . 79
4.2.2 The model of Montague et al. (1996; Schultz et al., 1997): overview . . . 79
4.2.3 The model of Montague et al. (1996; Schultz et al., 1997): physiology . . . 80
4.2.4 The model of Montague et al. (1996; Schultz et al., 1997): behavior . . . 81
4.2.5 The model of Suri and Schultz (1998, 1999): overview . . . 82
4.2.6 The model of Suri and Schultz (1998, 1999): physiology . . . 82
4.2.7 The model of Suri and Schultz (1998, 1999): behavior . . . 84
4.3 A new TD model of the dopamine signal . . . 84
4.3.1 Value learning in a broader functional context . . . 85
4.3.2 A fully observable semi-Markov model of the dopamine response . . . 87
4.3.3 A partially observable semi-Markov model of the dopamine response . . . 88
4.3.4 Modeling dopamine responses: inference models . . . 90
4.3.5 Modeling dopamine responses: vector versus scalar signaling . . . 91
4.3.6 Modeling dopamine responses: rectification of negative error . . . 92
4.3.7 Modeling dopamine responses: the effect of timing noise . . . 92
4.3.8 Limiting cases of the partially observable semi-Markov TD model . . . 93
4.4 Dopamine responses in the model . . . 94
4.4.1 Results: Free reward delivery . . . 95
4.4.2 Discussion: Free reward delivery . . . 98
4.4.3 Results: Signaled reward, overtraining, and timing noise . . . 98
4.4.4 Discussion: Signaled reward, overtraining, and timing noise . . . 100
4.4.5 Results: State inference and variation in event timing . . . 100
4.4.6 Discussion: State inference and variation in event timing . . . 104
4.4.7 Results: Tonic responding . . . 105
4.4.8 Discussion: Tonic responding . . . 108
4.5 Behavior: Timescale invariance in the model . . . 110
4.5.1 Scalar variability in timed responses . . . 110
4.5.2 Timescale invariance in acquisition . . . 111
4.6 Discussion and open issues . . . 113
4.6.1 Gaps in the present work . . . 114
4.6.2 The dopamine response magnitude . . . 115
4.6.3 Implications of the internal model . . . 116
4.6.4 Implications for experiment and comparison of Markov and semi-Markov models . . . 117
4.7 Appendix: Mathematical formulae and derivations . . . 118

5 The organization of reinforcement learning in the striatum 121
5.1 Introduction . . . 121
5.2 Experimental methodology . . . 122
5.2.1 Behavioral task . . . 122
5.2.2 Surgery . . . 124
5.2.3 Recordings . . . 124
5.2.4 Histology . . . 124
5.2.5 Data analysis . . . 126
5.3 Results: behavior . . . 128
5.4 Results: neurophysiological recordings . . . 130

5.5 Discussion 140

6 Concluding remarks **143**
6.1 Future directions 143
6.2 Summary of contributions 144
6.3 Summary of modeling results 147

Acknowledgments

I am very grateful for many crucial contributions from a variety of collaborators and mentors over the years. The members of my thesis committee have each provided important advice and guidance. They are David Touretzky (the committee chair and my advisor), Jay McClelland, Andrew Moore, and Peter Dayan and Bill Skaggs (external members). I thank Dr. Touretzky for six years of training, support, and guidance about all aspects of my scientific career, and not least about the various research projects that make up this thesis. Drs. Dayan and Skaggs have also gone particularly far beyond the call of duty, at times having been like second (and third) thesis advisors to me. In particular, I spent several summers in London visiting Dr. Dayan and his then-graduate student Sham Kakade; some of the direct fruits of our collaboration are described in Section 3.4, though they each additionally provided influential input about nearly every part of this thesis. I carried out the experiment described in Chapter 5 while visiting Dr. Skaggs' lab. It goes without saying that without his extensive guidance, training and assistance, I would have had none of the skills necessary to carry out such a project. In addition, Judith Joyce Balcita-Pedicino in his lab provided a great deal of technical assistance with the experiment, including performing all of the drive construction, surgery, and histology. Beata Jarosiewicz also helped with a variety of issues in the lab, and I worked with Mark Albert and Vikrant Kapoor on various parts of the project.

I have been very fortunate to collaborate with Aaron Courville. His ideas and input have greatly shaped my thinking about the problems considered in this thesis, and his contributions are particularly evident in Chapter 4, which builds on work we did together, and some parts of which are neurophysiological extensions of theoretical ideas he had developed in the behavioral domain. I have also collaborated on various projects with Rudolf Cardinal, Barry Everitt, Trevor Robbins, Mark Gluck, Geoff Gordon, Samuel McClure, and P. Read Montague. While that work is not directly represented in this thesis, all of it influenced subsequent work that is included here. I would also like to acknowledge the helpful input of Peter Shizgal, Rich Sutton, Andy Barto, Daphna Shohamy, David Redish, Mark Fuhs, Christopher Fiorillo, Wolfram Schultz, and Mark Wightman.

A number of people have generously provided advance access to unpublished data and theory that have helped shape this thesis, among them Christopher Fiorillo, Wolfram Schultz, Hannah Bayer, Paul Glimcher, Samuel McClure, P. Read Montague, David Redish, and Randy Gallistel.

Throughout my Ph.D., I was funded by the National Science Foundation, through a Graduate Research Fellowship and also by grants IRI-9720350, IIS-9978403, and DGE-9987588.

Chapter 1

Introduction

1.1 Introduction and motivation

This thesis aims to improve theories of how the brain functions and to provide a framework to guide future neuroscientific experiments by making use of theoretical and algorithmic ideas from computer science. Such an approach is not new; it has so far been pursued most successfully in sensory neuroscience and psychology, particularly vision and audition, where the dominant framework guiding research is derived from computational theories about signal processing and representation (e.g. Rieke et al., 1997, to choose one recent example). Examples of this approach include physiological experiments that treat visual neurons as linear filters, and related work that attempts to explain the measured characteristics of these filters in terms of computational theories of optimal image representation (e.g. Lewicki and Olshausen, 1999).

This thesis attempts to apply a similar program to the more abstract area of conditioning and its neural substrates. Conditioning is the study of how animals learn to predict significant events and select actions, which relates to computational ideas about optimal control, statistical prediction, and reinforcement learning. Strictly behavioral theories of animal conditioning have sporadically met with reinforcement learning; indeed, the temporal difference learning algorithm was originally conceived in part as a theory of Pavlovian conditioning (Barto and Sutton, 1982). But only recently has a convincing connection been drawn between reinforcement learning and the machinery in the brain that might implement it. This work, consisting of recordings from dopamine neurons made in the lab of Schultz (reviewed in Schultz, 1998), and their subsequent interpretation in terms of temporal difference learning by Montague et al. (1996) and others, is the foundation for the present thesis. The simple computational theories developed to date do a remarkably good job explaining the responses of dopamine neurons during appetitive conditioning, but they can also be seen as opening the door to a significant future program of relating brain systems and conditioning behavior to the theoretical framework of reinforcement learning, much as sensory brain systems and behaviors are now understood in signal processing terms. Particularly exciting is the promise, not yet fully realized, that the same theory could bridge both physiological and behavioral levels, providing a unified, normative account both of the neuronal responses in dopaminergic and related brain systems and of how the computations they carry out influence learning behavior.

My own contribution to this program is to investigate how a number of improvements in the computational theory can correspondingly advance the scientific explanations. In particular, this thesis considers three theoretical issues that have been studied in artificial reinforcement learning but so far ignored in the neuroscientific literature. These are the temporal horizon of predictions, variability in the timing between events, and partial observability of the state of the world. For each, I show how incorporating a solution improves the theory's explanations both of neuronal responses and of animal behavior, explaining a number of seemingly paradoxical neuronal responses and mending some mismatches between dominant theoretical ideas at physiological and behavioral levels. Key throughout all this will be issues of representation and timing, problems for which modelers have a number of potential theoretical approaches at their disposal, but have so far found few empirical constraints in the scientific data that would help adjudicate between them. As a result, these issues have so far received a treatment that is speculative compared to other aspects

of models that are more firmly constrained by data. In a final section, I discuss a fourth theoretical issue, that of the relation between prediction and action selection, which has received rather speculative attention in the modeling literature but for which the experimental foundations are weak. I report the results of an experiment designed to probe the anatomical organization of prediction and action selection.

1.2 On modeling

I wish here to clarify the general goals of this work and its contributions to computer science. The method of this thesis is to treat the task of explaining scientific data as an application for computational techniques not unlike more traditional computational applications such as robot vision or speech recognition. In this thesis my goal is to construct explanations of scientific data, drawing on tools that have been well studied in the artificial reinforcement learning community, and combining them in some novel ways.

It may be useful to consider various notions about the role of modeling in neuroscience, and how the present work fits within them. Marr (1982) famously distinguished three levels of neuroscientific modeling: the computational, algorithmic and implementational. The modeling in this thesis cuts across all three of these levels, though in some cases I develop computational and algorithmic models (i.e. theories aimed at understanding what computations the brain is trying to carry out and how, algorithmically, it accomplishes them), without suggesting any specific neural implementation. It is also common to subdivide implementational models further by their granularity of description, contrasting for instance theories about the interactions between different brain "systems" or modules versus those describing chemical or genetic events within neurons. To the extent this thesis presents implementational ideas, they remain on the systems level and eschew the lower-level biology.

We might also contrast *computational* modeling of the sort described in this thesis — theorizing about how the brain performs computations and represents information — against *mathematical* modeling of the sort common in other physical sciences (Dayan, 1994). In mathematical modeling, computational methods are used to simulate physical systems, but there is no notion that the systems are carrying out a computation. A related distinction is between *top-down* and *bottom-up* modeling — in the former, one starts with a computational goal and reasons about how the brain might be carrying it out; in the latter, one constructs a mathematical simulation of brain tissue and then performs "experiments" studying how this clockwork brain behaves. Here I am engaged in top-down modeling — constructing functional theories about how the brain computes — and not in building a detailed replica of brain activity.

Interactions between computer science and the biological sciences can go both ways — computational insights can inform biological theory-making, and insights from biology can also provide inspiration for solving engineering problems. Here, I am engaged in the former function — leveraging the relatively sophisticated machinery for prediction and control from computer science into a theory of brain function, rather than trying to use biological insight to guide novel algorithmic or practical work in reinforcement learning.

I am also, emphatically, not trying to present a single or definitive theory of brain function, nor even of brain function in conditioning. Instead, I study a family of related models, each of which has advantages and disadvantages in terms of explaining the data that are presently available (which are sparse and, on present understanding, sometimes seemingly inconsistent). A related point is that constructing scientific theories requires a great deal of abstraction in order to expose the core ideas that are supposed to explain the data, and the clearest understanding often comes when this core is stripped as bare as possible. In part for this reason, I take a gradual and componential approach to theory-building, isolating each small change to the theory and studying how it relates to some small body of data, rather than trying to assemble a single, giant theory to account for everything. One reason this approach works well is that it mirrors the way scientific experiments tend to be organized.

1.3 Organization of the thesis

The overall goal of this thesis is to advance theories of the dopamine system, an important and phylogenetically venerable brain system that is implicated in such general functions as motivation, decision-making and motor control, and whose dysfunction is associated with disorders such as schizophrenia, addiction, and

Parkinson's disease. The thesis is organized into chapters corresponding to three broad sets of theoretical issues in dopaminergic theories: the optimization of a return, uncertainty and timing, and the organization of action selection. Each of these theoretical issues has implications that extend existing theories in a number of different directions, and each thesis chapter thus touches on all of these directions to some extent. First, computationally, many simplifying assumptions are eliminated in order to produce a more sophisticated theory that copes with a number of complexities that had been previously ignored. Second, physiologically, the model's account of the behavior of dopamine neurons is made more faithful, explaining a number of previously anomalous findings. Also, the thesis extends the physiological reach of the theory by leveraging our relatively strong understanding of the dopamine signal into hypotheses about the function of some related brain systems. Finally, psychologically, this theory of brain function is brought into contact with bodies of behavioral data and theory that had previously been more abstract.

Chapter two reviews related work in three broad areas: computational work in artificial reinforcement learning, neuroscientific investigations of the dopamine system and related brain systems, and experiments and theory about animal conditioning behavior.

Chapters three and four focus on theoretical issues poorly addressed by existing dopamine system models: respectively, prediction horizons, and uncertainty about the the state of the world and about the timing of events. These chapters share a common organization. Each first lays out versions of the dopamine model that cope with the theoretical problem, then demonstrates several examples of how the new features enrich our understanding of both physiological and behavioral data, by explaining existing data, making new predictions for experiments that have yet to be performed, and forging new connections with existing, but previously disjoint, bodies of theory. Though their discussions are complementary in many respects, the two chapters also lay out in detail the expected behavior of the dopamine system under two rather different hypotheses about how the system should account for the costs of delays, which represent both extremes of a spectrum of possibilities. The data do not yet exist to decide between these alternatives, but a major contribution of the present work is to note the existence of these alternatives, flesh out their implications in detail, and suggest what experiments and analyses can be done to adjudicate between them.

Chapter five considers a fourth theoretical issue — mechanisms for action selection — under a more empirically focused treatment. In particular, the chapter reports experimental results from my own recordings of neurons in the rat striatum, a brain area that is an important input and output of the dopamine system and involved in many of the same functions. These data are discussed in terms of computational ideas about the anatomical organization of reinforcement learning in the brain.

Chapter 2

Related work

2.1 Related Work: Reinforcement Learning

This section reviews work in reinforcement learning. After some tutorial material introducing basic notation and algorithms, the discussion focuses specifically on work related to the theoretical issues important to the scientific data considered in this thesis. More complete overviews of reinforcement learning have been published in both book (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) and article form (Kaelbling et al., 1996), and I draw on all of these in what follows.

2.1.1 Reinforcement learning: background and notation

The central problem for reinforcement learning is optimal control: learning to choose actions that optimize some reward or cost metric. Typically, the control problem is modeled as a Markov decision process (MDP). Such a process consists of two functions, $\mathbf{R}$ and $\mathbf{T}$, defined over two sets, a set $\mathcal{S}$ of states and a set $\mathcal{A}$ of actions. The world, under this model, evolves stochastically under a simple, discrete temporal dynamics. At time t, the world is in some state, denoted by the random variable $s_t \in \mathcal{S}$. The agent chooses some action $a_t \in \mathcal{A}$, and the world transitions into some new state s_{t+1} at the next timestep. Given the values of the random variables s_t and a_t, the transition function $\mathbf{T}$ specifies a probability distribution over the successor state s_{t+1}; I write the probability $P(s_{t+1} = s' | s_t = s, a_t = a)$ as $\mathbf{T}_{s,a,s'}$. At each timestep the agent is also assessed some real-valued reward or cost r_t; the probability distribution from which this is drawn is specified by the reward function $\mathbf{R}$, and I similarly abbreviate $P(r_t = r | s_t = s, a_t = a)$ as $\mathbf{R}_{s,a,r}$. The fact that the transition and reward probabilities can be specified this way is known as the Markov property, from which the processes take their name. That is: the future behavior of the system after time t is wholly specified (up to the limits of the stochasticity of the process) by the state and action at time t; knowledge of previous states contributes no further information.

We can define a *Markov policy*, π, (usually just called a "policy") as a function mapping states to actions. Alternatively, it can map states to probabilities of actions; the mapping need not be deterministic. For deterministic policies, we write the action chosen in state s as π_s; for probabilistic policies, $\pi_{s,a}$ is the probability that action a is chosen in state s. The goal of reinforcement learning is to discover a policy that is optimal in a formal sense, which we need some further machinery to define.

If we fix some policy π, an MDP is reduced to a Markov chain. That is, with the agent's action probabilities for each state chosen in advance, we can think of the world as evolving on its own, using transition and reward functions that now depend only on the state (with probabilities $\mathbf{T}_{\pi,s,s'} = \sum_{a \in \mathcal{A}} \pi_{s,a} \cdot \mathbf{T}_{s,a,s'}$ and $\mathbf{R}_{\pi,s,r} = \sum_{a \in \mathcal{A}} \pi_{s,a} \cdot \mathbf{R}_{s,a,r}$). We can thus define another function, the value function $\mathbf{V}_{\pi,s}$, which provides some measure of expected future reward (known as a "return" or a "currency"), given that the process started in s and the agent is following the policy π. When a particular policy π is fixed, I often abbreviate the notation by omitting references to it, writing the functions as $\mathbf{V}_s$, $\mathbf{T}_{s,s'}$, and $\mathbf{R}_{s,r}$ rather than $\mathbf{V}_{\pi,s}$, $\mathbf{T}_{\pi,s,s'}$, and $\mathbf{R}_{\pi,s,r}$.

Let us also define a notion of *termination* for a Markov process. In MDPs, we can distinguish some subset

of states as *absorbing states*: those that deliver reward 0 and transition only to themselves, thus truncating the value sum at the time when they are entered. For now, let us restrict attention to *absorbing* Markov processes, those in which we are guaranteed to eventually reach an absorbing state (true if from every state, under every policy, there is a nonzero chance of reaching an absorbing state within some finite number of timesteps; c.f. Bertsekas and Tsitsiklis, 1996, sec. 2.2.1).

The simplest measure of long-run return is cumulative expected future reward, which is finite in an absorbing Markov process:

$$\mathbf{V}_{s_t} = \sum_{\tau \geq t} E\left[r_\tau | s_t\right] \tag{2.1}$$

where the expectation is over randomness in state transitions and reward delivery, i.e. $E[r_\tau|s_t] = \sum_s \sum_r r \cdot \mathbf{R}_{s,r} \cdot P(s_\tau = s|s_t)$, whose last term follows from repeated application of the transition function $\mathbf{T}$. We can explicitly expand the expectation over states while rewriting the function recursively:

$$\begin{aligned} \mathbf{V}_{s_t} &= E_r[\mathbf{R}_{s_t}] + E_{s_{t+1}}[\mathbf{V}_{s_{t+1}}] \\ &= E_r[\mathbf{R}_{s_t}] + \sum_{s_{t+1} \in \mathcal{S}} \mathbf{T}_{s_t, s_{t+1}} \mathbf{V}_{s_{t+1}} \end{aligned}$$

This recursive formula for the value function is known as the Bellman equation for $\mathbf{V}$, (after Bellman, 1957, who introduced an analogous equation for the value of the optimal policy) and some form of it underlies every reinforcement learning algorithm we will consider. We can write it more compactly in vector form:

$$\mathbf{V} = E[\mathbf{R}] + \mathbf{T}\mathbf{V} \tag{2.2}$$

where I have assumed $\mathbf{V}$ and $E[\mathbf{R}]$ are expressed as n-vectors (for n states) containing the value and $\mathbf{T}$ is expressed as an $n \times n$ matrix.

We can now define an optimal policy as one that maximizes the value function, that is some π^* such that for all other policies π and all states s, $\mathbf{V}_{\pi^*,s} \geq \mathbf{V}_{\pi,s}$. Many algorithms for finding optimal policies rely on separate steps of *policy evaluation* (computing V_π for candidate policies π) and *policy improvement* (updating π in light of the evaluation). A great deal of the modeling in this thesis will concentrate on how the brain solves the first, narrower problem of policy evaluation: reward prediction in the service of action selection. I will discuss policy evaluation now, and return to policy improvement in section 2.1.3.

2.1.2 Reinforcement learning: policy evaluation/reward prediction

Given the functions $\mathbf{T}$ and $\mathbf{R}$ — referred to as a *model* of the process — and a fixed π, the value function $\mathbf{V}_\pi$ can be recovered in a number of ways. (These approaches are known as *model-based* because they rely on explicit representations of $\mathbf{T}$ and $\mathbf{R}$.) Bellman's equation is really a system of linear equations (one for each state) which can be solved directly for the value function. When the number of states is large, making the equations' direct solution unwieldy, $\mathbf{V}$ can also be iteratively estimated using a form of the value iteration algorithm (Bellman, 1957). (Throughout this thesis, I use $\widehat{\mathbf{V}}$ to denote an estimate of the value function computed during some algorithm, to distinguish it from the target function $\mathbf{V}$.) First, initialize $\widehat{\mathbf{V}}_\pi^0$ to a vector of zeros, then repeatedly update each state's value using the Bellman equation:

$$\widehat{\mathbf{V}}_\pi^{n+1} \leftarrow E[\mathbf{R}] + \mathbf{T}\widehat{\mathbf{V}}_\pi^n \tag{2.3}$$

The correctness of this algorithm follows from the fact that it directly implements the function definition.[1] The value estimates produced by this algorithm are truncated versions of the sum in equation 2.1, with the n_{th} estimate $\widehat{\mathbf{V}}_\pi^n$ containing the first n terms of the sum.

The dopamine models under consideration here are all based on an asynchronous, stochastic, and *model-free* version of this algorithm, called temporal-difference (TD) learning (Sutton, 1988). An agent acting in an

[1] This reasoning applies to the *synchronous* form of value iteration as written here, in which the $n+1st$ value of every state is computed from the nth value of its successor states. It turns out an *asynchronous* form of value iteration also converges. In this version, each state is updated individually using whatever are the latest available values of its successors.

unknown environment does not, in general, have access to a model of the environmental contingencies. But the agent can observe a series of states $s_1...s_t$ and rewards $r_1...r_t$ that are *samples* of the random variables of the Markov process. One approach would be to use these samples to learn a model, that is to estimate the functions $\mathbf{T}$ and $\mathbf{R}$, and then solve as above for the value function. It is possible to forgo this intermediate modeling step and use the samples with various Monte Carlo methods to directly estimate the values. For instance, we could sum up all of the rewards over an entire sample trajectory, starting in state s, repeat this process many times and average the sums to estimate $\mathbf{V}_s$. This method is related to one case of the family of temporal difference algorithms, known as TD(1) (which will be discussed later), and to Monte Carlo methods for matrix inversion (Barto and Duff, 1994). Here I will discuss another special case of TD, TD(0), in which the samples are used to stochastically estimate the expectations of terms in the Bellman equation. Thus, unlike the very simple Monte Carlo scheme described above, there is no need to accumulate sums of sampled rewards over multiple timesteps. Assume we have some estimate of the value function $\widehat{\mathbf{V}}$, and we observe a transition from state s_t to s_{t+1} with reward r_t. Using r_t as a sample of $E_r[\mathbf{R}_{s_t}]$ and $\widehat{\mathbf{V}}_{s_{t+1}}$ as an estimated sample of $E_{s_{t+1}}[\mathbf{V}_{s_{t+1}}]$ in Bellman's equation, we can estimate a sample of the value $\mathbf{V}_{s_t}$ as $r_t + \widehat{\mathbf{V}}_{s_{t+1}}$. The difference between this estimate and our current estimate $\widehat{\mathbf{V}}_{s_t}$ is the *temporal-difference error* δ:

$$\delta_t = r_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \tag{2.4}$$

Assuming (as I often do in this thesis) that $\widehat{\mathbf{V}}$ is simply stored as a vector of values, one for each state, then we improve the value function estimate by nudging $\widehat{\mathbf{V}}_{s_t}$ in the direction that reduces the error:

$$\widehat{\mathbf{V}}_{s_t} \leftarrow \widehat{\mathbf{V}}_{s_t} + \nu\delta_t$$

for some learning rate ν. Under some technical assumptions about the learning rate schedule and the structure of the MDP (e.g., all states must be sampled infinitely often) this algorithm converges (Dayan, 1992). It is also possible to use arbitrary function approximation schemes in place of the table-lookup representation of $\widehat{\mathbf{V}}$. The algorithm converges (to some approximation of $\mathbf{V}$ whose error can be quantified) for linear function approximation (Bertsekas and Tsitsiklis, 1996); there are examples in which nonlinear approximators can be shown to diverge.

I have so far described the TD(0) algorithm, a special case of a more general algorithm known as TD(λ) (Sutton, 1988). This rests on the observation that the Bellman equation can easily be generalized to an equation relating the values of two states separated by any temporal delay. For instance, we can write a Bellman equation in which the recursion is unrolled by two timesteps instead of the customary one. In the discounted case, this is:

$$\mathbf{V} = E[\mathbf{R}] + \mathbf{T}E[\mathbf{R}] + \mathbf{T}^2\mathbf{V} \tag{2.5}$$

An analogous Bellman equation can can be written for any deeper unrolling; all these equations have the same fixed points, and they can all be used interchangeably (or even intermixed) in suitably modified versions of the algorithms we've described.

For instance, based on equation 2.5, we could define a TD algorithm in which after observing the state/reward sequence $s_1/r_1/s_2/r_2/s_3$, we nudge the value of s_1 toward $r_1 + r_2 + \widehat{\mathbf{V}}_{s_3}$. Similarly, we could use any n-timestep unrolled Bellman equation to define a TD algorithm that estimates the value of a visited state using an *n-timestep* backup: the sum of n subsequent rewards plus the value of the nth subsequent state. The idea of TD(λ) (Sutton, 1988; see also Watkins, 1989) is to blend TD updates involving all different numbers of timesteps backed up. If we weight an n-timestep backup exponentially in n (for some base $\lambda < 1$), then this turns out to be easily accomplished by maintaining a parameter known as an *eligibility trace* for each state, which is incremented whenever that state is visited and decayed exponentially (by λ) thereafter. One-timestep TD errors δ are computed as usual, but are applied at each timestep to each state, weighted by its eligibility. Apart from some details of batch versus ongoing updates, the net effect of this process can easily be shown to be equivalent to individually applying n-timestep backups for all different values of n, weighted by λ^n. I omit further detail, as TD(0) is sufficient for the modeling purposes of this thesis, but early work in reinforcement learning devoted a great deal of attention to the more general algorithm and to the effects of different choices of λ; for pointers, see Sutton and Barto (1998).

2.1.3 Reinforcement learning: policy improvement/action selection

Given knowledge of value functions, there are several related schemes for selecting actions. The basic intuition is that, insofar as the value function represents information about all future rewards from a state, it can guide an action selection process toward states with high expected future payoff, without requiring deep search through trees of future states. In particular, given the value of the *optimal* policy, $\mathbf{V}_{\pi^*}$ and a model of the process, it is easy to recover the optimal action π^*_s at each state with a one step lookahead:

$$\pi^*_s = \text{argmax}_{a \in \mathcal{A}}(\sum_r \mathbf{R}_{s,a,r} + \sum_{s' \in \mathcal{S}} \mathbf{T}_{s,a,s'} \mathbf{V}_{\pi^*,s'})$$

It is possible to solve directly for the optimal values $\mathbf{V}_{\pi^*}$ from a model using a value iteration method, but a slightly different model-based approach, known as *policy iteration*, is more directly extensible to the online, model-free case. Given some candidate policy π and its value $\mathbf{V}_\pi$, with a search one step deep, we can find a new policy that optimizes $\mathbf{V}_\pi$:

$$\pi'_s = \text{argmax}_{a \in \mathcal{A}}(\sum_r \mathbf{R}_{s,a,r} + \sum_{s' \in \mathcal{S}} \mathbf{T}_{s,a,s'} \mathbf{V}_{\pi,s'})$$

It is obvious that the new policy is either the same or better than the old one at every state. It can be proven that, if this cycle of policy evaluation followed by policy improvement is repeated and if the state and action spaces are finite, it will find the optimal policy after a finite number of iterations (see e.g. Bertsekas and Tsitsiklis, 1996).

This algorithm is still not directly applicable to the model-free, online situation, since although TD methods can be used for the policy evaluation step, the policy improvement step requires knowledge of the transition and reward functions. But policy improvement can be approximately handled in the online case using gradient methods and an algorithm known as actor/critic (Barto et al., 1983; Sutton, 1984; derived in the form presented here by Dayan and Abbott, 2001). We must assume some parameterization of the policy; for concreteness, assume a policy is stored as set of weights $m_{s,a}$. We define the probability that particular actions are taken in terms of the weights using the softmax function:

$$\pi_{s,a} = \frac{\exp(\beta m_{a,s})}{\sum_{a' \in \mathcal{A}} \exp(\beta m_{a',s})}$$

where $\pi_{s,a}$ is the probability of action a in state s and β is a temperature parameter. Note that we are now considering probabilistic rather than deterministic policies, to allow for small-step stochastic updates following the value gradient.

Now, given some policy π, its values $\widehat{\mathbf{V}}_\pi$, and sampled trajectories through the world, we can perform policy improvement with respect to $\widehat{\mathbf{V}}_{\pi,s}$ by stochastically following the gradient $d(E_r[\mathbf{R}_{\pi,s,r}] + E_{s'}[\widehat{\mathbf{V}}_{\pi,s,s'}])/dm$ using samples for the expectations. In particular, after the series of events starting in state s_t, receiving reward r_t, taking action a_t, and transitioning into state s_{t+1} the actor/critic update rule for the weights is:

$$m_{s_t,a} \leftarrow m_{s_t,a} + \xi(r_t + \widehat{\mathbf{V}}_{\pi,s_{t+1}} - c_{s_t})(\delta_{a,a_t} - \pi_{s_t,a_t})$$

for all $a \in \mathcal{A}$, where ξ is a learning rate parameter, c_{s_t} is an arbitrary state-dependent constant, and δ_{a,a_t} is the Kronecker delta (not the TD error). If we take $c_{s_t} = \widehat{\mathbf{V}}_{\pi,s_t}$ then this update is proportional to the TD error signal δ_t (equation 2.4) — i.e. the same error signal can be used both for policy evaluation and improvement.

This algorithm is just a small-step, stochastic version of the policy iteration algorithm so long as policy evaluation converges between each policy improvement step. But in actor/critic algorithms, policy evaluation and improvement are performed simultaneously, i.e. each sampled state transition is used to update both $\widehat{\mathbf{V}}$ and m. In this case, the algorithm has not been proved to converge, since the value estimates will generally lag the policies actually in use.

Actor/critic is the predominant action selection algorithm in the models considered in this thesis. I will quickly mention a couple of variations. First, it is possible to define a version of the value function that is parameterized over both states and actions. This function, $\mathbf{Q}_{\pi,s,a}$, measures the expected reward if action a

is taken in state s and then the policy π is followed thereafter. This function is most interesting when π is the optimal policy $\pi*$, in which case we write it as simply $\mathbf{Q}_{s,a}$. We can write a recursive Bellman equation for the optimal $\mathbf{Q}$-values:

$$\mathbf{Q}_{s,a} = E[r_{s,a}] + \sum_{s' \in \mathcal{S}} \mathbf{T}_{s,a,s'} \cdot max_{a' \in \mathcal{A}}(\mathbf{Q}_{s',a'})$$

A TD algorithm based on this recursion can be used to learn the function (Watkins, 1989; Watkins and Dayan, 1992). When it converges, the optimal policy can be determined from $\mathbf{Q}$ directly (for each state, the optimal action is the one that maximizes $\mathbf{Q}_{s,a}$). Conveniently, unlike traditional actor/critic TD for $\mathbf{V}$, Q-values learned in this manner are not dependent on the policy being carried out by the agent during learning.

We can also define yet another function, the *advantage* function. $\mathbf{A}_{\pi,s,a}$ is the expected reward gained by performing action a in state s rather than the action specified in the policy π (Baird, 1994; this simplified version of the algorithm is due to Dayan, 2002). This can be defined in terms of the difference between $\mathbf{Q}$ and $\mathbf{V}$ functions:

$$\mathbf{A}_{\pi,s,a} = \mathbf{Q}_{\pi,s,a} - \mathbf{V}_{\pi,s}$$

Since the TD errors are just samples of the advantages, the function can be estimated as:

$$\widehat{\mathbf{A}}_{s_t,a_t} \leftarrow (1-\xi)\widehat{\mathbf{A}}_{s_t,a_t} + \xi \cdot \delta_t$$

Given some method of translating advantages to action probabilities (e.g. softmax), this can be seen as a policy improvement rule for an actor/critic method that uses a more interpretable parameterization of its policy.

2.1.4 Reinforcement learning: returns and temporal horizons

The discussion thus far ignores a serious issue in defining optimal behavior: how do we measure optimality? So far, we have defined the return to be optimized as cumulative expected reward (Equation 2.1). This is only a meaningful notion of value if the sum converges, i.e. in an absorbing Markov process.

Given a problem without an obvious endpoint, one approach is to recast it as an *episodic* problem, by defining some set of states as endpoints for learning trials and then, whenever they are encountered, declaring a learning trial complete, and restarting learning afresh with their successors. But the criterion this approach measures, cumulative reward per trial, is not necessarily a sensible notion of optimality, since it fails to take into account trial length in addition to reward amount. That is, this criterion can disfavor strategies that increase the payoff rate by collecting less reward per trial while finishing the trial (and moving on to the next one) more quickly. This problem comes up explicitly in attempting to understand experiments in which animals choose between rewards of different amounts after different delays, as will be discussed in Chapter 3.

This reasoning suggests that a better measure of optimal behavior in continuous problems is the long-term reward rate, or more properly, the discrete analog of rate, average reward per timestep:

$$\rho_{s_t} = \lim_{n \to \infty} \frac{1}{n} \sum_{\tau=t}^{t+n} E[r_\tau] \tag{2.6}$$

There are some subtleties to this notion of optimality, having to do with distinguishing between policies whose differences affect only a finite prefix of the sum inside the limit, and hence have no effect on the long-run average reward. These are beyond the scope of this thesis but are the focus of a review by Mahadevan (1996) and are also covered by Puterman (1994). In any case, it is easy to see that for *finite recurrent* MDPs (those in which under all policies, all states have some nonzero probability of reaching all other states after some number of transitions), ρ_{s_t} is the same for all states. This is because starting in some s_{t+k} in place of s_t just truncates a finite number of terms from the sum in Equation 2.6, which has no effect on the limit. For the same reason, this property also holds true for the somewhat broader class of *unichain* MDPs, those for which, under every policy, there is only a single set of states that can all reach each other, but which is maximal in the sense that no state inside the set can ever reach any state outside the set. (The

idea is that the process must eventually enter one of these states, and then the average reward of this set of states controls the limit sum in Equation 2.6; for a fuller discussion, see Mahadevan, 1996) The result of this property is that the average reward ρ is not a useful value function in that it doesn't carry any information about which states are richer than others, information that could guide action selection. But it does represent a reasonable metric for judging policies.

One way to approximate the infinite horizon reward rate ρ might be to define a value function that measures cumulative reward over some fixed temporal window, say, the next 50 timesteps. If this window is long with respect to the mixing time of the MDP[2], this should be a reasonable approximation to ρ. This specific idea, however, fails since no recursive Bellman equation can be written for such a function (it equals immediate reward plus the expected sum of the next 49 rewards, which is not the same function). But we can instead define a "soft" window by *discounting* delayed rewards. Traditionally, rewards are discounted exponentially in their delays, by some parameter $\gamma < 1$, redefining the value function as:

$$\begin{aligned} \mathbf{V}_{s_t} &= E\left[\sum_{\tau \geq t} \gamma^{\tau - t} r_\tau\right] \\ &= E_r[\mathbf{R}_{s_t}] + \gamma \cdot E_{s_{t+1}}[\mathbf{V}_{s_{t+1}}] \end{aligned} \tag{2.7}$$

which is guaranteed to converge. The vector version of the Bellman equation is $\mathbf{V} = E[\mathbf{R}] + \gamma \mathbf{T}\mathbf{V}$, and the corresponding TD error signal is:

$$\delta_t = r_t + \gamma \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \tag{2.8}$$

Given these definitions, the reinforcement learning methods previously described work as before. Discounting effectively introduces a soft temporal horizon on the value function, with a timescale of roughly $1/(1-\gamma)$. Moreover, it can be proved that for any MDP, there is some threshhold discounting factor γ' such that for all $\gamma \geq \gamma'$, the policy that optimizes the exponentially discounted value function (Equation 2.7) will also optimize the long-run average reward (Equation 2.6). However, for any particular γ, it is also easy to construct an MDP in which optimizing the discounted value function produces a policy which is suboptimal with respect to long-run average reward.

Thus a discounted return can be used as a convenient device for finding policies that optimize the average reward. There are also situations in which a discounted return is itself normatively justified. For instance, exponential discounting is called for if rewards earn interest once received, or if there is a constant probability of $(1-\gamma)$ per timestep that future rewards will be lost altogether (e.g. if the process terminates: see the construction of a discounted MDP from an undiscounted one in Bertsekas and Tsitsiklis, 1996, sec. 2.3).

In the psychology literature, a different form of temporal discounting predominates. With hyperbolic discounting, rewards lose value proportionally to the reciprocal of their delays. This suggests a value function of the form:

$$\mathbf{V}_{s_t} = E\left[\sum_{\tau \geq t} \frac{r_\tau}{\tau}\right] \tag{2.9}$$

However, this function cannot be written recursively, so it is not amenable to a reinforcement learning treatment. Moreover, as I will discuss further (and see Kacelnik, 1997), the psychological theories envision a choice episode that terminates with a single reward, in which case Equation 2.9 just measures the average reward per timestep. Under this understanding, Equation 2.6, rather than 2.9, properly extends the return to the infinite horizon case.

There is a final return that will be important in this thesis, one which is undiscounted but has an infinite horizon. It is defined by the Bellman recursion:

$$\mathbf{V} = E[\mathbf{R} - \rho] + \mathbf{T}\mathbf{V}$$

[2]For a Markov process, we can define the *stationary distribution* over states as the distribution $\mathbf{s}$ that is invariant under state transition, i.e. satisfying $\mathbf{T}\mathbf{s} = \mathbf{s}$. We can then define a formal measure of the overall cycle time of the process, the *mixing time*. Starting in some state and running the process forward by some number of timesteps induces a distribution $\mathbf{s}'$ over the states; the mixing time is the time it takes for $\mathbf{s}'$ to come ϵ-close to the stationary distribution, for some tolerance ϵ and some measure of distance between distributions such as the Kullback-Leibler divergence

for some scalar ρ, i.e.

$$\mathbf{V}_{s_t} = E_r[\mathbf{R}_{s_t} - \rho] + E_{s_{t+1}}[\mathbf{V}_{s_{t+1}}] \tag{2.10}$$

which is known as the *average-adjusted*, or *relative* value function. For unichain MDPs, the solutions of this equation have ρ equal to the average reward per timestep (Equation 2.6); moreover, a policy which maximizes this value function will also maximize ρ (Puterman, 1994). Algorithms based on this value function are thus often called *average-reward* reinforcement learning algorithms. The solution of Equation 2.10 is not unique: for any vector of values $\mathbf{V}$ that satisfies it, so will $c + \mathbf{V}$ for any constant c. (This is unimportant, since what matters when choosing among actions is the *difference* between their expected outcome values.) With the recursion unrolled, the relative value function has the form

$$\mathbf{V}_{s_t} = c + E\left[\sum_{\tau \geq t} (r_\tau - \rho)\right] \tag{2.11}$$

Returns of this form were introduced to the reinforcement learning literature by Schwartz (1993), with his **R**-learning algorithm (a version of **Q**-learning based on the relative value function). However, the same ideas date back several decades in the study of dynamic programming (for recent overviews, see Puterman (1994) on dynamic programming and Mahadevan (1996) on reinforcement learning). Some refinements to Schwartz' scheme were suggested by Singh (1994), while the precise TD algorithm that will be at issue in this thesis was described and analyzed by Tsitsiklis and Van Roy (1999, 2002). The TD error corresponding to Equation 2.10 is:

$$\delta_t = r_t - \rho_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \tag{2.12}$$

where the average reward ρ is now time-dependent because it must be estimated online. For this, the authors use an exponentially windowed running average with learning rate σ:

$$\rho_{t+1} \leftarrow (1 - \sigma)\rho_t + \sigma r_t$$

There is clearly a close relationship between the exponentially discounted and average-adjusted returns. In an absorbing Markov process (i.e. one that is guaranteed to eventually enter an absorbing state), the undiscounted return obviously represents the limit of the exponentially discounted return as $\gamma \to 1$. In a process that is not absorbing, the average-adjusted return plays a similar role: In a certain formal sense (elaborated below) algorithms for learning the exponentially discounted return behave like algorithms for learning the average adjusted return in the limit as $\gamma \to 1$. The exponentially discounted return, Equation 2.7, can be rewritten:

$$\mathbf{V}_{s_t} = E\left[\sum_{\tau \geq t} \gamma^{\tau - t}(r_\tau - \rho)\right] + \frac{\rho}{1 - \gamma}$$

In this equation, the first term can be seen as a discounted approximation to the average adjusted return (Equation 2.10), whose error approaches zero as $\gamma \to 1$, while the second term is a constant (which increases with γ). Thus we can view the value of each state, under the exponentially discounted approach, as the sum of two values: a state-dependent measure of the state's average-adjusted value plus a *state-independent* encoding of the long-term reward rate, $\rho/(1 - \gamma)$. If $\mathbf{V}$ is linearly approximated (and a table lookup representation is a special case of this), it is reasonable that the shared baseline value $\rho/(1 - \gamma)$ should be associated with a common bias element in the encoding. Working from this idea, Tsitsiklis and Van Roy (2002) proved that TD algorithms based on exponential discounting and the average-adjusted return are equivalent on an *update-by-update* basis in the limit as $\gamma \to 1$. For the proof, it is assumed that both algorithms use the same linear representation of the states, but the average-adjusted version uses no bias element (which would in any case be superfluous because the value function is only defined up to a constant), while the exponentially discounted version contains a bias whose magnitude is specially chosen so that learning about its value mirrors the updates to ρ in the average-adjusted version.

2.1.5 Reinforcement learning: timescales and temporal variability

The simple temporal dynamics underlying the algorithms we have so far considered are unsatisfactory. The main problem is that the MDP formalism is a *discrete time* model controlling events at a single, fixed timescale, to which the algorithms I have discussed are rather deeply wedded. It seems obvious that learning and planning effectively in any realistic environment (and even *unrealistic* environments like the toy gridworld or the laboratory situations at issue in this thesis) will require the consideration of many different timescales. One such difficulty we have already discussed is choosing between different candidate returns accumulated over different timescales. Another problem is finding and executing policies that involve temporally extended sequences of action; simply building these up out of the smallest possible units by the methods described thus far is grossly intractable. For instance, while it is possible to treat driving a car formally as a series of choices taken every 100ms, using such a microscopic view alone would make planning hopeless; however, planning exclusively on a much coarser timescale would surely risk collision! A related issue more directly relevant to this thesis concerns *variability* in the timing between events. While it is possible to treat such variability formally by building up varying extended intervals out of different numbers of small Markov timesteps, such a device is awkward, complicates reasoning about temporal intervals, and interacts poorly with simple approximation methods one would otherwise be tempted to employ.

One early attempt to address some of these issues was due to Sutton (1995), who used model-based reinforcement learning based on multiply-unrolled Bellman equations like 2.6 on page 9 to model the world at different timescales. He notes that Equation 2.6 has exactly the same form as the usual Bellman equation, if we group its first two terms into a new two-timestep expected reward vector and define a new, two-timestep transition matrix $\mathbf{T}^2$. Thus these can be viewed as a slower-timescale model of the world dynamics that nonetheless defines the same value function as the one-timestep model. Sutton goes on to point out that analogous models can be built involving any timescale or quite general mixtures of timescales and gives algorithms (based on TD-λ) for doing so; agents equipped with a collection of such models could perhaps use them to plan at different timescales. This edifice of elegant theory does not seem to have translated into much in the way of practical applications, but see Precup and Sutton (1997, 1998) for further development.

A related cluster of attempts to address timescale issues relies on the same trick of using Bellman equations involving arbitrary time delays, but in a temporally more flexible formal model. This work is exemplified by two recent doctoral theses (Parr, 1998; Precup, 2000; see also Parr and Russell, 1998; Sutton et al., 1999) that sought to unify and extend previous work on planning with temporally extended or hierarchical actions under the framework of *semi*-Markov decision processes (SMDPs). This is another venerable tool from dynamic programming (see, e.g., Puterman, 1994 for a modern overview), which was only later imported to the reinforcement learning community (by Bradtke and Duff, 1995).

The idea is to replace the discrete-time MDP dynamics with discrete-*event* dynamics under which discrete state transitions occur irregularly in continuous time. These are determined by a third function, $\mathbf{D}$, which maps state/action pairs (or state/action/successor state triplets) to a probability distribution over the time dwelt in the first state before transitioning to the successor state (which is drawn in the usual way from $\mathbf{T}$). The process is known as semi-Markov because the chance of a transition depends not just on the state but on the time that has been spent there. In this thesis, I treat rewards as point events that occur on state transitions; in general, they are allowed to occur at some state- and action-dependent rate during the dwell time at a state. It can at times be useful to index random variables either by their time t in continuous time, or by a discrete index k that counts state transitions. For instance, if the system enters the kth state s_k and receives reward r_k at time τ, dwells in the state for d_k timesteps, and enters the next state s_{k+1} at time $\tau + d_k$, then we can also write that $s_t = s_k$ for all $\tau \leq t < \tau + d_k$ and $r_t = r_k$ for $t = \tau$ while $r_t = 0$ for $\tau < t < \tau + d_k$.

It is rather simple to adapt standard reinforcement learning algorithms to the richer formal framework, a task first tackled by Bradtke and Duff (1995). In the discounted case, the Bellman equation has the form:

$$\begin{aligned} \mathbf{V}_k &= E_r[\mathbf{R}_{s_k}] + E_{s_{k+1},d_k}[\gamma^{d_k}\mathbf{V}_{s_{k+1}}] \\ &= E_r[\mathbf{R}_{s_k}] + \sum_{s_{k+1}\in\mathcal{S}} \mathbf{T}_{s_k,s_{k+1}} \int_0^\infty \mathbf{D}_{s_k,d_k}\gamma^{d_k}\mathbf{V}_{s_{k+1}}\,dd_k \end{aligned} \tag{2.13}$$

where $\mathbf{D}_{s_k,d_k} = \mathbf{D}_{s_k,d_k,\pi(s_k)} = P(d_k|s_k,\pi_{s_k})$, the probability that the dwell time was d_k given the state and

action. Since $V_{s_{k+1}}$ is independent of the dwell time d_k, we can move it out of the integral. What remains is just the standard Bellman equation with a state-dependent discounting factor defined by the integral, a construction that Parr (1998) used to import standard correctness proofs for reinforcement learning methods to the semi-Markov framework.

TD for SMDPs works as usual (Bradtke and Duff, 1995), except that learning occurs only on state transitions and it is necessary to discount by the sampled transition time d_k:

$$\delta_k = r_k + \gamma^{d_k} \widehat{\mathbf{V}}_{s_{k+1}} - \widehat{\mathbf{V}}_{s_k} \tag{2.14}$$

This relies on the use of $\gamma^{d_k} \widehat{\mathbf{V}}_{s_{k+1}}$ as an estimated sample of $E_{s_{k+1},d_k}[\gamma^{d_k} \mathbf{V}_{s_{k+1}}]$. Mahadevan et al. (1997; Das et al., 1999) extend this in the obvious way to the average reward, Q-learning case.

I should note that the MDP and SMDP models are very close to being formally equivalent. That is, any SMDP can be simulated (with arbitrarily small time-discretization error) by an MDP that tracks dwell time by subdividing each semi-Markov state into a series of Markov states. However, the use of a truly continuous underlying time variable, the potential for different timescales in different areas of the environment, and the separation of timing logic from state-transition logic all make the SMDP formalism much more natural for reasoning about situations where timing or timescale is important.

2.1.6 Reinforcement learning: partial observability and non-Markovian contingencies

Difficulties quickly arise when one attempts to apply the abstract MDP model to real-world situations such as robot control or animal behavior. A crucial question is what corresponds to a *state* of the process. We might hope that states would correspond to the animal or robot's immediate sensory observations (which I will denote o_t) — that is, that at each timestep, the agent could measure the complete state of the system directly and anew. But in realistic situations, this is unlikely to be the case: immediately available sensory information is usually insufficient by itself to satisfy the Markov property of containing all available information about the future behavior of the system. This could be due to non-Markovian contingencies or to *partial observability*: situations in which the relevant state evolves as a Markov process but is not directly observable.

An example is an experiment known as *trace conditioning*, which will be important in this thesis. Here, animals learn that a momentary signal (such as a flash of light) predicts the availability of reward after a delay of, say, two seconds. Suppose we bin time into 100 ms increments and take the state of the world to be the animal's sensations during each. Since the reward signal is fleeting, nothing immediately observable distinguishes the time bin directly preceding reward delivery from other time bins such as the one directly following reward delivery. But, though the observed "state" is the same in both cases, $o_t = o_{t'}$, the distribution over subsequent rewards is not: $P(r_t|o_t, a_t) \neq P(r_{t'}|o_{t'}, a_{t'})$. This violates the Markov property.

There are two (related) approaches to this problem: changing what counts as a state, or modifying the underlying formal model. One approach is to augment the state to include previous as well as current observations, defining it as an n-tuple $s_t = \{o_t, o_{t-1}, ..., o_{t-n+1}\}$. (Note that each o_t may itself be vector-valued, containing for instance readings from various sensors.) This scheme, proposed by Sutton and Barto (1990) in the context of modeling classical conditioning, captures the idea that observations made at previous timesteps can (as in the case of the flash of light) have persistent effects on the behavior of the process. If this vector includes *all* observations $o_1 ... o_t$, then the Markov property is trivially satisfied (and rendered useless), since there is no further information of which transitions and rewards could be independent. But increasing the depth of the history vector increases the number of states (exponentially, unless function approximation is employed to reduce the effective size), making the problem of learning values for all of them more difficult. More sophisticated schemes (McCallum, 1995) use statistical tests to identify the minimal amount of observational history that approximately preserves the Markov property, which may even vary between different subareas of the same MDP.

These memory-based approaches can be seen as grounded in a slightly elaborated formal model: a *higher-order* MDP. In an nth-order MDP, the successor and reward distributions are conditioned on the previous n states. The construction underlying the history-vector methods is that an nth-order MDP can be reduced to a traditional first-order MDP by introducing new states representing n-tuples of the original

states. If a first-order MDP is recovered from an nth order MDP, then otherwise unmodified reinforcement learning algorithms can be used with the usual guarantees in the augmented state space. In natural language processing, higher-order Markov models are common, known there as bigram and trigram (or biword and triword) models.

A more useful view of these issues can be had by adopting a formal model that explicitly incorporates the notion of observation. A partially-observable MDP (POMDP; see Kaelbling et al., 1998, for background) is a process whose state evolves as an MDP, but cannot be observed directly. Instead, at each timestep, the agent receives an *observation* o_t drawn from some set $\mathcal{O}$ under some probability distribution $\mathbf{O}_{s,a,o} = P(o_t = o|s_t = s, a_t = a)$. This observation typically will not uniquely identify the underlying hidden state of the process. Note that o_t may in general be structured; e.g. it may be set-, real-, or vector-valued – in fact, in the POMDP framework, rewards are just real-valued observations of a privileged type. For this reason, the observation model $\mathbf{O}$ in a POMDP encompasses the reward function $\mathbf{R}$ from an MDP, and the observation o_t includes the reward r_t, though it is sometimes still useful to refer to the latter separately. The POMDP is related to the familiar hidden Markov model (HMM) used in many areas of computer science such as natural language processing. In an HMM, the state evolves as a Markov chain, but is observed only indirectly. A POMDP is a hidden Markov model with actions that influence the evolution of the underlying unobservable state variable.

An agent in a POMDP faces a new obstacle to optimal action: that of *state estimation*. But this is easily solved, at least given a world model (which now must include the observation function $\mathbf{O}$ as well as the transition function$\mathbf{T}$). A *belief state* — a distribution over the hidden states given the history of observations and actions — can be recursively maintained by conditioning on each new observation using Bayes' rule. Defining $\mathbf{B}_{s,t} = P(s_t = s|o_{1...t}, a_{1..t})$, the update rule is:

$$\begin{aligned}\mathbf{B}_{s,t} &\propto P(o_t|s_t = s, a_t)P(s_t = s|o_1...o_{t+1}, a_1...a_{t+1}) \\ &\propto \mathbf{O}_{s_t,a_t,o_t} \sum_{s' \in \mathcal{S}} \mathbf{B}_{s',t-1}\mathbf{T}_{s',a_{t-1},s}\end{aligned}$$

where I have made use of the Markov property and omitted renormalization. The fact that the belief state can be computed recursively using only the most recent observation and action (and the previous belief state) means that this approach is one solution to the problem of maintaining all necessary history in memory: the belief state is a sufficient statistic for the process' history. This prompted some early work on algorithms that learn a model of the process (e.g. using HMM techniques) to use in maintaining a belief state during performance (Chrisman, 1992; McCallum, 1993). The systems use that belief distribution as a state representation for otherwise standard Q-learning.

Hope that this approach of applying standard MDP algorithms to belief states might work comes from the fact that the belief states in a POMDP form an MDP. (Its reward and transition functions can be computed by integrating over the hidden states, and the Markov property holds for them as a result of the Markov property of the underlying POMDP.) A practical problem, however, is that the space of belief states (unlike the state spaces for MDPs considered so far) is continuous, and so traditional table lookup methods for exactly representing values and policies are inapplicable. In general, the agent's exact, continuous degree of uncertainty at a given moment should inform its policy: as its belief state shifts smoothly from some state to some other state, the optimal action can change repeatedly, for instance because in uncertain belief states "information gathering" actions may be worthwhile. It turns out that the optimal action can change only a finite number of times if the return accumulates over only a finite number of timesteps; for this reason such value functions are piecewise linear in the space of belief states. Offline algorithms for exactly solving POMDPs (in the finite horizon case) rely on this fact (Kaelbling et al., 1998). Exact solution is extremely computationally expensive, and tractable only for problems consisting of perhaps a dozen states or less.

Attempts to solve the belief-state MDP online using standard, infinite-horizon methods must instead use function approximation to cope with the continuity of the belief space. The previously mentioned Q-learning algorithms of Chrisman (1992) and McCallum (1993) both used linear function approximation for the value function; more recently Rodriguez et al. (1999) and Thrun (1999) discuss similar systems using more elaborate function approximators (both also employ approximations to the belief state). Hauskrecht (2000) exhaustively reviews methods for approximating the value function in POMDPs; however, in contrast to the few references just mentioned, this discussion is geared at offline approximate model solution by

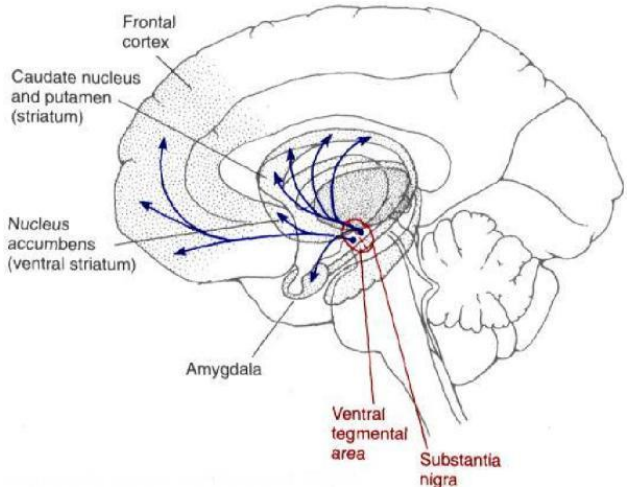

Figure 2.1: Schematic of the dopamine system and its projections. Adapted from Kandel et al. (1991).

methods like value iteration rather than online Monte Carlo methods like TD.

2.2 Related work: neurophysiology and modeling

This thesis centers around the signal carried by a small group of neurons, the dopamine-containing neurons of the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) (Figure 2.1). The dopamine system is phylogenetically venerable — related systems are seen even in insects (Montague et al., 1995) — and implicated in a variety of important functions including learning, motivation, and motor control. In humans, its dysfunction is associated with a number of disorders including Parkinson's disease, schizophrenia, and addiction. The dopamine neurons have diffuse, ascending projections, supplying dopamine to large swathes of the brain. Here I shall review (in a somewhat intermixed fashion) data and theory about the characteristics and function of this system. In addition, I will discuss in less detail work about two other, related brain systems that are also at issue in this thesis: the striatum, which is one of the predominant efferents and afferents of the dopamine neurons, and the serotonin system, another neuromodulatory system that has ascending (though also descending) projections and parallels dopamine in a number of ways.

2.2.1 Neurophysiology and modeling of the dopamine system

Theories about the function of the dopamine system have long focused on two areas: reward and motor control. There is a third hypothesis, that dopamine is involved in attention. Of course, these hypotheses are by no means mutually exclusive, though they often are presented as rivals in the literature. This thesis concentrates on a version of the reward hypothesis that holds that dopamine is specifically involved in the *prediction* of future rewards.

The general hypothesis the dopamine is involved in some aspect of reward builds on such facts as that a number of addictive drugs (e.g. cocaine and amphetamine) seem to exert their primary action on the dopamine system, through some combination of promoting dopamine release, preventing the transmitter's reuptake, or mimicking dopamine in the brain. Similarly suggestive are brain stimulation reward (BSR) experiments, a paradigm in which animals will work to receive electrical stimulation activating neurons in certain parts of the brain, and in fact often prefer such stimulation to natural reinforcers such as food or water. The sites where brain stimulation is rewarding cluster around dopamine-related circuitry (including the VTA itself and the fiber bundle through which the dopaminergic projections ascend), though the exact chain of events by which BSR exerts its rewarding properties is complex and controversial (Murray and Shizgal, 1996a,b).

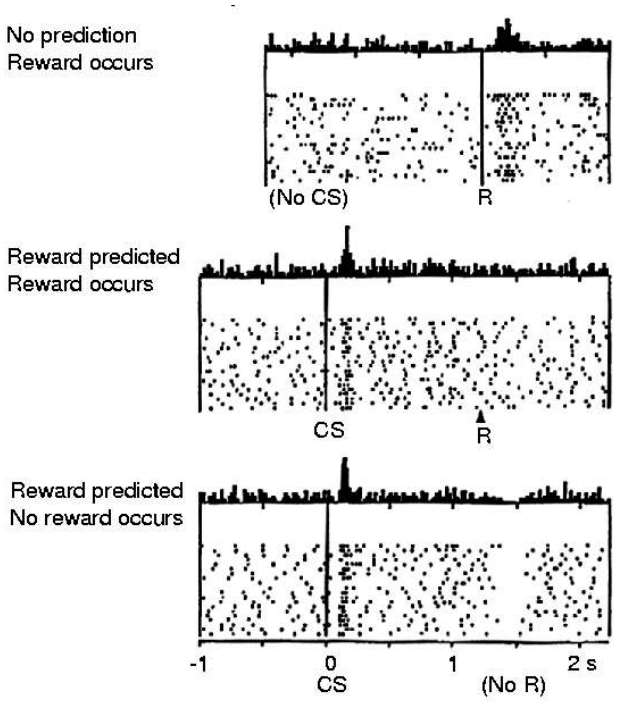

Figure 2.2: Canonical phasic dopamine responses, from Schultz et al. (1997).

The purest, classic version of the reward hypothesis (curiously known as the "*an*hedonia hypothesis," referring to the effects of *blocking* dopamine) holds that dopamine directly reports the hedonic qualities of natural reinforcers to, for instance, brain systems responsible for behavioral control. The hypothesis was introduced by Wise (1982), who now disowns it. According to the hypothesis, drugs of abuse and BSR exert their reinforcing effects by hijacking this signal. A number of more recent reviews have attacked this thesis (Berridge and Robinson, 1998; Ikemoto and Panksepp, 1999; Salamone et al., 1997), but nevertheless have proposed refinements that preserved the broad idea that dopamine is involved in some processes by which rewards or reward expectations influence behavior.

Early research into the response properties of dopamine neurons seems to have been driven more by the idea that dopamine is important for pure motor function. This follows from the most obvious effects of brain dopamine depletion (as famously in Parkinson's disease, in which the dopamine neurons die), which are gross motor deficits: slow movements, difficulty initiating movements, tremor, paralysis, all of which are ameliorated by treatment with L-Dopa, a dopamine precursor which is rendered into dopamine in the brain. But early attempts, inspired by these findings, to correlate the recorded activity of primate dopamine neurons with specific motor actions or muscle activations were largely unsuccessful (DeLong et al., 1983; Schultz et al., 1983).

Instead, the reward hypothesis proved to be a more useful guide to experimenters. A series of recording studies (reviewed by Schultz, 1998), revealed that large percentages of dopamine neurons across both the VTA and SNc respond with a burst of spikes to unexpected primary rewards (such as juice dripped in the mouths of thirsty monkeys) and to stimuli predictive of reward (such as tones cuing the animal to make a response for reward). The neurons do not, however, respond to primary rewards whose delivery is expected, due to being signaled by prior cues (e.g. Mirenowicz and Schultz, 1994). When cued rewards fail to arrive, many dopamine neurons exhibit a momentary pause in their background firing, timed to the moment reward was expected (Schultz et al., 1997). These responses are summarized in Figure 2.2.

Together, these results suggest that the neurons are involved not with reward per se but with the *prediction* of future reward. Specifically , they suggest that dopamine neurons carry an error signal for the prediction of future reward, prompting a series of influential computational models (Montague et al., 1996; Schultz

et al., 1997; Houk et al., 1995; Friston et al., 1994) to propose that dopamine activity carries a TD error signal (Equation 2.4). Montague et al. (1996; Schultz et al., 1997) give the most specific and now standard formulation, using the TD algorithm of Equation 2.4 to update the weights $\mathbf{w}_t$ in a linear estimate of the value function $\widehat{\mathbf{V}}_t = \mathbf{w}_t \cdot \mathbf{s}_t$. The state vector $\mathbf{s}_t$ is a tapped-delay line representation of stimulus history, where each element s_{ijt} of $\mathbf{s}_t$ is one if stimulus i was observed at time $t - j$, and zero otherwise. This corresponds roughly to the higher-order MDP scheme of Section 2.1.6, though using a linear approximation of value as a function of the delay line stimulus history.

This model explains (qualitatively) the dopamine response properties mentioned thus far and depicted in Figure 2.2: unpredicted primary rewards or stimuli that increase the animal's reward expectancy cause momentary positive TD error, corresponding to phasic neuronal excitation, while fully predicted rewards cause no error and no change in firing. When expected rewards fail to arrive, the TD error is negative, corresponding to inhibition of the neurons. Assuming that the prediction error influences action choices, either using an actor-critic approach (Houk et al., 1995; Suri and Schultz, 1999) or more direct hill-climbing (called "learned klinokinesis") on the error signal (Montague et al., 1996), the theory also connects naturally with the ideas about drug addiction and BSR mentioned above. In this sense, it provides a formal counterpart to the anhedonia hypothesis, providing a computationally specific formulation for the informal distinctions (e.g. appetitive versus consummatory conditioning, Ikemoto and Panksepp, 1999, and incentive salience versus hedonic value, Berridge and Robinson, 1998) that lie behind more psychologically driven refinements of the hypothesis (McClure et al., 2003).

Before preceding with a discussion of how the Montague et al. (1996) model has fared in light of more recent experiments, it is worth teasing apart several distinct theoretical claims of the model, in order to evaluate them individually. The core idea is that the dopamine signal reports a TD(0) error in reward prediction, which seems generally to be consistent with the dopaminergic recording studies already mentioned (Schultz, 1998). This error signal is assumed to control learning through changes in synaptic strength at dopaminergic targets. As a general matter, this idea is also supported, as dopamine has been implicated in plasticity in a number of experiments (e.g. Bao et al. 2001). This basic foundation is dissociable from the particularities of the prediction process that the Montague et al. (1996) simulations assumed, which are clearly oversimplified. There, predictions were computed using a linear function approximation scheme based on the standard behavioral conditioning model of Rescorla and Wagner (1972; see Section 2.3.1). This approximator operated over a tapped delay line representation of recent stimulus history (Sutton and Barto, 1990). These particular elements embed a number of further theoretical claims and predictions about experimental outcomes. For instance, linear function approximation implies that the reward associations of multiple simultaneously presented stimuli combine additively; this has been only indirectly tested in dopaminergic recordings (discussed below), and has a mixed record in purely behavioral experiments (e.g. Myers et al., 2001, vs. Pearce et al., 1997). Meanwhile the tapped-delay line representation gives rise to a number of specific predictions about the behavior of dopamine neurons in situations where event timing varies, which are largely not borne out by experiment, and will be a major subject of Chapter 4. In considering experimental tests of the model and how they influenced its further development, it is important to distinguish data that pose serious, in-principle challenges to the core general claims of the model, and data that disagree with the specific instantiation of these claims, but that can be addressed with various (more or less serious and more or less plausible) tweaks to the details. While no single model (including any of the ones presented in this thesis) fits all available data exactly, I would contend that the TD models' failings are not of the serious sort. I know of no data that seriously challenge the core hypotheses of the TD models, though below I discuss some results that have at various times been thought to be candidates for this position.

Turning to the data, I will first discuss experiments that support the model's various claims or challenge them only in relatively superficial ways. Several experiments have supported the relationship between the dopaminergic response and prediction error, by showing that the level of dopaminergic bursting is graded in a way that follows various manipulations of the magnitude of error. This has been demonstrated with error levels manipulated through variability in reward magnitude (Bayer and Glimcher, 2002; Tobler et al., 2002) and also occurrence/nonoccurrence (partial reinforcement; Fiorillo et al., 2003). As expected, these experiments demonstrate that the level of dopaminergic excitation to a reward-predicting stimulus varies proportionally to the associated reward probability and magnitude; similarly, the response to uncued rewards

is graded by their magnitude and the response to cued rewards is graded by their conditional probability given the cues. However, an experiment (Tobler et al., 2002) failed to detect any effect of reward magnitude on the residual response to a cued reward in a partial reinforcement task. In this task, the cue revealed the reward magnitude (small, medium, or large) but was unrewarded on 50 percent of trials, so rewards were partially unpredictable and, when they occurred, excited the neurons. Equivalent dopamine bursting was seen on all rewarded trials regardless of magnitude, though the theory predicts that the response should be graded.

On the issue of the dopaminergic error contributing to learning, Hollerman and Schultz (1998) showed that dopamine responses to rewards declined during learning of a behavioral discrimination in a manner that roughly tracked behavioral improvement. The same article (and later work, so far published only in abstract form; Fiorillo and Schultz, 2001) also addressed the tapped-delay line representation by examining how dopamine neurons respond to variation in reward timing. The results are inconsistent with a prediction of the tapped-delay-line timing mechanism used by (Montague et al., 1996). Suri and Schultz (1998, 1999) offer an ad-hoc fix, and we (Daw et al., 2002a, 2003) have suggested a theoretically better motivated and more general solution, discussed in detail in Chapter 4 of this thesis.

The use of a linear function approximator for $\widehat{\mathbf{V}}$ has also encountered mixed results. Neurons behaved as predicted in a blocking experiment (Waelti et al., 2001), apart from some issues with overgeneralization of response, which are discussed below. Less consistent were results in a conditioned inhibition experiment (which so far has appeared only as an abstract; Tobler et al., 2001); specifically, dopamine neurons failed to show excitation corresponding to predicted positive TD error at the time of expected reward omission after a conditioned inhibitor presented alone. Both the conditioned inhibition and the blocking experiments are also noteworthy from a modeling perspective in that they confirm that the dopamine neurons exhibit more or less normal responses in purely classical conditioning experiments (ones in which animals learn direct stimulus-reward associations instead of behavioral responses); this is important because the models have often tended to ignore the instrumental (response learning) aspects of prior experiments on dopamine neurons and instead idealized them to classical conditioning tasks. (The distinctions between classical and instrumental conditioning are discussed further in Section 2.3.)

There are several other, earlier experiments that some investigators have argued pose a more fundamental challenge to the models and support an alternative, attentional, hypothesis of dopamine function. These are dopamine responses to novel neutral stimuli, aversive stimuli, and stimuli that share qualities with reward-predictive stimuli but are not themselves predictive of reward — none of which the original model obviously accounted for. This evidence was laid out most cogently by Horvitz (2000), who advanced the idea that dopaminergic responses in appetitive situations are just special cases of a broader and more motivationally neutral attentional function, in which dopamine activation is envisioned to tag all salient or surprising events regardless of their motivational value. Redgrave et al. (1999b) make a number of similar points in the service of a related hypothesis about the involvement of dopamine in signaling salient events that cause a shift in animals' behavior. I will now discuss how TD models can explain the specific responses flagged by Horvitz (2000) and Redgrave et al. (1999b), but note first that their alternative attentional hypotheses are not particularly appealing because they would seem to have no way of explaining dopaminergic inhibition in response to omitted rewards or to conditioned inhibitors (Tobler et al., 2001).

Dopaminergic excitation to novel neutral stimuli was reported in early experiments in Schultz's lab (Ljungberg et al., 1992) and later examined more systematically in cats (Horvitz et al., 1997). The level of these responses is graded by stimulus salience and decays with repeated presentations (disappearing after around 100 trials for stimuli of the type used as cues in Schultz's early dopamine experiments, though more slowly for more salient stimuli; by contrast responses to unexpected juice reward persist over months of training; Schultz, 1998). Horvitz (2000) takes this finding as central evidence for his attentional hypothesis of dopamine.

In contrast, Kakade and Dayan (2001a, 2002b) suggest that the dopaminergic response to novelty connects naturally with a body of work in reinforcement learning on the tradeoff between exploiting existing knowledge about reward contingencies and exploring areas of the state space whose reward possibilities are not yet well known. (Incidentally, animals are capable of making quite sophisticated tradeoffs of this sort, e.g., Krebs et al., 1978.) A standard rough-and-ready approach to this problem in practical reinforcement learning applications, supported by some theoretical work (Dayan and Sejnowski, 1996), is to deliver fictitious bonus

rewards to agents encountering novel situations, thereby encouraging exploration. This approach can be problematic because introducing fictitious rewards willy-nilly can distort the optimal policies; Kakade and Dayan exploit a more sophisticated version (Ng et al., 1999) in which the novelty bonus is calculated as the change over each state transition in a potential function ϕ_s, which measures the novelty of state s. The full error signal is then:

$$\delta_t = r_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} + \phi_{s_{t+1}} - \phi_{s_t}$$

This is provably nondistorting because positive bonuses received for encountering novelty will be offset by later negative bonuses when the novel state is left. Kakade and Dayan connect this algorithm to the observation that, at least anecdotally, dopaminergic burst responses to novel neutral stimuli are typically followed by phasic inhibition. Such a response pattern could encourage exploratory behavior, e.g. saccades to novel stimuli, without distorting optimal policies in the long run. (The common trick in computational reinforcement learning of encouraging exploration by optimistically initializing $\widehat{\mathbf{V}}$ with high values is a special case of this approach, and has been used in a dopamine model by Suri et al., 2001.)

Kakade and Dayan (2001a, 2002b) also provide a TD account for another allegedly anomalous class of dopamine responses, known as generalization responses. It has long been known that dopamine neurons respond (again with a burst-pause pattern) to stimuli that resemble those predictive of reward, but which do not themselves have rewarding associations. These responses were first noted by Schultz and Romo (1990) in the context of a task involving two boxes, one consistently containing food and one consistently empty. Dopamine neurons were found to respond to the box door opening (or to a cue light signaling subsequent door opening, in another version of the task), for both the rewarded and the unrewarded box. This finding is central to the critique of TD models by Redgrave et al. (1999b), who also note that the neurons respond *prior* to saccades that animals make toward the opening box, and thus before animals can be certain whether reward is available. Redgrave et al. (1999b) use this point to argue that the neurons respond *too early* to signal reward prediction error. To the contrary, such an overgeneralized excitatory response is exactly what the TD model would predict if the brain's initial sensory indications of a box opening (perhaps the sound of the door or a flicker of movement in the periphery) do not distinguish which box has opened — this is an example of partial observability. These initial sensory hints, though sketchy, would still predict reward roughly half the time and are themselves unpredictable because they occur irregularly. Thus they should immediately evoke positive TD prediction error, quickly followed (if, upon foveation, the unrewarded box is revealed) by negative error. Kakade and Dayan (2001a, 2002b) laid out this argument, but the fact that they could do so using the original, unmodified model of Montague et al. (1996) shows that the entire controversy centered around a misunderstanding of how a TD error signal would behave in the face of ambiguous information. In particular, it is not the case (as Redgrave et al., 1999b, seem to assume) that there is no TD error until reward receipt is certain. Rather, TD models envision (and experiment confirms: Fiorillo and Schultz, 2001) positive error to cues predicting reward only probabilistically.

The third class of anomalous response considered by Horvitz (2000) and Redgrave et al. (1999b) is hints of dopamine activation by aversive events such as footshock or tail-heat in rats. Unfortunately, the literature on this issue is fairly incomplete and confusing. Most data come from microdialysis at dopamine target sites such as the striatum, experiments which measure chemical markers of dopamine activity in samples of brain fluid taken on a timescale of minutes. Such experiments have repeatedly shown evidence of elevated dopamine activity in samples collected during periods in which rats were electrically shocked or similarly punished (for a full review of the many experiments along these lines, see Horvitz, 2000). Given the slow timescale of these measurements, it is not at all clear whether this response bears any resemblance to the quick, phasic dopamine responses that are the subject of the TD theories; instead it is often assumed that the apparent aversive response is a much slower, tonic excitation (e.g. in reviews by Horvitz, 2000, and Schultz, 1998).

Ideally, this hypothesis could be verified by electrophysiological recordings of dopamine neurons in aversive situations, but there are only a few such experiments published, all of which are problematic for various reasons. Mirenowicz and Schultz (1996) found little dopaminergic responding in a task in which animals had to take a cued action to avoid mildly aversive stimuli such as an airpuff to the hand, but the mildness of the punishment and the fact that the animals were able to avoid receiving it with near perfect accuracy argues against taking the negative result as definitive. (About a third of neurons did respond with phasic

pauses to the aversive cues, reminiscent of the negative prediction error response for missed reward; a few also responded with excitation to the cues or the airpuffs themselves, but the article contains no information about the time course of these infrequent responses.) Previously Schultz and Romo (1987) recorded prolonged dopaminergic responses, both excitatory and inhibitory, when they pinched anesthetized monkeys — the inhibitory responses were more common, but the excitatory responses more prolonged. It is unclear to what extent the use of anesthesia complicates the interpretation of this result. Finally, Guarraci and Kapp (1999), reported sustained dopaminergic responses, again both excitatory and inhibitory, to conditioned stimuli predicting electric shock. These experiments were performed in rabbits. Because of electrical interference, they could not record responses to the shock itself.

TD modelers thus generally treat dopamine responses to aversion as tonic, and tonic responses as outside the scope of the theory (e.g., Schultz, 1998). We (Daw and Touretzky, 2002; Daw et al., 2002b) have proposed an alternative account of these phenomena, which also assumes the responses have a slow timescale, but explains them under a unified TD theory of both phasic and tonic dopamine responses. In particular, excitation to aversive stimuli could result from their effect on long-timescale average reward predictions; this theory is presented in detail in Chapter 3.

Voltammetry is another method of measuring dopamine release at target areas, which seems more promising than microdialysis in that it is much faster (subsecond resolution). The most advanced version of the method was developed by Wightman and collaborators (Garris et al., 1997), who have used it to measure striatal dopamine release during brain stimulation reward experiments (Garris et al., 1999; Kilpatrick et al., 2000). Those results were initially widely interpreted as demonstrating modulation of dopamine release by the predictability of the brain stimulation event (as envisioned by TD theories); however a more recent abstract (Phillips et al., 2002) has cast some doubt on this understanding. The group has only just begun to move into recording dopamine release in other operant tasks such as leverpressing for liquid reward, a program which should, in the future, provide data relevant to models of the dopamine system. At present, the available data seem most useful for understanding the dynamics of dopamine habituation and reuptake (Montague et al., 2003), phenomena for which I present some early hints of a computational explanation in Chapter 3.

Finally, I shall review some other models that have built on the original series of TD models. Apart from refinements already mentioned, there have been few attempts specifically to improve the account of dopamine responses *per se*. Suri and Schultz (1998, 1999) focused on the ability of an actor/critic system based more or less on the original model of Montague et al. (1996; Schultz et al., 1997) to simulate animal response learning in an instrumental task similar to those in which the behavior of dopamine neurons had been studied. Unlike the model of Montague et al. (1996; Schultz et al., 1997), the error signal in this version incorporates generalization and novelty responses of the sort recorded in dopamine neurons, and responds correctly to temporal variability of the sort studied by Hollerman and Schultz (1998). However, all of these responses are simply built-in ad-hoc; the model is thus not at all revealing about *why* the neurons behave as they do, nor even about how they compute the mysterious aspects of the signal. Indeed, the only obvious conclusion of this aspect of the work is that these aspects of the error signal do not interfere with response learning. This model also contains an interesting variation on the tapped-delay-line timing scheme, which I will discuss further in Chapter 4.

A much more intricate attempt to develop the action selection side of the TD models was recently published by Dayan (2002); Dayan and Balleine (2002). These articles build a connection between reinforcement learning and a great deal of behavioral experimentation on such issues as the role of motivational state in energizing actions, the development of stimulus-response habits, and the sensitivity of animals to devaluation of previously rewarding outcomes: all phenomena that are not captured in a straightforward optimal policy learning account. The theory includes a direct effect of dopamine on energizing responses, an advantage-learning policy representation scheme to explain the formation of habits, and an internal stimulus-stimulus model based on the successor representation (Dayan, 1993) to account for the sensitivity to outcome devaluation of some instrumental actions.

Several models have attempted to push outward from the identification of the dopamine system as a TD error signal to infer TD-related functions for some brain systems that are interconnected with the dopamine system. A good deal of modeling aimed at identifying further anatomical substrates for the TD model, which will be discussed in the next section, has focused on the striatum. We (Daw et al., 2002b) have tried

to infer a functional role for the neuromodulator serotonin based on its interactions with dopamine, which I will elaborate further in Chapter 3. Doya (2002) presents a similar project of using the TD model of dopamine as a basis for speculating about the functions of other neuromodulators (including serotonin, and also noradrenaline and acetylcholine, which have important attentional and learning functions but are not discussed at all in this thesis). Braver et al. (1999) envision a role for dopamine projections to prefrontal cortex in the gating of information into and out of working memory. This idea, and a long line of subsequent work by the same authors, has been particularly fruitful in understanding the behavior of humans with various brain dysfunctions such as schizophrenia. (If we view learning a memory control policy as an action selection problem, this model closely resembles the learned klinokinesis method of Montague et al. (1996); O'Reilly (2003) has extended these to an actor/critic approach for learning an explicit memory-gating policy.) By far the most physiologically detailed mapping of the entire TD model onto its implementation in terms of brain systems was that of Brown et al. (1999), in work which represents itself (somewhat surprisingly) not to be such a mapping at all but rather an alternative to TD models. Indeed, their approach does not specifically draw on the TD formalism, but it has essentially the same computational shape: the dopamine signal is assumed to be the sum of an instantaneous primary reward signal and the time derivative of a sustained reward prediction signal that resembles TD's $\widehat{\mathbf{V}}$. The main structural elaboration of this model is that it separates into distinct anatomical pathways the positively and negatively rectified portions of this derivative; this is anatomically plausible (see Joel et al., 2002) but does not seem to have any obvious *computational* implications.

2.2.2 Neurophysiology and modeling of the striatum

One of the brain structures most closely associated with the dopamine system is the striatum, which receives a strong dopaminergic input and reciprocates with a projection back to the dopamine neurons. The striatum is the main input structure for a group of brain areas known as the basal ganglia — it receives extensive input from cortex and projects onward to other basal ganglia structures such as the globus pallidus, which eventually (by way of the thalamus) project back to cortex. Thus, these areas are often conceived of as being organized in a loop, from cortex through the basal ganglia and back to cortex. The corticostriatal projections are organized topographically, and connection tracing studies suggest this topography is preserved throughout the entire loop through the basal ganglia, which has given rise to the predominant notion that there are actually a number of distinct loops joining various territories of cortex with associated areas of striatum (Alexander et al., 1986).

Thus the striatum is not an undifferentiated whole; its inputs and outputs are instead highly organized. Moreover, a number of subdivisions can be distinguished on anatomical grounds, and these may also have functional import. Broadly, striatum can be divided into dorsal and ventral subareas; these have particularly distinctive cortical afferents, with the dorsal striatum tending to receive most input from sensorimotor cortex, while ventral striatum receives input from "limbic" structures such as prefrontal and orbitofrontal cortex, the hippocampus and the amygdala (for the most part, areas broadly associated with motivation and emotion). The dorsal striatum is also known as the caudoputamen (after two further subdivisions of the area: the caudate and the putamen); the ventral striatum is often referred to as the nucleus accumbens, and contains a *core* region surrounded by a *shell* with somewhat different inputs and outputs. At a more local level, striatum (particularly the dorsal part) consists of a collection of neurochemically distinguishable "patches" (or striosomes) against a background known as "matrix" or matrisomes.

The predominant cell type in striatum is the medium spiny neuron; these receive cortical and dopaminergic input and project to deliver the inhibitory neurotransmitter GABA to targets outside the striatum. In intracellular recordings made in vitro, these neurons alternate between excited and quiescent voltage states (known as up and down states); in vivo, putative medium spiny neurons recorded extracelluarly display bursty firing patterns, which may result from this alternation. Several types of interneurons (i.e. neurons that connect only locally within striatum) appear in smaller numbers. These are thought to correspond to neurons displaying smoother tonic firing in electrophysiological recordings; slowly firing tonic neurons (known as TANs, for tonically active neurons), have generated particular interest due to their firing correlates, which are discussed below. TANs are thought to correspond to one subset of striatal interneurons that contain the neuromodulator acetylcholine (C. J. Wilson and Kitai, 1990; Aosaki et al., 1995).

Functionally, like the dopamine system (and for similar reasons) the striatum is associated both with reward and with motor control. Due to the functional anatomy already mentioned, there is a natural assumption that the ventral striatum subserves a more motivational role while the dorsal striatum is more motoric. In a classic review, Mogenson et al. (1980) saw the ventral striatum as a "limbic-motor interface," through which motivational or emotional information from the limbic system gained access to the behavioral control functions of the basal ganglia. Everitt and Robbins (1992; Robbins and Everitt, 1996) have advanced this thread, though they focus more specifically on a role suggested by lesion studies for the nucleus accumbens as the route by which information about *conditioned* reinforcers (e.g. lights predictive of reward) controls behavior.

On the more strictly motoric side, the paralysis and other motor symptoms of Parkinson's disease are thought to be due to loss of dopaminergic transmission in (initially, dorsal) striatum. A seemingly complimentary disorder is Huntington's disease, in which medium spiny neurons progressively die, resulting in chorea: spasmic, uncontrolled movements. In animals, lesions or drug manipulations of the dorsal striatum also have motor effects ranging from paralysis to subtler effects on motor learning. On the basis of their own lesion work (McDonald and White, 1993) and an extensive review of work in other laboratories (White and McDonald, 2002), McDonald and White suggest that the dorsal striatum is involved in learning to take motor behaviors cued by stimuli ("stimulus-response learning").

Something similar (but more computationally specific) is suggested by a straightforward mapping of TD policy learning onto striatal anatomy, since stimulus-response learning is akin to determining a state-action mapping in reinforcement learning. Recall that an actor critic model involves an error signal δ, which is used to learn weights underlying both value estimates $\widehat{\mathbf{V}}$ and an action selection policy π. These are all functions of a sensory state representation s. As we know, the error signal is assumed to be carried by dopamine. We can imagine (for the time being) that the state representation is cortical. But what is the site of plasticity where the values and policies are learned, and how do these representations control behavior? Houk et al. (1995), in their actor/critic model of the basal ganglia, envisioned projections from cortex to striatum as the site where dopamine's action on synapses produced learning of policies π and values $\widehat{\mathbf{V}}$, with these two functions divided between matrisomes and striosomes. The article notes that dopaminergic and cortical axons often synapse together on medium spiny neurons, which would be appropriate for a three-factor learning rule in which a dopaminergic error signal trained the corticostriatal synapses to represent weights for value prediction or action selection. The broad idea for behavioral control in this model is a competition between potential actions: each medium spiny neuron, when active, would trigger a particular motor action, and its synapses from cortex represent the extent to which that action is appropriate in different sensory states s. Moreover, in the model, the medium spiny neurons form an inhibitory, competitive network (they project inhibitory collaterals to each other as well as outside the striatum), so the actions compete with each other to be selected. This notion of the striatum as a competitive action selection network occurs in a number of models, including some that do not otherwise adhere to the TD formalism (e.g. Redgrave et al., 1999a). And to the extent that other TD models speculate about anatomical substrates outside the dopamine neurons themselves, the notion of the striatum as subserving action selection and policy learning is ubiquitous. Montague et al. (1996) are less specific about the anatomical substrates of their model, which is also not strictly an actor/critic model. However, they envision that learning underlying value prediction could take place at the targets of the VTA dopamine neurons (e.g. amygdala and ventral striatum; see also Dayan et al., 2000), and that dopamine projections from SNC to dorsal striatum would be involved in behavioral control. Thus this version of the theory also envisions the striatum as involved in both value learning and action selection, though it envisions these functions segregated on dorsal/ventral lines (with action selection in the dorsal striatum), rather than patch/matrix. This breakdown seems sensible in light of the general functional anatomy discussed at the beginning of this section, and fares better than the patch/matrix proposal in a recent, more detailed anatomical review (Joel et al., 2002). The experiment reported in Chapter 5 was designed in part to test this theory. The more recent work of Dayan (2002; Dayan and Balleine, 2002) instead focuses on the roles of nucleus accumbens substructures (core and shell) in subserving various facets of a more fractionated account of action selection.

As already mentioned, TD modelers frequently envision that corticostriatal synapses encode weights not just for policies π but in some parts of the striatum for values $\widehat{\mathbf{V}}$ instead (though additional candidate substrates for this latter function include the orbitofrontal cortex and the basolateral amygdala, see e.g.

Dayan, 2002; Dayan and Balleine, 2002). There have been some attempts to connect this idea with the response properties of striatal neurons in electrophysiological recordings (which, as will be discussed below, are fairly convoluted and inconclusive). Suri and Schultz (2001) set this case out most clearly, noting that, much like value functions, a subset of striatal neurons ramp up, roughly exponentially, in anticipation of reward (and then, again like a value function, drop off rather abruptly when the reward occurs). However, this account doesn't explain why similar anticipatory buildups are seen preceding events and actions other than reward. Montague and collaborators (Montague and Berns, 2002; Montague and Baldwin, 2003) address this issue in a model of value prediction by striatal neurons which is not overtly connected with TD; on their account non-reward events serve as reward proxies that are themselves predicted like rewards. Suri (2001) instead envisions that anticipatory firing for non-reward events is due to the striatum learning an internal stimulus-stimulus model of the sort described in the reinforcement learning literature by Dayan (1993) and Sutton and Pinette (1985). Dayan et al. (2000) (see also Kakade and Dayan, 2002a, 2000) focus on a different aspect of value function learning: the interaction between multiple, differentially reliable pieces of evidence in forming a combined estimate of overall reward expectancy. The authors speculate that, analogous to the competition between actions envisioned in models of dorsal striatum, these sorts of reliability interactions could be subserved by mutually inhibitory competitions between medium spiny neurons in ventral striatum representing the reward predictions associated with different stimuli.

Finally, I will briefly review the results of electrophysiological recordings of striatal neurons. Unlike dopamine neurons, cells in striatum exhibit a blinding variety of response types, and there is currently no unifying theoretical idea about what information they represent. There are also very few distinctions drawn in the literature between the firing properties of ventral and dorsal striatal neurons; in the following I lump results from both areas together, but the experiment reported in Chapter 5 aims to distinguish them. (In light of the topography of the corticostriatal projections, one might expect even finer distinctions between different subareas of dorsal or ventral striatum, but again, such distinctions do not emerge clearly from the available literature on striatal recordings, and would be technically difficult to investigate.)

Striatal responses have been recorded during simple tasks such as cued button pressing or eye movements with monkeys and simple maze traversals with rats. I have already mentioned that striatal neurons (probably medium spiny neurons) display ramping firing preceding rewards, experimenter-controlled stimuli, or motor actions (Apicella et al., 1992). It should be noted that many of these apparent correlates of firing are sometimes experimentally conflated; for instance, neurons firing around the time of reward could be associated with reward per se, or alternatively with tongue or jaw movements involved in consumption. Moreover, instead of anticipatory responses, other striatal medium spiny neurons instead fire during or after rewards, stimuli or actions (see Schultz et al., 1995, for a review of all of these sorts of responses). Meanwhile, TANs often show phasic decreases in firing related to stimuli and rewards, which will be discussed further below. It is common to see all of these different sorts of responses carefully cataloged (e.g. Schultz et al., 1995, figure 2.8, reproduced here as Figure 2.3, or Carelli and Deadwyler, 1997) and neurons classified by response type as, say, "pre-reward excitatory" or "post-cue inhibitory". However, we should be cautious of taking these categories too literally as a complete picture of striatal representation, in particular because they include more or less every firing pattern that might be discernible on peri-event time histograms keyed only to experimenter-controlled events. One reason this is dubious is that assuming that neurons represent the controlled event nearest with their response ignores the possibility that they are actually selective for something else occurring in between. I advocate a different and broader view of the family of striatal responses in Chapter 5.

Several researchers have shown that the general firing patterns just described can be further modulated by various contextual factors. For instance, in a task in which monkeys must complete a series of several trials to receive reward, striatal firing (to cues or actions) is often modulated by the phase of the animal's progression through the series; e.g., neurons may respond preferentially on the first or last trial of the series (Shidara et al., 1998). Somewhat similarly, in a task in which animals are cued to perform either rightward or leftward saccades, but only one side is rewarded, the activity of striatal neurons anticipating the cue is higher during blocks of trials when the contralateral (rather than ipsilateral) side is rewarded (Takikawa et al., 2002; Lauwreyns et al., 2002). In these experiments, the neuronal responses are recorded before the animal receives the cue indicating whether a particular trial is rewarded or unrewarded, and so at issue is the broader context of which direction is associated with reward over a block of trials. Similar experiments

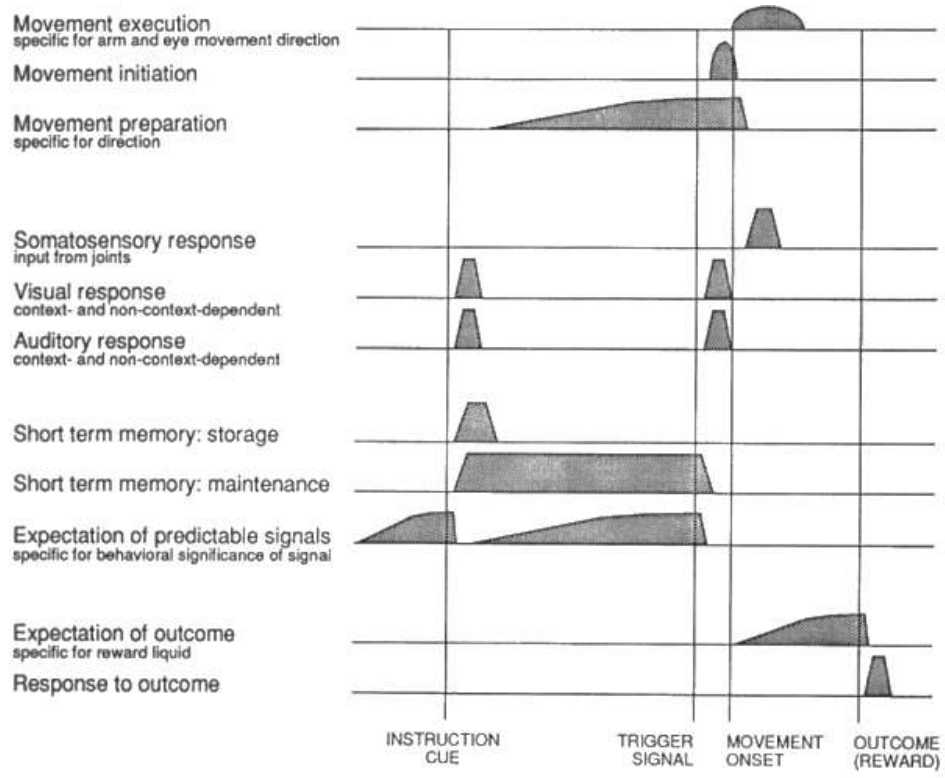

Figure 2.3: Classification scheme for striatal neurons suggested by Schultz et al., 1995.

examining firing after the animal is aware of the pending outcome of a particular trial have shown modulation of striatal responses by whether the trial is to be rewarded or unrewarded (Hollerman et al., 1998), or by which of a number of different types of juice will be used as the reward (Hassani et al., 2001).

Striatal neurons also seem (on limited data) to change their firing properties with learning. Notably, Jog et al. (1999) showed that as rats learned a T-maze task, neurons in the dorsolateral striatum stopped responding during the execution of a trial, and instead began to cluster their firing at the beginnings and ends of trials. The authors interpreted this in terms of a striatal involvement in "chunking" or automization of behavioral sequences. While the results discussed so far in this section probably for the most part concern medium spiny neurons (to the extent that the experimenters have made the distinction), changes in response properties with learning have also been shown in TANs, putative striatal interneurons that fire with a regular baseline rate of 3-8 hz. TANs' tonic firing is typically interrupted by phasic inhibitory responses to cues and rewards, and the percentage of TANs responding to particular cues increased when those cues were conditioned to predict reward (Aosaki et al., 1994). As it turns out, the reason for this may be that the input causing TANs to pause is dopaminergic — e.g., the responses are attenuated by drugs that interfere with dopamine receptors (Watanabe and Kimura, 1998) — and so the responses may just be a sort of negative image of excitatory dopamine responses. In accord with this hypothesis, TAN pauses are modulated by predictability of rewards and stimuli (Apicella et al., 1998; Ravel et al., 2001), similarly to dopamine responses.

2.2.3 Neurophysiology of the serotonin system

Serotonin is a neuromodulator whose detailed properties are as yet much more mysterious than dopamine, but which is implicated in a wealth of important phenomena, ranging from analgesia (LeBars, 1988; Sawynok, 1988) to hallucinations (Aghajanian and Marek, 1999) to a variety of mood disorders such as anxiety and depression (Westenberg et al., 1996; Stanford, 1999). Serotonin is transmitted throughout the forebrain in ascending projections from two midbrain nuclei — the dorsal and median raphe nuclei — and also to the spinal cord in a descending projection from separate nuclei located more caudally. There are also serotonergic projections from the raphe to another midbrain area, the periacqueductal gray, which is involved in the control of defensive and aversively motivated behaviors (see Graeff, 2003, for a review). In this thesis I consider only a small corner of serotonin's numerous functions — its involvement in aversion, and its interactions with the dopamine system.

Of particular interest here are the ascending serotonergic projections from the dorsal raphe, whose targets parallel dopaminergic projections. There is a variety of suggestive evidence for interactions between serotonin and dopamine at these targets, which this section reviews. (Though it is generally assumed that the dorsal raphe is responsible for serotonin's interactions with dopamine, there is some disagreement on this point, and indeed even on whether the dorsal/median breakdown is functionally relevant; Fletcher, 1995; Imai et al., 1986.)

Electrophysiological recordings of serotonergic neurons (Jacobs and Fornal, 1997, 1999; Mason, 1997; Gao et al., 1997, 1998) have been unsuccessful at detecting changes in firing related to much other than general arousal (e.g. the wake-sleep cycle). These results are perhaps not totally conclusive since the studies tended to focus on measuring gross changes in tonic firing rate (averaged over fairly long periods or eyeballed using only a single trial) rather than subtler transient responses of the sort seen in dopamine neurons.

As a result of this poverty of neurophysiology, there has been little detailed modeling of the system, and theorizing on a more general level is driven mostly by the results of pharmacological and lesion interventions. In contrast to the electrophysiological studies, these implicate serotonin in a broad variety of phenomena, some of which were mentioned above. Two broader themes that pervade a number of theories of serotonergic function are that it is involved in behavioral withholding (Soubrié, 1986) and in aversive situations Deakin (1983); Deakin and Graeff (1991); Graeff (2003). Ideas about withholding are based on the fact that treatments which block or deplete serotonin are typically associated with increased impulsivity (and vice versa for treatments increasing serotonin), on measures such as the willingness to choose delayed over immediate reward (Wogar et al., 1993) or the ability to perform "differential reinforcement of low rates" tasks, in which animals are rewarded for pressing a lever, but only if they do so slowly (Fletcher, 1995). In contrast, dopamine is associated with behavioral activation, in part because dopamine manipulations tend

to produce the opposite results on such tasks (and also, more simply, because treatments that enhance or mimic dopamine activity can produce flagrant motor hyperactivity).

Some of the same experiments suggesting serotonin's role in behavioral inhibition are also used to argue for a role in aversive situations: e.g. blocking serotonin inhibits the ability to withhold punished responses (Geller and Seifter, 1960). Conversely, stimulation of the serotonergic median raphe nucleus mimics aversive stimulation by causing animals to withhold rewarded (and otherwise unpunished) behaviors, and also induces responses characteristic of aversion such as teeth-chattering and piloerection (Graeff and Filho, 1978). In addition to withholding ("passive avoidance"), serotonergic manipulations also affect *active* avoidance behaviors such as pressing a lever, jumping, or running away in order to switch off aversive stimulation such as electrical stimulation of the periacqueductal gray (for reviews, see Deakin, 1983, and Graeff, 2003). Data such as these led Deakin and Graeff (1991) to postulate that serotonin controls adaptive responses to aversive situations (such as by mediating fight vs. flight). Again, dopamine has rather the opposite associations: with rewards and appetitively motivated responses.

Thus, at least on the level of theories of broad function, dopamine and serotonin seem to serve opposing roles. In fact, there is a great deal more evidence for opponency between the two transmitters. First, as already noted, treatments that decrease serotonin often have similar behavioral effects to those that increase dopamine (Fletcher, 1993, 1995; Fletcher et al., 1999), and vice versa. Moreover, the opposing effects of multiple drug treatments can be combined: for example, activating serotonin receptors in the nucleus accumbens can block the behavioral effects of dopaminergic agonists like amphetamine (Fletcher and Korth, 1999), while depleting serotonin increases amphetamine sensitivity (Lucki and Harvey, 1976).

Besides the indirect but suggestive evidence from pharmacological manipulations, methods such as dopamine measurement via microdialysis provide more direct evidence (reviewed by Kapur and Remington, 1996) that serotonin inhibits dopamine release, both at the level of the dopamine cell bodies in the midbrain and at shared target structures such as the striatum. In one study (Jones and Kauer, 1999), the activation of serotonergic receptors on dopamine neurons was shown to reduce the excitability of the neurons by decreasing the efficacy of their glutamatergic inputs, apparently from prefrontal cortex. Comparable data supporting the suppression of serotonin release by dopamine do not seem so far to be available.

Though the available data are by no means complete or definitive, we (Daw et al., 2002b) have used this framework of dopamine/serotonin opponency, together with TD models of dopamine, to infer a candidate computational role for serotonin. This work will be discussed at length in Chapter 3. Doya (2002) also presents a TD model with a candidate role for serotonin; in this work, inspired by data about serotonin and impulsive choice (rather than the data data about dopamine/serotonin interactions and aversive functions of dopamine that guided us), he suggests that serotonin is responsible for coding the discounting factor γ.

2.3 Related work: Behavioral experiment and theory

Here I review some items from the fields of behavioral psychology and ethology that relate to the issues discussed in this thesis. This thesis centers around developing careful, algorithmic account of the response properties of dopamine neurons. Concurrently, I attempt to bring this neurophysiological theory into contact with animal behavior, both in order to provide a neuronally grounded account for behavioral data, and to exploit a novel source of empirical constraints on the neurophysiological theories. This is not a radical idea; many of the tasks in which the dopamine neurons have been studied (e.g. Waelti et al., 2001) are adapted directly from behavioral conditioning tasks of the sort described here. And TD had been used to model animal behavior before it was proposed as a model of brain function (Barto and Sutton, 1982; Sutton and Barto, 1990). The goal in this thesis will be to flesh out and extend these nascent connections.

By necessity, the connections I draw will be by no means exhaustive, if only because a theory of the dopamine system is manifestly not a theory of the entire brain or of the vast field of animal behavior. Thus I have selected a number of discrete areas from animal conditioning that seem particularly informative with regard to dopaminergic theories, focusing particularly on some that are difficult to reconcile with existing TD accounts. Moreover, to cope with the vastness of the field (and also with experimental results that are at times contradictory, incomplete, or subject to dispute), my method is to exploit existing bodies of psychological theory by drawing connections between TD models and these theories, and thus a bit indirectly with the behavioral data they describe. In keeping with this approach, the following review highlights a number of

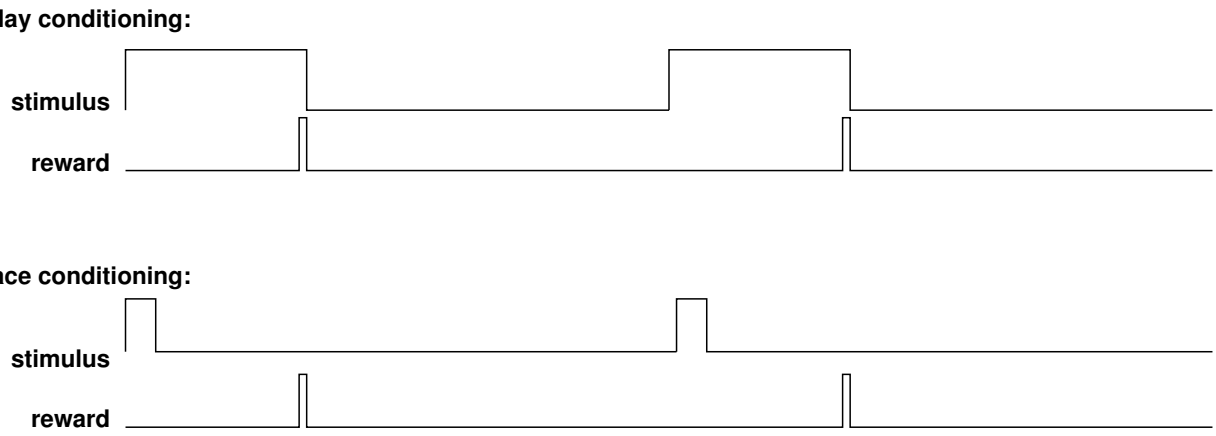

Figure 2.4: Schematic comparing trace and delay conditioning paradigms.

important theoretical ideas from psychology and describes a somewhat selective sampling of the data on which they rest.

2.3.1 Classical conditioning

The study of animal conditioning divides, broadly, into two areas: *classical* or *Pavlovian* conditioning (e.g. Pavlov, 1927), and *instrumental* or *operant* conditioning (e.g. Skinner, 1938). Classical conditioning experiments aim to measure behaviors (such as salivation or freezing) that are thought to reflect directly the expectation of reward or punishment. The rewards or punishments are delivered regardless of the subjects' actions; the experiments instead manipulate the contingencies of their relationship with other experimenter-controlled stimuli and study how the measured expectations change. This contrasts with instrumental conditioning experiments, in which animals make voluntary actions (such as pressing a lever) that result in delivery of reward or withholding of punishment.

Classical conditioning experiments focus on the prediction of some biologically important stimulus such as food or electric shock, which is known as the *unconditioned stimulus* (US). This delivered in some contingent relationship with an (initially) motivationally neutral stimulus, such as a buzzer or a cue light, known as a *conditioned stimulus* (CS). In *delay conditioning*, the CS is presented for a period of time, and the US is delivered coincident with it (normally the US is shorter than the CS and their offsets are aligned.) In another variation, *trace conditioning*, there is a delay between CS offset and US onset. These experiments are illustrated in Figure 2.4. If a CS and US are repeatedly paired, animals will show some characteristic behavior, known as a conditioned response (CR) to the presentation of the CS alone. The degree of CR is assumed to reflect, in a graded manner, the degree to which an animal expects the US. Table 2.1 outlines a number of classic results from Pavlovian conditioning.

In this presentation (following the majority of psychological theories), I will treat the CR as a unitary phenomenon. In fact, there are many different sorts of responses, which sometimes have dissociable neural foundations (e.g. Killcross et al., 1997 — though the authors interpret one of the responses studied there as instrumental rather than Pavlovian), and whose form can sometimes reveal further details about what the animal is expecting or what sort of stimulus underlies the expectation (e.g. Ross and Holland, 1981). Another important distinction is between "preparatory" responses such as approaching a stimulus paired with reward, and "consummatory" responses such as pecking or licking. It has been argued that dopamine is involved in the former but not the latter (see Ikemoto and Panksepp, 1999, who also provide pointers to the rich psychological literature on this distinction).

Classical conditioning experiments are interesting in that they study, more or less directly, how animals make predictions about future events. An important approach to understanding these data is a *normative* one: comparing animal behavior against formal, statistical approaches to prediction that are optimal in some sense. Of course, animals may not be optimal predictors — one might indeed be inclined to assume that rats

Phenomenon	Phase 1	Phase 2	Phase 3	Test⇒Response	Reference
acquisition	A→US			A⇒CR	Pavlov (1927)
extinction	A→US	A→ ·		A⇒ ·	Pavlov (1927)
overshadowing	A,B→US			B⇒reduced CR	Pavlov (1927)
blocking	A→US	A,B→US		B⇒ ·	Kamin (1968)
conditioned inhibition	A→US; A,B→ ·	C→US		C,B⇒ ·	Rescorla (1969)
nonextinction of cond. inhib.	A→US; A,B→ ·	C→US	B→ ·	C,B⇒ ·	Zimmer-Hart and Rescorla (1974)
second-order conditioning	A→US	A,B→ ·		B⇒CR	Rizley and Rescorla (1972)
sensory preconditioning	A,B→ ·	A→US		B⇒CR	Brogden (1939)
negative patterning	A→US; B→US; A,B→ ·			A⇒CR; B⇒CR; A,B⇒ ·	Woodbury (1943)
latent inhibition	A→ ·	A→US		Retarded acquisition of A⇒CR	Lubow (1973)

Table 2.1: Some major effects in classical conditioning. Those above the double line are accounted for by the unelaborated model of Rescorla and Wagner (1972). A, B, and C are CSs. Typically, several repetitions of each trial type are given during the training phases. In the "response" column, "·" denotes a CR that is either significantly diminished (compared to some relevant control) or altogether missing.

and birds are instead rather stupid — but insofar as prediction plays an important role in avoiding danger and achieving nourishment, evolution should favor it. The normative approach is also useful because it can provide an explanation for *why* animals predict in the way that they do. Of course, there is not a single optimal approach to prediction — there are many different statistical models and methods that may also vary in their priors and parameters. But by starting with the assumption that animals perform optimally with respect to *some* statistical model, we can investigate its properties.

One bread-and-butter model of classical conditioning is that of Rescorla and Wagner (1972), which accounts qualitatively for the phenomena in Table 2.1 above the double line. The model aims to learn a linear weight vector that maps the CSs present on a trial to an aggregate reward prediction. Weights are updated using the delta rule (Widrow and Hoff, 1960). In the notation of this thesis, the Rescorla/Wagner model can be written:

$$\widehat{r}_k = \mathbf{w}_k \cdot \mathbf{s}_k$$

$$\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \alpha \cdot \beta \star \mathbf{s}_k \cdot (r_k - \widehat{r}_k)$$

In this, $\widehat{r}_k$ and r_k are the predicted and observed US on trial k, $\mathbf{s}$ and $\mathbf{w}$ are observed stimulus and learned weight vectors. The model incorporates two learning rates, the global learning rate α and a vector β of stimulus-specific learning rates. The operator $\star$ denotes element-wise vector multiplication, so that each stimulus, when present, has its weight updated according to its own learning rate. This is required to model overshadowing. The error-driven learning approach is suggested by the phenomenon of blocking (Kamin, 1968; see Table 2.1), in which CS B does not acquire a predictive association despite being paired with the US in the second phase. On the Rescorla and Wagner (1972) account, this is because the US is already predicted by CS A, so there is no prediction error and no further learning in the second phase. Because the aggregate prediction is a linear function of the CSs present, the model also accounts for conditioned inhibition (Rescorla, 1969; see Table 2.1) by assigning a negative weight to the conditioned inhibitor B, which can offset the positive associations of other CSs presented in the same trial. However, without modification, the model (erroneously: Zimmer-Hart and Rescorla, 1974; Detke, 1991) predicts that repeated presentation of the conditioned inhibitor alone will extinguish its inhibitory power. In general, issues involving conditioned inhibition, the interactions between appetitive and aversive conditioning and between acquisition and extinction are best studied in terms of *opponent process* models (Solomon and Corbit, 1974; Wagner, 1981; Grossberg and Schmajuk, 1987; Daw et al., 2002b), to which I return in the following section.

A wealth of alternatives or successors to the Rescorla and Wagner (1972) model have since sprung up. One important line of work from the perspective of this thesis concerns the fact that the original model was a *trial-level* model; that is, it ignored all effects of timing to focus on the rather abstract problem of predicting whether a US will occur at some point during a trial, given only information about which CSs were present at some point during the trial. A number of classical conditioning models (Barto and Sutton, 1982; Sutton and

Barto, 1990; Moore et al., 1998) have used temporal-difference learning to extend the Rescorla and Wagner (1972) model into the temporal domain, focusing particularly on issues of CR timing and animal sensitivity to the length of the CS-US interval. The extension is rather straightforward because the TD rule can be viewed as an instance of the delta rule, where the target output is not just the US magnitude in the current trial (as in Rescorla and Wagner, 1972), but the US on the current timestep plus the value prediction $\widehat{\mathbf{V}}$ expected to be made on the subsequent timestep.

The TD models improve on their predecessor in that they predict second-order conditioning (Rizley and Rescorla, 1972; see Table 2.1) — an experiment in which an excitatory CS, A, is paired in otherwise unreinforced trials with a neutral CS, B, after which B produces conditioned responding even though it has never been directly paired with primary reinforcement. The TD models predict this effect only in specialized conditions: when the presentation of CS B *precedes* the presentation of CS A in their pairings. In fact, second order-conditioning can also occur when the stimuli are presented simultaneously (Rescorla, 1982), which points to a shortcoming in the TD models of conditioning. Specifically, these models learn only to map CS representations to aggregate USs expected in the future, but learn nothing about correlations between CSs. Learning about such correlations would be necessary to explain phenomena such as sensory preconditioning, (Brogden, 1939), a version of second-order conditioning in which the training session order is reversed, so that CS A and B are paired while both are neutral, and only afterwards is A paired with reward.

As real-time models, TD models can account for some timing effects, though generally not convincingly. In order to account for the fact that some CRs such as rabbits' eyeblinks occur at the time of expected US delivery (some time after the CS onset; see e.g. Kehoe et al., 1989), the models make use of a stimulus representation that incorporates timing information about past stimuli. Specifically, a CS onset initiates a cascade of delayed internal representations of the event, which can be individually associated with later US occurrence. This device was originated by Grossberg and Schmajuk (1989), who called it the Spectral Timing Model, in the context of their real-time conditioning model, which is not based on TD. Sutton and Barto (1990) brought the idea into the TD framework and called it the "complete serial compound representation." However, at least in the TD version, further work (see Section 4.2) is needed for this simple mechanism to reproduce experimentally observed variability in the timing of animals' responses (Gibbon, 1977). A similar problem occurs with the TD model's explanation — using decaying eligibility traces or a decaying stimulus representation — for the observation (Smith, 1968) that animals learn conditioned responses more slowly as the CS-US delay increases. The TD explanation (and similar ones, such as the Sometimes Opponent Processes model of Wagner, 1981, and the Spectral Timing model of Grossberg and Schmajuk, 1989) fails to take into account a related but countervailing phenomenon: that conditioning occurs more quickly if the interval between trials (the US-CS delay) is *increased* (Gallistel and Gibbon, 2000; discussed further in Section 2.3.5). This turns out to be a symptom of a deep problem of *timescale* with these models, a recurrent theme in this thesis.

TD models are also incomplete in that they do not learn anything directly about the expected timing of US events, just a prediction of aggregate future reward. TD models using a tapped delay line representation do learn some CS-US timing information implicitly. However, even these models are unable to account for a variety of experiments in which animals demonstrate rather explicit reasoning about the temporal interrelationships of CS and US events in conditioning. Several fairly abstract theoretical articles (Matzel et al., 1988; Miller and Barnet, 1993; Gallistel and Gibbon, 2000) have reviewed these data and argued for the centrality of temporal information to understanding conditioning; these ideas are sometimes referred to as the "temporal coding hypothesis." Courville and Touretzky (2001) present a more concrete embodiment of the hypothesis, a hidden Markov model account of how animals could learn how the stimuli in a conditioning experiment are arranged on a timeline of events. The model in Chapter 4 develops this work in a number of directions.

The laboratory that originated the temporal coding hypothesis also suggested the "comparator hypothesis," a conditioning theory that focuses less on the process of learning about CS-US associations and more on the response rule that determines when CRs are emitted (Miller and Matzel, 1988; Denniston et al., 2001). On this theory, a number of cue competition phenomena (such as blocking) that Rescorla and Wagner (1972) would explain as acquisition deficits, are instead understood as expression deficits, resulting from comparisons made at the time of *testing* rather than *training*. The theory thus makes a number of predictions,

some of which have been empirically verified, about how subsequent experience can unmask the expression of a seemingly blocked or overshadowed association. Unlike many other early theories reviewed here, to my knowledge this theory has not been addressed from a more rigorous statistical or computational perspective, though the work of Kakade and Dayan (2001b) on backwards blocking is somewhat reminiscent.

A set of issues with the Rescorla and Wagner (1972) model that is normally considered separate from the temporal issues discussed above concerns the representation of stimulus configurations. Because the model uses a linear function approximator to derive an aggregate prediction when multiple CSs are presented, it famously cannot solve nonlinear discriminations such as the XOR ("negative patterning") problem, though animals can (Woodbury, 1943). Several models have envisioned augmenting the model's state representation with "configural" representations of CS tuples, constructed by some heuristic (Pearce, 1994; Touretzky et al., 2002) or by backpropagation of error in a nonlinear neural network (Schmajuk and DiCarlo, 1992; Gluck and Myers, 1993). Dayan and Long (1998) examined the problem of how to construct aggregate predictions given multiple CSs from the perspective of statistical mixture models (though they did not consider configural representations). The present thesis considers the problem of learning representations appropriate to a task in a manner that unifies both configural and temporal representations, though it focuses on temporal aspects of the representation.

There also exists a group of classical conditioning models (Mackintosh, 1975; Pearce and Hall, 1980) that envision that learning about the associative value of each CS is governed by a modifiable, stimulus-specific learning rate (known as the CS's "associability"), and focus on rules for adapting these associabilities and on explaining classical conditioning phenomena in terms of associability. These models are motivated by attentional phenomena such as latent inhibition (in which unreinforced pre-exposure to some CS A retards later conditioning of an A→US association, Lubow, 1973), and provide an alternative explanation for blocking, based on the idea that association is governed by surprise. Kakade and Dayan (2000; 2002a; Dayan et al., 2000) use Kalman filtering to provide a statistically well founded version of these ideas. Holland and collaborators have conducted a series of experiments (reviewed by Holland, 1997) that probed the behavior of animals with various brain lesions on conditioning tasks of this sort; the findings interestingly suggest a dissociation between separate brain systems involved in incremental versus decremental changes stimulus associability.

2.3.2 Opponency

The notion of opponency has appeared repeatedly in both neuroscience and psychology. In neuroscience it addresses the problem that neuronal firing rates are positive values, so that any quantity they directly represent is bounded below by zero. One approach is to represent the quantity in separate, rectified, positive and negative channels, instantiated in the firing rates of distinct groups of neurons. The opposing channels might then engage in various suppressive or competitive interactions. In fact, such a scheme is well documented in the early visual system, where neurons as early as retinal ganglion cells represent the difference between the local image brightness and the brightness of the surrounding area using dual rectified channels, producing either "on-center, off-surround" or "off-center, on surround" receptive fields (Kuffler, 1953). In this and later stages of visual processing, opponent interactions are also thought to take place between channels representing different colors, giving rise to effects like after-images (Hurvich and Jameson, 1957; DeValois et al., 1966), and between channels representing different directions of motion, producing various illusory effects (Heeger et al., 1999).

More germane to the topics of this thesis are ideas about opponency in motivation and conditioning. Many conditioning studies, reviewed by Dickinson and Balleine (2002), suggest that animal motivation is reasonably described as being driven by a single appetitive and a single aversive channel. The existence of multiple appetitive or aversive channels is argued against by the phenomenon of transreinforcer blocking (Ganesan and Pearce, 1988), in which a CS predicting one reward can block the association between another CS and a different appetitive event. The existence of both appetitive and aversive channels (rather than a single channel) is suggested by conditioned inhibition studies in which CSs appear to simultaneously carry appetitive and aversive associations (Williams and Overmier, 1988). Finally, a sort of commensurability or opponent interaction between the two channels is shown by another sort of transreinforcer blocking (Dickinson and Dearing, 1979; Goodman and Fowler, 1983), in which a conditioned inhibitor for an aversive

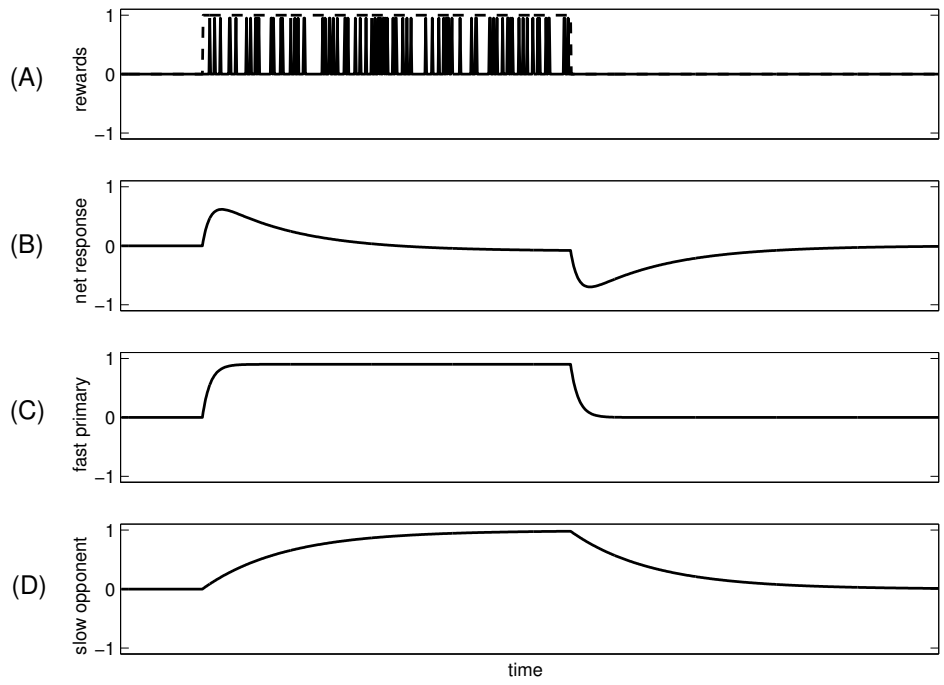

Figure 2.5: Diagram of the opponency model of Solomon and Corbit (1974). Punctate motivationally significant events are treated as a continuous envelope (A), whose time course is reflected in the brain by a fast primary (C) and a slow opponent (D) response. The difference between these two channels gives rise to the net behavioral response (B).

event can block learning between another CS and an appetitive US.

So far we have mostly discussed a *static* notion of opponency, grounded in essentially computational ideas about the neural representation of signals containing both a sign and a magnitude. The interaction between opponent channels at different timescales may also give rise to more *dynamic* effects, as in the example of after-images. However, these effects are usually described at a more phenomenological level. One goal of this thesis, in Chapter 3, will be to offer a more computational explanation for these phenomena.

The classic theory of dynamical opponency in conditioning is that of Solomon and Corbit (1974). They note that on a variety of measures and in a variety of situations, motivational or emotional phenomena display a stereotyped dynamic pattern of habituation, rebound and re-habituation (Figure 2.5 B). To understand the idea, consider a hypothetical experiment in which a subject is fed a series of M&Ms at some rate. According to the theory of Solomon and Corbit (1974), his motivational state (assessed either subjectively or through some measurement like heart rate or skin galvanic response) will initially rise in response to the candy, but then will habituate and level off. Moreover, when the candy is withdrawn, the motivational state will rebound below baseline, and only gradually rehabituate, as shown in the figure. Similar phenomena occur at a variety of timescales and in situations ranging from classical conditioning to drug addiction. Solomon and Corbit's phenomenological model involves treating the punctate candy as a continuous envelope of reward (Figure 2.5A), which excites a fast primary channel (Figure 2.5C) but also a slower opponent channel (Figure 2.5D). The inhibitory interaction between these two channels produces a sort of opponency between *timescales* and produces the rebound and habituation effects. To the limited extent that Solomon and Corbit discuss *why* the system should work this way, they view the system as being geared toward homeostasis.

Similar ideas about the dynamics of opponent processes were subsequently incorporated in models geared more specifically at explaining classical conditioning, along the lines of those discussed in the previous section. The Sometimes Opponent Processes model of Wagner (1981) is the most direct attempt to incorporate the ideas of Solomon and Corbit (1974) into this area; one upshot of this, as mentioned in the previous section,

is an (inadequate, in light of the data reviewed by Gallistel and Gibbon, 2000) explanation for the effects of the CS-US interval on response strength. The classical conditioning models of Grossberg (1984; 1988; 2000; Grossberg and Schmajuk, 1987) also incorporate opponent interactions (though set up in a somewhat different mechanistic fashion), and Grossberg and his collaborators have extensively explored the relationship between the dynamical effects of the device and various phenomena in classical conditioning, particularly focusing on phenomena surrounding extinction and conditioned inhibition, which might be expected to exercise both opponent channels. As just one example (which I will pick up in Chapter 3), the model explains how extinguished CS-US associations can be more quickly relearned or even spontaneously recover, by postulating an active extinction process in which extinction excites the opponent channel rather than inhibiting the primary one. The original CS-US association thus remains encoded in the primary channel, masked by a sort of anti-association in the opponent channel. Thus, after extinction, a variety of learning or attentional phenomena can weaken the opponent channel, causing the primary channel to regain control of behavior and the association to recover. Somewhat reminiscent of the focus of Solomon and Corbit (1974), Grossberg's modeling appears mostly geared toward cataloguing the mechanistic behavior of the opponent system rather than answering the computational question of *why* the system should behave as it does. One of my goals in Chapter 3 will be to build a model that can draw on many of the same mechanistic insights but is computationally founded.

2.3.3 Instrumental conditioning: free operant tasks

We turn now to instrumental conditioning experiments (e.g. Skinner, 1938), those in which an animal learns to take actions in order to acquire rewards. At first glance, reinforcement learning would seem an excellent candidate for explaining these data, as it is essentially a theory of action selection. Needless to say, on a closer look, things are more complicated, mainly because most instrumental conditioning tasks are not easily or usefully viewed as ones in which an animal makes a series of discrete choices between discrete alternatives. Here I first review some basic effects and ideas from traditional instrumental conditioning tasks, concentrating on their tenuous relationship with reinforcement learning. In the following section I turn to discrete choice experiments more obviously suited to the reinforcement learning framework. Note that there is a close relationship between both kinds of tasks, and many of the issues discussed here in are also relevant, to a greater or lesser extent, to the discrete choice tasks discussed in the following section.

In a paradigmatic instrumental conditioning experiment, known as a "free operant" task, an animal is placed in a small computer-controlled box (a "Skinner box") with one or more levers, a food or liquid dispenser, and some cue lights. Responses on each lever are reinforced, or not, according to some programmed schedule. In a ratio schedule, for instance, one out of every n leverpresses is rewarded. In an interval schedule, a leverpress is rewarded if the time since the last rewarded leverpress exceeds some threshhold interval. (In an interval schedule, leverpresses during the interval are ignored and do not reset the timer; the experiment in which every response restarts the interval timer is known as "differential reinforcement of low rates" of responding.) The ratios and intervals may be either fixed or variable, i.e. drawn randomly from some distribution. So, for instance, on a variable interval (VI) schedule, the interval before which the lever is re-armed after each reinforcement is randomized, usually Poisson. The animal may respond as often or as rarely as it wishes; free operant tasks are not explicitly divided into trials or discrete choices.

The dependent variable most often studied is response rate. For instance, a classic result in operant conditioning is the "fixed interval scallop" (e.g. Dews, 1970) — if responses are reinforced on a fixed interval schedule, e.g. once every 60 seconds, then animals' leverpress rates will be lowest just after a rewarded response, and will ramp up sharply as the reinforcement time nears. Experiments also play various schedules off against one another, studying relative response rates between pairs of schedules running simultaneously on two levers. The key result here, and the most important regularity in instrumental conditioning, is the *matching law* of Herrnstein (1961, 1970), which states that the ratio of responses on alternatives will match the ratio of rewards received at each. An important theory that explains this global regularity in terms of local decision-making is the *melioration* model (Herrnstein and Vaughan, 1980), which envisions that animals track local reward rates between alternatives and shift their attention toward the most rewarding alternatives in an attempt to perform a sort of gradient ascent on overall reward rate. At the fixed point, such a strategy will clearly result in matching. Note that such a strategy need not actually *achieve* a globally optimal reward

rate; it is easy to construct response rate-dependent payoff rules that punish matching (Herrnstein, 1991). Nonetheless, in the case of concurrent VI schedules, where matching is most famous and best studied, it is also roughly optimal (Baum, 1981).[3]

There is a subtle aspect to this last result. Strictly speaking, *given a fixed overall level of responding*, near-optimal payoff on the concurrent VI task will occur when those responses are apportioned between the alternatives in accord with matching. This caveat is necessary because it is almost always the case in free operant tasks that faster responding increases the payoff. From the perspective of reinforcement learning or other optimal decision making accounts of behavior, this is a most unfortunate aspect of the experimental designs. Of course, there are physical limits to how fast animals can respond, and there are presumably also energetic costs to responding, so that the optimal policy, in terms for instance of caloric intake versus expenditure, need not be the trivial policy of responding infinitely quickly. Thus, a hypothetical optimal action selection model of these data could be constructed using cost/benefit tradeoffs, but it would rest necessarily and crucially on speculation about the cost side of the equation. All this is to say, it is extremely tricky to address optimal decision-making using free operant tasks and rate-of-responding measures. In this thesis, I concentrate more on discrete choice experiments, in which the issues are somewhat more straightforward.

A second problem for optimality models of operant tasks is that response rate measures are exquisitely sensitive to arousal, motivational and attentional factors outside the cost/benefit framework. For instance, there is a well-studied phenomenon known as Pavlovian-instrumental transfer (Estes, 1948; Lovibond, 1983), in which an animal leverpressing on some schedule will increase its response rate if an appetitive Pavlovian CS is presented noncontingently. Needless to say, the noncontingent presentation of a CS does not alter the costs or benefits of a leverpress in any way, so this effect is extrinsic to an optimal decision-making account. These data are nonetheless interesting from the perspective of reinforcement learning, since they show that Pavlovian stimulus→reward associations (reminiscent of a value function) can affect instrumental behavior. However, Pavlovian-instrumental transfer is crucially *not* an effect on *learning* instrumental responses of the sort that an actor/critic model would predict, but instead a direct arousal of ongoing behavior by Pavlovian predictions. An actor/critic model would predict that animals should learn to choose actions that achieve *contingent* presentation of Pavlovian CSs (because these signal states with elevated predicted value), even if the CSs are unreinforced during training. This is an instrumental conditioning counterpart to the Pavlovian phenomenon of second-order conditioning, and in this context it is known as conditioned reinforcement. The data are unclear: although some effect is documented in a number of tasks (e.g., Zimmerman et al., 1967; see Williams and Dunn, 1994, for a review), all published experiments to date fail to control properly for the possibility that the effect is actually Pavlovian in nature (see, e.g., the discussion by Dayan and Balleine, 2002).

Pavlovian-instrumental transfer and conditioned reinforcement point to a bedrock idea of instrumental conditioning: that instrumental behavior can be fractionated into experimentally dissociable processes dependent on different brain systems and on the representation of a number of separate sorts of information (Dickinson, 1994). Pavlovian-instrumental transfer and conditioned reinforcement depend on Pavlovian representations of stimulus→reward contingencies, and can also be partially affected by motivational states (such as whether the animal is hungry or thirsty or has developed a conditioned aversion to a particular sort of food).

More interestingly, some (but not all) instrumental behaviors are demonstrably *goal-directed,* that is, they depend on knowledge of action→reward contingencies, and are again modulated by the animal's motivational state relative to the specific reward (e.g. food vs. water) expected. This is demonstrated in experiments in which a previously rewarding outcome is *devalued* (e..g by satiation or through pairing it with induced illness), after which animals will cease to perform an action that had previously delivered the former reward (Adams and Dickinson, 1981). (Animals must in fact experience the devalued reward in the new motivational state before they will cease working for it — this process is known as incentive learning — but they need not

[3]It is also the case that in this task, matching only occurs if rapid switching between levers is discouraged with a *changeover delay* — an unrewarded timeout lasting a few seconds that is imposed whenever the animal switches sides. Otherwise, animals tend to alternate rapidly between sides. From the perspective of optimal behavior allocation, this makes sense, since if switching carries no costs, the problem is degenerate in that any allocation of behavior will collect every reward nearly as soon as it becomes available, so long as it involves rapid switching. However, a more suspicious viewpoint is that matching may not be so fundamental if it requires such tinkering with the contingencies to make it emerge.

experience any further instances of the action bringing about the devalued goal; Dickinson, 1987; Dickinson and Dawson, 1988.) This result may not seem surprising, but in fact it challenges simple actor/critic models on multiple fronts. First, traditional reinforcement learning models don't distinguish between different reward types at all, and they certainly don't learn or store any information about specific outcomes — only a numeric summary of the "value" of a situation, together with a state→action policy. Second, because these models use dynamic programming (i.e. learn a Markov policy) rather than performing any sort of dynamic forward planning, they are unable to adjust stored polices when motivational states change, without a great deal of relearning. This inflexibility contrasts with the ability of animals to quickly adjust behavioral policies in the experiment of Adams and Dickinson (1981). A similar ability of animals to plan actions based on representations of task contingencies beyond simple Markov policies was demonstrated by the famous "latent learning" experiments of Tolman (1932), in which, for instance, animals who were allowed to explore a maze freely without reinforcement were faster than naive animals at subsequently learning to traverse a particular route in the maze for reward. A simple actor/critic system would learn nothing during the unrewarded pre-exposure period and thus be equally slow as the untrained controls at subsequently learning the reversal.

There is yet another experimentally dissociable source of instrumental behavior, more reminiscent of TD's policies: stimulus-response habits (Adams, 1982; Dickinson, 1985). These are responses that (e.g. through excessive repetition) have evidently become independent from the representation of their specific outcomes; as a result they persist despite devaluation of the outcome.

Dayan (2002; Dayan and Balleine, 2002) has produced a revision of the actor/critic model, which attempts to account for all of these issues. Specifically, it includes the necessary representational and computational machinery to explain Pavlovian-instrumental transfer, incentive learning and the effects of devaluation, and stimulus-response habits. There have been a few other attempts to connect reinforcement learning models with instrumental conditioning data, though they often (as in this thesis) concern discrete choice rather than free operant tasks. The actor/critic model very closely resembles melioration: though the former allocates behavior based on expected future values rather than local reward rates, it also has fixed points when local reward rates match. Thus Montague et al. (1996) give a TD model for a human card choice task that originated in the melioration literature as a task for which matching is suboptimal. We (Daw and Touretzky, 2001) also exploited the parallels between melioration and actor/critic to study representational issues in TD models of choice between concurrent variable interval schedules. This work demonstrated that the melioration model of the task used too impoverished a state space to predict the detailed temporal structure of animals' behavior, while a TD model using unrestricted tapped delay lines would be able to perform *better* than animals by introducing well-timed responses.

2.3.4 Instrumental conditioning: discrete choice tasks

In discrete choice tasks, animals are asked repeatedly to decide between a pair of alternatives, in order to measure their relative preferences for such parameters as reward amount, delay, and variability. The hope is that by systematically studying how animals trade off these parameters, insight can be gained into into their decision-making processes, and in particular, what sort of return they are optimizing.

One set of experiments concerns time discounting. It is not possible to measure directly how the subjective value of a future reinforcer falls off with its expected delay, but it is possible to get an indirect idea by studying how animals trade off reward amounts and delays when choosing between pairs of reinforcers which differ in these parameters. The predominant understanding of these experiments is that they reject exponential discounting and instead support an alternative model, hyperbolic discounting (in which the value of a delayed reinforcer falls off with the reciprocal of the delay rather than exponentially in the delay). These conclusions are valid only in the context of quite specific models of choice, however, and I will return in Chapter 3 to their implications for choice in more general settings.

Here are the choice models the experiments were intended to test. Assume animals must decide between two rewards of different amounts, r_1 and r_2, after different delays, d_1 and d_2. Then, under exponential discounting, animals should choose based on whether

$$\gamma^{d_1} r_1 > \gamma^{d_2} r_2 \tag{2.15}$$

In the hyperbolic discounting model, choice instead turns on whether

$$r_1/(d_1 + \theta) > r_2/(d_2 + \theta) \tag{2.16}$$

for some parameter θ. Thus the assumption is that animals choose the reinforcer that delivers the larger single-trial return, appropriately discounted. Note that these equations assume that value increases linearly with reward amount, which is not necessarily the case. I will return to this question below in the context of animals' risk sensitivity, where it is more consequential.

Experimental results reject the exponential model of Equation 2.15 because it predicts that relative preferences should be invariant to additive delay, that is if $\gamma^{d_1} r_1 > \gamma^{d_2} r_2$ then $\gamma^{k+d_1} r_1 > \gamma^{k+d_2} r_2$. In fact, many experiments reveal a preference shift from the smaller reinforcer to the larger as the delay to both is increased while maintaining their difference $d_2 - d_1$. For an example and a review, see Bradshaw and Szabadi (1992). A more sophisticated experiment along the same lines (Mazur, 1987) studied indifference points: pairs of delays for which the animal is indifferent between two reinforcers of fixed delay, which were found by titrating the delay to one reward until it was chosen approximately as often as another control reward with a fixed delay. In accord with hyperbolic discounting, these indifference points lay on a line with slope approaching the ratio of reward amounts r_2/r_1 (which was three in the experiment), rather than with slope one, which is predicted by exponential discounting independent of the reward magnitudes. (These data are reproduced in Figure 3.16 on page 68.) Besides experiments using birds (Mazur, 1987; Kacelnik, 1997) and rats (Wogar et al., 1992; Bradshaw and Szabadi, 1992), experiments on humans (Rodriguez and Logue, 1988; Green et al., 1994; Myerson and Green, 1995) have revealed results that appear similar. Because there are serious methodological differences, in this thesis I will not consider the human experiments.

There have been several theoretical attempts to derive or justify the models of Equations 2.15 and 2.16. As discussed in Section 2.1.4, exponential discounting arises when accumulated rewards receive interest, or alternatively if there is some chance, constant per timestep, that future rewards will be lost. The hyperbolic return seems to rest primarily on phenomenology. That is, the model fits animals' choices well, but it is unclear *why* animals should choose this way. However, some attempted justifications for hyperbolic discounting rest on the similarity between Equation 2.16 and a measure of the rate of reward (Rachlin, 1989; Kacelnik, 1997). The most detailed version of this idea is due to Kacelnik (1997), who suggests a normative account for hyperbolic discounting by introducing an infinite-horizon return (note that the return of Equations 2.15 and 2.16 is truncated at a single trial). The original experiments typically failed to control for the effect of an animal's choice on the time at which the next trial arrived. That is, choosing the earlier reward would also bring about the next trial, and hence all subsequent rewards, that much sooner, a fact that would only be reflected in a long-term return. If we assume animals are maximizing their long-term average reward expectation (Equation 2.6) computed over multiple trials, then it is easy to see that they will choose as in Equation 2.16, with θ equal to the delay between trials.

However, the rate maximization account has a serious quantitative deficiency: measured values of the parameter θ in both Mazur's and Kacelnik's own experiments are in the range of about 1 second (this parameter can be estimated because it controls the intercept of the indifference line in Figure 3.16 on page 68). In contrast, the theory predicts that the parameter should match the delay between trials, which was 60 seconds. Moreover, manipulations of the intertrial interval in similar choice experiments have revealed no systematic effect on animals' preferences (Mazur, 1989). Thus Kacelnik's theory must resort to ad-hoc assumptions to the effect that the animals are somehow ignoring the interval between trials when computing their expected reward rates.

Finally, I should note that it is not clear that these experiments actually measure pure preference between delays at all: like free operant tasks, discrete choice experiments may also be affected by the multitudinous behavioral influences discussed in the previous section, such as conditioned reinforcement or habitual motor responding (Cardinal et al., 2003). For instance, we (Cardinal et al., 2002) attempted to reproduce Mazur's (1987) results using rats rather than pigeons, and found that animals' choices were wholly insensitive to the shifts in delay contingencies that were supposed to allow us to titrate to the indifference points. We suggested that responding was instead dominated by stimulus-response habits, which would explain its insensitivity to changes in the outcome contingencies. Also, in most experiments of this type, cue lights were used during the delay periods to span the gap between choice and reward. Thus these experiments might in fact be measuring conditioned reinforcement, i.e. the relative preference for one cue light over another,

rather than anything direct about the expected delay to the primary reward. In accord with this hypothesis, a number of experiments (reviewed by Mazur, 1997) have shown that animals' preferences can drastically be shifted by manipulating the durations of the cues, without changing the magnitudes or delays of the subsequent rewards. Since most data about hyperbolic discounting were recorded from experiments using just such confounding cues, it is at present unclear to what extent true temporal discounting, uncorrupted by conditioned reinforcement, would follow the hyperbolic pattern. That said, Cardinal et al. (2000) compared versions of a time-discounting task with and without bridging cues, and saw qualitatively similar discounting behavior.

A related set of experiments concerns animals' sensitivity to *variability* in reward amounts or delays. The results reveal two general deviations from strict optimization of expected average reward. In studies using a wide variety of methodologies, animals tend to be *risk-prone* (favoring variability) for delays and *risk-averse* (disfavoring variability) for amounts (see Kacelnik and Bateson, 1996, for an exhaustive metastudy and review).

I will first discuss variability in delays. Animals in free-operant tasks will prefer a lever on a variable interval schedule over one with a fixed interval schedule with the same average payoff rate (Herrnstein, 1964); discrete choice tasks along the line of the titration experiments discussed above (which determine the fixed delay at which animals are indifferent to some probabilistic mixture of delays) provide a more quantitative measure of the same risk-prone tendency Mazur (1984, 1986b). Animals' indifference points between fixed and variable-time reinforcers are well fit by the expectation over the delay of the hyperbolically discounted reward (Mazur, 1984, 1986b; Bateson and Kacelnik, 1996). That is, a choice which delivers one of the reward amounts $r_1 ... r_n$ at the corresponding delays $d_1 ... d_n$ with probabilities $p_1 ... p_n$ is valued as:

$$\sum_{i=1}^{n} p_i \frac{r_i}{d_i + \theta} \tag{2.17}$$

a quantity known as the "expectation of the ratios" (EOR). This expression predicts (risk-prone) sensitivity to variability in delay but no sensitivity to variability in amount, which must be imposed separately (see below). Note that this expression is incompatible with long-term average reward optimization, which should be obvious since the average reward is risk-neutral. Long-term average reward optimization would instead predict that animals should choose based on the "ratio of expectations" (ROE), the expected amount divided by the expected time to retrieve it (including not only the delays prior to reward but also the intertrial intervals). In the present notation, ROE would be written $\sum_{i=1}^{n} p_i r_i / \sum_{i=1}^{n} p_i (d_i + I_i)$, where $I_1 ... I_n$ are the intertrial intervals corresponding to each possible delay, but this is just a case of Equation 2.6. In any case, although there is controversy in the ethology literature about the normative basis for the EOR rule (see Bateson and Kacelnik's, 1996, discussion of the "fallacy of the averages" and the "fallacy of the fallacy of the averages"), there does not seem to be any convincing justification for why animals' choices should follow this rule. Bateson and Kacelnik (1996) suggest that EOR may be easier to compute than ROE, but differ little from it in realistic, natural situations.

The data about variability in reward amount are less consistent than those for delay. About half of the studies reviewed by Kacelnik and Bateson (1996) demonstrate aversion to variability in reward amount; the rest show no preference or (less frequently) risk-proneness. The best known demonstrations of risk-averse choices involve bumblebees, which preferentially forage on flowers delivering a small volume of nectar rather than flowers that deliver the same volume on average, but sometimes as a large amount and sometimes as none at all (Waddington et al., 1981; Real et al., 1982; Real, 1991). Previously in this section, I have repeatedly cited Mazur's discrete-choice titration studies as rigorous quantitative counterparts to more widely demonstrated qualitative findings. However, Mazur's (1985; 1989; 1991) studies of variability in reward amount are unrepresentative in that they are among the few studies demonstrating risk-prone choices. Part of the discrepancy may be due to the fact that the amount variability in Mazur's experiments is exclusively binary rather than graded, i.e. on each trial either a set reward amount is delivered or nothing. (However, this was also true of the bumblebee experiments that demonstrated the opposite result; for an example of risk aversion with graded reward amounts; see Bateson and Kacelnik, 1995.) The case of binary variability is special, as originally noted by Rachlin et al. (1986), because it can equally well be viewed as a case of variability in reward *timing*. That is, since choices on the risky alternative are reinforced only sporadically, there is a variable delay — often spanning multiple unrewarded trials — between the risky alternative being

chosen and the reward delivered. In accord with this hypothesis, Mazur's (1985; 1989; 1991) data are well fit by the EOR model. For this fit, the delays in Equation 2.17 were taken as the cumulative durations of the conditioned reinforcer during runs of trials in which the variable alternative was repeatedly chosen and the reward finally received. This should remind us that another explanation for the discrepancy between Mazur's findings of risk-proneness and the more common finding of risk-aversion is the confounding effect of conditioned reinforcement, which may not have been a factor in some other designs.

A further explanation for the seeming inconsistency of the literature on tolerance of variability in reward amount emerges when we consider how risk sensitivity for amounts has been modeled. The simplest way to explain risk sensitivity with respect to reward amount is to assume that subjective values do not scale linearly with reward amount. If two pellets of food delivered at once are worth less than twice as much as one pellet delivered alone, then animals will be risk-averse with respect to amount. Montague et al. (1995) used this approach in a TD model of risk-averse bee foraging. But why should this be the case? While it seems that rate of caloric intake would be a reasonable currency for animals to optimize, there are various normative suggestions (reviewed by Kacelnik and Bateson, 1996) for values scaling nonlinearly with amount. One is *reward handling time*: Larger rewards typically take longer to consume. If temporal discounting is in use, this time has a cost, and smaller rewards will be worth proportionally more than larger rewards expected to start at the same time. More interestingly, animals may be limited in their ability to store calories (so very large meals are useless) or they may need to meet a certain energetic budget to survive. In this latter case, if the fixed choice meets their needs they should not gamble on the risky choice if it has a chance of failing to do so (Stephens, 1981). Inspired by this example, a number of studies have found that animals' preferences for or against variability in reward amount can be shifted, or even reversed, by manipulating their energetic budgets (e.g., Caraco et al., 1990); of the studies examined by Kacelnik and Bateson (1996) that attempted this manipulation, most were able to modify the subjects' risk tolerance, often reversing it altogether. Thus some discrepancies between the outcomes of studies (including Mazur's) in which the energy budgets were uncontrolled may be due to accidental differences in this factor.

Finally, several authors (Bateson and Kacelnik, 1996; Niv et al., 2002) have noted that risk aversion can arise due to the way that the learning process estimates the value of alternatives and the way the choice rule makes use of these estimates. If algorithms make decisions based on a local estimate (e.g. computed by a windowed average, weighted average, or Hebbian rule) to the long-run reward rates at the alternatives, then in experiments like these, they will frequently underestimate the value of a risky alternative, and only less frequently overestimate it. To see how this works, consider the bee experiments (e.g. Waddington et al., 1981, in which the variable alternative delivered nothing 90% of the time and a large amount of nectar 10% of the time). In the extreme case, if an algorithm estimates the reward availability at each flower in the most immediate fashion — simply as the amount they received on the last visit — then it will estimate zero reward available on the risky flower 90% of the time, and the large amount for the remaining 10% of visits. In comparison, the reward rate estimate for the constant alternative will be uniformly lower but still positive, making it larger than the variable alternative (and thus the preferred choice) nine times out of ten.

2.3.5 Timing and timescale

In the previous section, we discussed one aspect of time-related behavior: how animals' preferences are affected by the delay before a potential reward would be delivered. Many experiments have examined, more directly, how animals track elapsed time. Generally, the idea is to study the temporal structure of conditioned behaviors (such as leverpresses or Pavlovian CRs) in order to infer the characteristics of animals' timing processes. For example, I have mentioned that animals' responding on a fixed interval operant schedule (in which a leverpress is reinforced only if the time since the last reinforced leverpress exceeds some threshhold) shows a characteristic scalloped pattern, with average response rates ramping up as the payoff time approaches (e.g. Dews, 1970). The shape of this function can be assumed to reveal something about the animal's sense of the temporal imminence of reward. (Note that since there is no explicit penalty for early responding, animals tend to begin vigorously responding early, i.e. *before* they presumably expect the interval to have elapsed.) One disadvantage of this task is that reward delivery interrupts the timing process so that responding corresponding to the remainder of the animal's reward expectancy cannot be observed; there is a discrete-trials version of this task known as the peak procedure

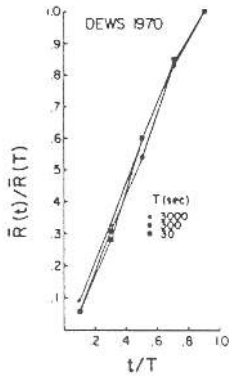

Figure 2.6: An example of scalar timing. FI response rates for three different intervals superimpose when plotted as a function of the proportion of the elapsed schedule interval. Data from Dews, 1970 replotted by Gibbon, 1977.

that addresses this problem by occasionally omitting reinforcement. (In this case, responding wanes after the time of expected reinforcement.) Finally, animal temporal measurements have been addressed even more directly using psychophysical tasks that require animals to compare various elapsed intervals, e.g. to decide whether the duration of a test light is nearer to one or four seconds (Church and Deluty, 1977; Gibbon, 1981).

An important regularity that pervades the results of many of these experiments is known as the *scalar property* of animal timing (Gibbon, 1977). Across many tasks and many measures, animals' responses follow a temporal profile which is proportional to the interval being timed. For instance, if an animal is exposed to fixed interval schedules with different interreinforcement intervals, its response rates subsequent to a reward will ramp up more slowly for schedules with longer delays between rewards. But if those response profiles are plotted as a function not of clock time, but instead of the *proportion* of the inter-reward interval that has elapsed, then response rates recorded for many different inter-reward intervals superimpose (Figure 2.6; data from Dews, 1970, replotted by Gibbon, 1977). That is, responding follows the same profile in each case; this profile is *scaled* to each inter-reward interval.

One upshot of this superimposability is that if we study the *distribution* of some timed animal response (say, the time of peak responding over multiple trials of the peak procedure) as a function of the length of the interval being timed, then the *variability* of that distribution has a characteristic, scalar form: the standard deviation is proportional to the mean (Gibbon, 1977). Thus we say animals' interval timing processes contain scalar noise, i.e. variability whose magnitude, measured by the standard deviation, is proportional to the length of the interval being timed.

These regularities were set out by Gibbon (1977), who took them as the core of his timing model, Scalar Expectancy Theory (SET). In this theory, animals measure elapsed intervals using internal time variables that are linearly related to true clock time but with multiplicative Gaussian noise (producing scalar variability), and respond based on various rules involving ratios between intervals (e.g. time elapsed on the FI experiment divided by the reinforcement interval; such responses will superimpose at different timescales). An obvious alternative is to assume that animals represent elapsed time on a logarithmic rather than a linear scale, with additive noise and comparisons based on subtraction (Staddon and Higa, 1999). Such models are obviously closely related (indeed, strictly equivalent models using linear and logarithmic subjective timescales could obviously be produced); nonetheless SET partisans offer two pieces of experimental evidence that they believe argue in favor of a linear scale. Animals' temporal generalization functions are roughly symmetric on a linear scale but skewed on a log scale (Church and Gibbon, 1982), and choice experiments that pit a long, partially-elapsed interval to reinforcement against a fresh comparison interval do not reveal a bias toward the partially elapsed interval (Gibbon and Church, 1981). (This latter experiment is based on the idea that on a logarithmic scale, a partly elapsed interval will seem closer to completion than a new interval of duration equal to the time remaining, e.g. because $log(30) - log(1) > log(60) - log(31)$.) At least this

second line of evidence seems quite dubious (see the criticisms of Staddon and Higa, 1999).

While the original SET was a fairly abstract theory, it was later developed into a more concrete information processing model, with a number of interacting units such as counters, memories and switches (Gibbon and Church, 1984). The various modules and parameters of this model have served as the basis for explaining the effects of various drugs (Meck, 1996) or brain disorders and injuries (Gibbon et al., 1997) on timing behavior. One point about this work is of particular interest in the present thesis: the analysis of the sources of variability in timing behavior (Gibbon, 1992; Gibbon and Church, 1984), and particularly of variability due to the theory's use of what is known as a pacemaker/accumulator process to measure intervals. Such a process has two components: a "pacemaker" that ticks at some rate, and an "accumulator" that counts up those ticks in order to measure elapsed time. An obvious source of timing variability is then pacemaker noise: i.e., variability in the inter-tick intervals. However, such noise cannot, in itself, explain the scalar variability in animal timing. If the pacemaker is a Poisson emitter, for instance, then the amount of real time it takes to accumulate some threshhold number of clock ticks will be gamma-distributed, and so the standard deviation will scale sublinearly with the mean (specifically, it scales with the square root of the mean). This is actually a general fact, not just for Poisson emitters but for *any* distribution of intertick intervals, assuming the ticks are independent and identically distributed (i.i.d.). In this case, sublinear noise scaling follows directly from the central limit theorem.

The SET literature contains a few strategies for evading this difficulty (Gibbon, 1992; Gibbon and Church, 1984). One is to assume correlated timer noise. If (say) the Poisson tick rate is randomly drawn on each trial from some distribution, then the above argument does not apply because the tick intervals are not i.i.d.. This approach does produce scalar variability (since the rate variability affects the timed intervals multiplicatively); however it depends on the rather artificial notion of a trial. It is often said (see, e.g., Church, 1999) that this trial dependency could be eliminated simply by assuming a randomized timer rate that drifts around (or discretely jumps) at some timescale, not specifically keyed to the progression of trials. However, this only works if the timescale of the drift is roughly the same as the timescale of the trials: much faster drifting will cause sublinear noise scaling (consider the case when a new timescale is drawn for every pacemaker tick, in which case we are back to i.i.d. ticks), while much slower drifting will cause timer noise to be correlated between adjacent trials, and require proportionally more trials to be observed to reveal the full scalar variability. Thus it seems most unlikely that there is a single such timescale that is compatible with all available data, since scalar timing effects are robustly observable with "trials" ranging from sub-second timing to an hour or more. Thus, this general approach can probably only be rescued by operationalizing the notion of trial, so that events in the world that reflect the speed of trials also control the speed of timer correlations. For instance, one could assume the noise timescale is related to the reward rate (which, at least for simple experiments, usually reflects the trial length). This approach is related to one used in the Behavioral Theory of Timing (BET: SET's main competitor, Killeen and Fetterman, 1988), which assumes that the tick frequency is proportional to the reward rate. This device produces scalar noise because trials for different intervals will nonetheless contain around the same number of pacemaker ticks.

A different strategy for explaining the data, which seems to be the preferred one in SET, is to assume other sources of timing variability in addition to variability in the pacemaker intervals. The information processing version of SET assumes that accumulated pacemaker counts are "stored" in memory during learning and "retrieved" from memory during testing, and that multiplicative (i.e., scalar) noise can intervene during either of these steps. Gibbon (1992) demonstrated that such multiplicative noise dwarfs other sources of non-scalar variability (such as pacemaker noise) in the model.

A further broad theoretical idea about scalar effects in time-related behaviors, due to Gallistel (1999; Gallistel and Gibbon, 2000) is that they are instances of a more general and fundamental *timescale invariance* in many aspects of animal learning and behavior. The superimposability of distributions of timed animal responses across different timescales is taken as one example of this. The other main line of evidence (reviewed by Gallistel and Gibbon, 2000) is the effects of timescale on acquisition and extinction in classical conditioning. (Most quantitative work on this issue has been done in a classical conditioning preparation called autoshaping, in which the CS is illumination of a button, the US is access to seed, and the CR is pecking on the lit key.) Consider a delay conditioning experiment in which a CS is presented for some time T, paired with reward, and then followed by an intertrial interval of length I (Figure 2.7). (If these intervals vary, we refer to their averages.) The number of trials it takes animals to learn to emit CRs that

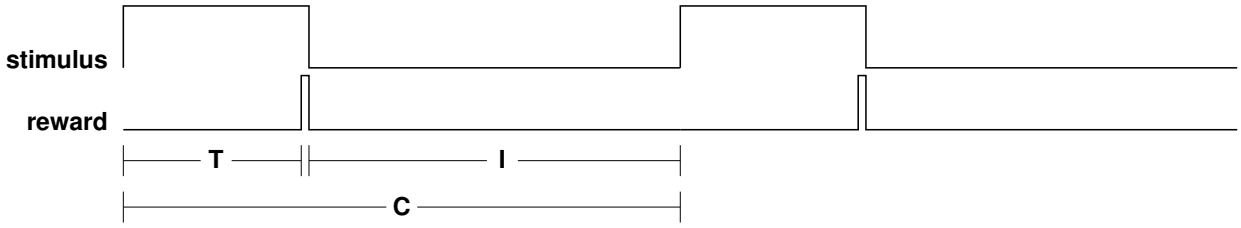

Figure 2.7: Schematic of a delay conditioning experiment, illustrating the notation for time intervals.

meet some acquisition criterion in such an experiment is *invariant* to contractions or expansions of the timescale of events. That is, doubling both T and I (slowing events down by half), or halving both intervals (thus speeding up events) has no effect on the number of trials to acquisition, and such invariance has been demonstrated in timescales ranging over several orders of magnitude. In contrast, changing I or T alone without equivalently changing the other interval can have vast effects on the time to acquisition. In fact, acquisition time is roughly proportional to the ratio T/I of the CS duration to the length of the inter-trial interval (or perhaps T/C, where the "cycle time" $C = I + T$). Gallistel and Gibbon (2000) explain this result as originating from statistical coincidence detection. For instance, if the CS is active most of the time (high T/C), then CS and US would often be coincident just by chance even if the US were delivered at random with no explicit relationship to the CS, and it would take more trials of evidence to detect a significant correlation. Meanwhile, if the CS is rare (low T/C), such coincidences are unlikely and it will take few trials to confidently detect a correlation. The paper attacks associative learning models such as TD for failing to respect timescale invariance; a major goal of Chapter 4 of this thesis will be to answer this criticism by developing a timescale invariant version of TD that is consistent with much of the data reviewed in this section. Kakade and Dayan (2000, 2002a) recast Gallistel and Gibbon's (2000) theory in a Bayesian form, and develop it into a Kalman filter model of rate estimation that explains how attentional effects can modify the speed of acquisition.

Chapter 3

Returns and temporal horizons

This chapter presents a unified attack on three seemingly unrelated issues with models of the dopamine system. The fundamental theoretical question is what return, or measure of expected future reward, the system is learning to predict, an issue that has implications both for the response properties of dopamine neurons and for the decision-making behavior of animals. The particular focus will be on how the system could represent *long-term* predictions about rewards expected to arrive far in the future, an issue whose implications for the dopaminergic signal were not addressed by previous models. A second question is how the brain's prediction systems represent *negative* prediction error — due, for instance, to aversive events or to rewards that are expected but fail to arrive. Neurophysiological data indicate that the negative portion of the dopaminergic prediction error signal is partially rectified, suggesting the need for a companion, opponent neural system to represent the negative error. Here I propose that the dorsal raphe serotonin system could fulfill this role. Intersecting both the issues of negative error and long-term predictions, there is also a wealth of psychological data and theory concerning the opponency between appetitive and aversive motivational systems — and, additionally, between predictions learned at different timescales. The consideration of opponency prompted by these intersecting psychological and computational concerns leads to a substantially enriched model that addresses a third, physiological issue: the functional roles of different *timescales of response* in the dopaminergic and serotonergic systems. Specifically, the new theory provides a framework for explaining tonic dopamine responses, while previous theories had addressed only phasic responding. The material in this chapter recapitulates and expands on theories presented by Daw and Touretzky (2002) and Daw et al. (2002b).

The next chapter, Chapter 4, further develops some of the same threads of prediction and timescale in the service of addressing some other theoretical concerns involving timing and uncertainty. In particular, that chapter explores an alternative scheme for accounting for the costs of delays (an important issue when optimizing measures of long-run reward), which gives a richer and more flexible account of learning about events at different timescales than the one considered here.

3.1 Background

Here I provide a more thorough summary of the different strands of investigation in this chapter. Though many models agree that the dopamine system is involved in learning to predict reward, there has been little attention to exactly what form those predictions take. That is, as discussed in depth in Section 2.1.4, reinforcement learning algorithms exist for learning to predict a variety of returns (differing, for example, with respect to discounting and temporal horizons), and it is not obvious which one(s) animals might be using. This raises issues both of behavior and of physiology. If, as the actor/critic model would suggest, animals are making choices that attempt to *optimize* some return, then their behavioral choices will differ depending on which return they are optimizing. On a physiological level, dopamine neurons should behave differently depending on what error signal they carry.

In fact, most previous dopamine system models (notably Montague et al., 1996; Schultz et al., 1997) were based on the simplest possible return, the undiscounted value function of Equation 2.1 on page 6, with the

sum implicitly truncated at trial boundaries. The use of such a temporal horizon is suspect, both for reasons of optimality (cumulative reward per trial is a poor currency to optimize, since trial lengths may vary), and simple practicality (it is not at all clear how or indeed whether animals segment their continuous experience into a series of discrete trials). Previous articles largely took the position that this episodic approach to learning was a simplification that would nonetheless be easy to fix, by adopting the discounted return of Equation 2.7 on page 10. But while the discounted return has actually been used in some simulations of dopaminergic behavior (Suri and Schultz, 1998, 1999), no work has examined the consequences of this modification on the behavior expected of dopamine neurons, beyond the obvious observation that dopamine responses signaling the expectation of future reward will be attenuated by the expected delay.

A seemingly separate weakness of dopamine system models is that they have tended to focus exclusively on modeling phasic dopamine events (bursts and pauses), while treating the neurons' tonic baseline firing rates as constant. Authors have gone so far as to explicitly declare slower timescale dopamine effects (e.g. the apparently slow dopamine response to aversive stimuli and the effects of tonic dopamine depletion in Parkinson's disease) outside the scope of the theory (Schultz, 1998). Unfortunately, there are few available data that are useful to constrain modeling of slow-timescale dopamine dynamics: electrophysiological studies have largely focused on phasic firing, while, due to their poor temporal resolution, neurochemical studies are short on detail. At the same time, though, there is a lack of computational theory relating the reward prediction hypothesis to slower timescale dopamine activity that might guide the design and interpretation of further experiments. This chapter develops one such theory.

A third issue surrounding models of the dopamine system is how the signal represents negative prediction error and (presumably related) whether it processes aversive as well as appetitive events. In rate code models, in which neurons are assumed to fire at a rate proportional to some number they are signaling, an issue often arises concerning the representation of negative numbers, since neurons can't fire at negative rates. One common device (suggested, for instance, by representations in the visual system) is to imagine that groups of neurons come in opponent pairs, with one set representing a positively rectified and one a negatively rectified signal (e.g. Grossberg and Schmajuk, 1987). Existing dopamine system models have taken a different approach, inspired by the observation (Schultz et al., 1997) that events causing negative prediction error often produce a pause in neuronal background firing. The models thus assume that negative error is represented by excursions below the baseline rate of firing. However, it seems improbable that this is the whole story, because the baseline firing rates are already quite low, and because of the dynamic range of dopamine release, which is strongly nonlinearly driven by *increases* over baseline firing (Chergui et al., 1994; Garris and Wightman, 1994). It is thus unclear whether phasic pauses in dopamine firing produce changes in transmitter release that are plausibly detectable at dopamine targets. Moreover, an experiment that regressed estimated prediction error against firing rate (Bayer and Glimcher, 2002) strongly suggested that the dopamine signal is indeed positively rectified, and that much information about negative error is missing completely from the signal. Finally, there are a range of psychological data and theory (recently reviewed by Dickinson and Balleine, 2002) that suggest that motivation is driven by the interaction between two separate opponent systems, an appetitive and an aversive system.

This chapter describes an attempt to address together all three of these issues — the temporal horizon, tonic dopamine activity, and the representation of negative error. I argue that TD models have all along contained the germ of an account of slow timescale dopamine activity, in that changing existing TD models of dopamine only so far as necessary to remove the artificial horizon on returns introduces a small, slowly changing component to the TD error signal. This allows me to make a number of predictions about dopamine neuron behavior at slow timescales. Moreover, the requirement for processing information about long-term reward predictions, like the requirement for representing negative phasic prediction error, suggests the need for a neural system to serve as an opponent to dopamine. This chapter speculates that the dorsal raphe serotonin system may perform both functions, and also explores the relationship between these ideas and multiple notions of opponency in psychology. Given the paucity of available data, many of the neurophysiological predictions presented here remain speculative, but some also provide new explanations for previously puzzling results, notably the seemingly paradoxical dopamine response to aversive events.

The rest of the chapter is organized as follows: Section 3.2 reviews the returns used in previous dopamine system models. In Section 3.3, a horizonless, average reward TD version of the dopamine model is introduced. Simulation results for this model are presented, followed by a discussion of how they help explain

some puzzling data on dopaminergic response. Section 3.4 discusses how this model relates to ideas about opponency in psychology and neuroscience, and proposes a family of extended models in which a serotonergic signal serves as an opponent to dopamine in representing the TD error signal. The question of what return animals are optimizing is revisited from a behavioral perspective in 3.5, and Section 3.6 recaps and discusses some remaining open issues

3.2 Previous models

Here I review the returns used in previous models of the dopamine system as well as in one relevant neural decision model that did not include a dopaminergic component. This is a slightly tricky endeavor, largely because in several models, temporal horizons were introduced implicitly.

All three of Houk et al. (1995), Montague et al. (1996), and Schultz et al. (1997) used the undiscounted return of Equation 2.1, which (in a continuous setting without temporal horizons) introduces the danger of the value function diverging. The model of Houk et al. (1995) has no mechanism for avoiding this (other than indicating that the discounted formulation could be used), but as no simulations were run, there was no practical problem.

The model of Montague et al. (1996; Schultz et al., 1997) contained no *explicit* temporal horizon, but due to a side effect of the state representation, the return was truncated at the end of every trial. Specifically, the model used a linear approximator for the value function, $\widehat{\mathbf{V}}_t = \mathbf{w}_t \cdot \mathbf{s}_t$, where $\mathbf{w}$ was a set of trainable weights and $\mathbf{s}$ was a tapped delay line representation of stimulus history (a vector of binary elements each of which equaled one if the stimulus was visible a certain number of timesteps in the past, and zero otherwise). In their simulations, the number of elements in $\mathbf{s}$ was smaller than the number of timesteps between trials. Thus, after a trial, the stimulus representation eventually emptied out so that $\mathbf{s}$ was all zeros, and the value estimate $\widehat{\mathbf{V}}$ was consequently forced to zero until the the stimulus reappeared at the beginning of the next trial. This prevented predictions of rewards in subsequent trials from backing up over the intertrial interval, effectively forcing a horizon of a single trial on the learned value function. Had the state representation included a bias element, or had eligibility traces been used in training (see Section 2.1.5), then the learned value function could still have spanned multiple trials, but the model as presented contained neither of these.

The model of Suri and Schultz (1998, 1999) used a representational scheme somewhat more complicated than the tapped delay lines used by Montague et al. (1996; Schultz et al., 1997). However, much like the previous models, this representation also became quiescent between trials, enforcing a temporal horizon on the learned value function. Thus, even though the model of Suri and Schultz (1998, 1999) used the exponentially discounted return (Equation 2.7), which could accommodate an infinite horizon without divergence, the way the model was simulated prevented any examination of the effects on the dopamine signal of rewards expected outside the current trial.

There is one further theory that is neither a TD model nor aimed overtly at modeling the dopamine system, but will nonetheless be relevant in this chapter. The "predictor valuation model" of Montague and Berns (2002; Montague and Baldwin, 2003) asks what return should animals use to assess the value of a stimulus (or perhaps a decision alternative) that predicts some pattern of future reward. The authors propose a return that extends the standard exponentially discounted value function by adding a temporal diffusion to expectations about the timing of future rewards, prior to discounting.

In this theory, which is expressed in model-based terms, a stimulus is supposed to evoke a full timeline of future reward expectancy. This is a function of time, reporting the expected reward amount (probability times magnitude) at every future instant. For instance, a flash of light might predict two food pellets with probability 50% after five seconds, and one more food pellet, with certainty, five seconds later, as illustrated in Figure 3.1, left.[1] By analogy with the reward function in standard reinforcement learning, I will call this function r_t in this discussion, though it is different in that it incorporates an expectation over reward amount, and in that the time variable is assumed to be continuous. I further assume for simplicity that $t = 0$ at the stimulus onset.

In order to assess the value of the stimulus, the authors propose that the reward function first be

[1]For the purpose of computing the return, these two situations (two food pellets with 50% probability versus one with certainty) are equivalent, and the model need not distinguish them.

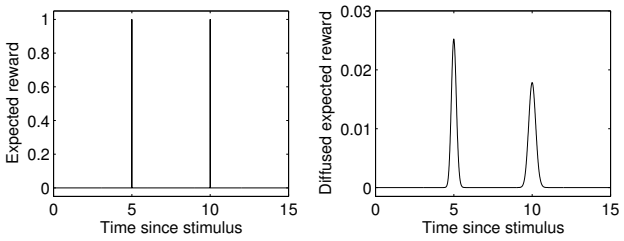

Figure 3.1: Example of temporal diffusion in the predictor valuation model of Montague and Berns (2002; Montague and Baldwin, 2003). Left: Reward expectation function r_t consists of two impulses, corresponding to rewards expected at two different delays after the stimulus presentation. Right: The same function after temporal diffusion by Equation 3.1.

temporally diffused by convolution with a Gaussian kernel whose variance scales with its mean:

$$r'_t = \int_{-\infty}^{+\infty} d\tau \cdot N(t-\tau, \sqrt{c \cdot t}) \cdot r_\tau \tag{3.1}$$

where $N(t-\mu, \sigma)$ is a Gaussian evaluated at t and c is a scaling constant. This convolution has the effect of temporally smearing-out reward expectations, more so the further in the future that a reward is expected (Figure 3.1, right). Then exponential discounting is applied as usual to the temporally diffused reward expectations:

$$\begin{aligned} V &= \int_0^{\infty} dt \cdot \gamma^t \cdot r'_t \\ &= \int_0^{\infty} dt \cdot \gamma^t \cdot \int_{-\infty}^{+\infty} d\tau \cdot N(t-\tau, \sqrt{c \cdot t}) \cdot r_\tau \end{aligned} \tag{3.2}$$

The original papers are vague as to the exact normative justifications for this procedure, though the authors note connections with the Black-Scholes equations that economists use for options valuation. But this return is important in this thesis because it has some interesting connections with empirical data. Notably, the effect of noise in animals' measurements of time intervals (Gibbon, 1977) may be to diffuse their expectations about reward timing in a manner similar to the Gaussian convolution of Equation 3.1 (though with the standard deviation, rather than the variance, scaling with the mean). In Section 3.5 I will study the implications of this phenomenon for animals' preferences on choice experiments without reference to any specific algorithm for learning these values from experience; in the following chapter I will discuss TD models capable of this.

3.3 Dopamine and timescales of response

In the following sections I specify a model of the dopamine system based on an infinite horizon return, explore the effects of long-term reward predictions on the modeled behavior of dopamine neurons, and, finally, discuss how these simulations relate to results seen in dopaminergic recordings.

3.3.1 An average-reward TD model of the dopamine signal

Here I outline a version of the dopamine system model of Montague et al. (1996; Schultz et al., 1997), modified to learn an infinite horizon return. The most straightforward way to do this (and the one repeatedly suggested in previous articles on TD models of dopamine) would be to use the exponentially discounted return, Equation 2.7. I instead suggest a model based on the average-adjusted value function of Equation 2.10. This version is preferable for our purposes not because of any *algorithmic* advantages over the more

common exponentially discounted formulation (indeed, the two algorithms are strictly equivalent in a limit, as discussed in Section 2.1.4 and by Tsitsiklis and Van Roy, 2002), but rather for reasons of *analysis*: the average-adjusted formulation segregates aspects of the error signal's behavior that are due to rewards outside the current trial, a distinction which is obscure in the exponentially discounted version. Also note that a number of aspects of this model will be significantly revisited in the following chapter. I have minimized other changes over the standard model to ease the comparison, as the goal of the analysis here is to examine the specific effect of long-term predictions on the dopaminergic signal. As I introduce further refinements, I will revisit their effects on the results presented here.

The average-reward TD error signal (from Section 2.1.4) is:

$$\delta_t = r_t - \rho_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \tag{3.3}$$

For the present analysis, the average reward estimate ρ is computed as an exponentially weighted moving average:

$$\rho_{t+1} \leftarrow (1-\sigma)\rho_t + \sigma r_t$$

In the next chapter I will suggest an alternative which eliminates this rule's dependence on a particular choice of timescale (determined by the learning rate σ).

As in all previous dopamine system models, I use a linear function approximator for the value function:

$$\widehat{\mathbf{V}}_t = \mathbf{w}_t \cdot \mathbf{s}_t$$

and update the weights with the TD(0) learning rule:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \nu\delta_t\mathbf{s}_t \tag{3.4}$$

for some learning rate $\nu \gg \sigma$.

The state representation is the same tapped delay line vector used by Montague et al. (1996; Schultz et al., 1997). This, too, is provisional: the next chapter will advocate a more elaborate representational scheme. In any case, the experiments in question here have at most a single stimulus, and so we can take the *ith* element s_{it} of $\mathbf{s}_t$ as one if the stimulus was last seen at time $t-i$, and zero otherwise. (Note that if the same stimulus recurs while a previous presentation of it is still represented, the earlier representation is cleared.) The remaining key modification over the model of Montague et al. (1996; Schultz et al., 1997) is that I ensure that this vector is long enough that it does not empty out between trials, so that value can accumulate over trial boundaries. In this scheme, so long as we are working with only a single stimulus, $\mathbf{s}_t$ is always a vector of all zeros except for a single one. Thus for the present purposes, this algorithm using this representation is equivalent to the table lookup methods discussed throughout Section 2.1; the linear function approximation has no effect. One exception to all this is in modeling experiments where no stimuli are delivered, but instead rewards are delivered randomly and unsignaled. In this case, $\mathbf{s}_t$ and thus $\widehat{\mathbf{V}}_t$ and $\mathbf{w}$ are all zero for all t, and everything the algorithm learns about the reward structure is expressed in ρ (instead of $\mathbf{w}$). This is sensible because these tasks can be viewed as controlled by a single state, whose average-adjusted value is arbitrary (because of the fact, discussed in Section 2.1.4, that the average-adjusted return is only defined up to an arbitrary additive constant). In a more realistic model, reward events might be represented in $\mathbf{s}$ (even though they are of no predictive utility here), and ρ itself might have a structured representation, for instance, computed from weights representing the influence of different contextual factors.

These algorithmic improvements also necessitate a change in experimental procedures: the interval between trials is randomized in all simulations here, which was not done in previously published models. If the trial length is constant, and the representation of the stimulus that starts the trial persists between trials, then each new stimulus can itself be predicted due to its stable temporal relationship with the previous stimulus. In this case, after learning there will be no prediction error at all, and so no dopamine response to the stimulus would be expected. This is not in accord with standard dopaminergic experiments, in which the intertrial intervals are randomized and the neurons, accordingly, fire asymptotically to the unpredictable stimuli (an exception is reported by Ljungberg et al., 1992, which I will discuss in the next chapter). Since, in previous models, the representation of a stimulus did not persist from one trial to the next, stimuli were never predictable on the basis of the representation, and it was not necessary to randomize the interval between trials.

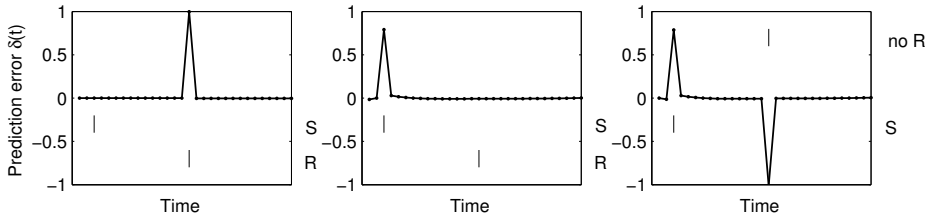

Figure 3.2: Performance of the average reward TD model on a task that demonstrates the basic phasic responses of dopamine neurons. The results are not obviously different from those simulated by Montague et al. (1996) using the undiscounted episodic model. Left: Before learning, positive TD error (simulated dopaminergic excitation) is seen to a reward but not to the stimulus that precedes it. Middle: After learning, the error response transfers to the time of the stimulus. Right: When expected reward is omitted, negative TD error (simulated dopaminergic inhibition) is seen at the time reward had been expected.

The results presented here are meant to reflect the effects of long-term predictions on expected dopamine behavior, in settings more general than the somewhat unorthodox formulation I present here. Specifically, a dopamine system model based on a more traditional untruncated exponentially discounted return would behave quite similarly to the average-reward model presented here, at least for a large γ. This is due to the fact that under some assumptions about the state representation, Tsitsiklis and Van Roy (2002) have proved that the algorithms coincide in the limit as $\gamma \to 1$. (This proof does not apply to previous models such as that of Suri and Schultz (1998, 1999), because the state representation prevents reward predictions from spanning trial boundaries, as discussed in Section 3.2.) Also note that even some systems not meeting all of the requirements for Tsitsiklis and Van Roy's proof (specifically, the exacting requirements involving basis function scaling) would nonetheless qualitatively reflect the behaviors I will present. One reason for this is that most of the results to follow will concern residual prediction error once the prediction parameters $\mathbf{w}$ and ρ have reached asymptote. In this situation the relationship between the two algorithms' error signals is simpler, and can be established (as in the argument sketched in Section 2.1.4) with weaker assumptions.

3.3.2 The behavior of the model under an infinite temporal horizon

Here I present simulations that demonstrate the behavior of the average reward, infinite-horizon TD model in a number of experimental situations. The goal is to examine situations where the model predicts dopaminergic responses that would be noticeably different from those predicted by the undiscounted, episodic TD model of Montague et al. (1996; Schultz et al., 1997). In this way, we can hope to isolate the effect that long-term reward predictions have on the modeled dopaminergic signal (since the truncated return used in the earlier models incorporates no such predictions). In the following section, I will discuss how some of these simulations help to explain the results of dopaminergic recordings. Many of the other predictions described here concern experiments that have not yet been performed; their purpose is to inform future experimentation.

The average reward error signal, $\delta_t = r_t - \rho_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t}$, and its undiscounted counterpart $\delta_t = r_t + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t}$ differ only in that the average reward estimate ρ_t is subtracted from the former. Thus, the experiments we will be studying will involve situations in which this subtraction is evident.[2] Since ρ_t is subtracted at every timestep and changes slowly, it is reasonable to envision it as a nearly constant baseline against which the phasic error signal appears — that is, we can associate this portion of the error with tonic dopamine activity. However, given that dopamine activity is recorded against an arbitrary baseline firing rate, if ρ_t were constant, it would not be possible to distinguish the two error signals on this basis. The effects of the infinite horizon return will instead be manifest mostly in situations where there are large changes in the average reward estimate ρ_{t} , producing noticeable shifts in the modeled dopamine baseline.

[2] I should note that it is not strictly the case that the two error signals always differ exactly by ρ_t. Because the return is defined differently in each case (e.g. truncated in the undiscounted model), there are also differences between the learned values $\widehat{\mathbf{V}}$ that occur in the two equations. In practice, the effects of the ρ_t term will be of the most interest to us.

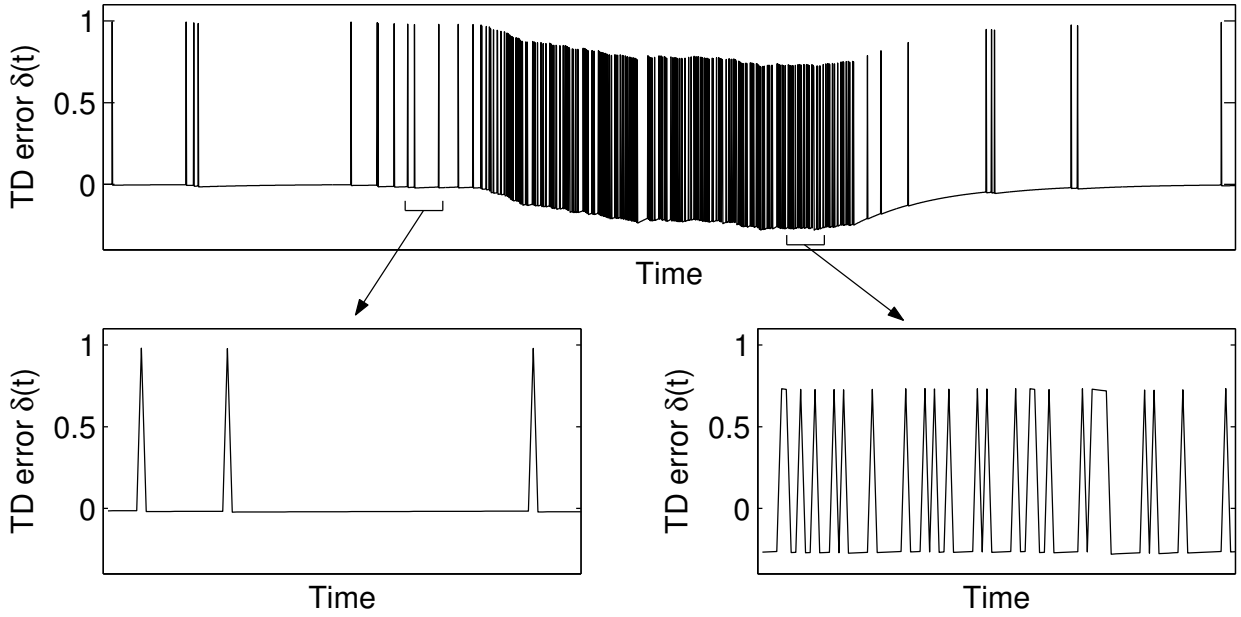

Figure 3.3: The effect of a change in the rate of unsignaled rewards on TD error in the average reward model. Unsignaled rewards occur with a Poisson rate that shifts from 1% per timestep to 25% per timestep and back again; their timing can be seen because they engender a positive phasic "spike" in the prediction error. This phasic response occurs against a background tonic error level determined by the rate of reward delivery. When the reward rate suddenly increases, the background error level gradually ramps down, and ramps up again when the reward rate decreases. Insets enlarge the time base of the error signal during two periods to show local structure.

This is not, for the most part, true of the dopamine experiments performed thus far, in which ρ_t has been both small and fairly stable. For this reason, the new model behaves like the previous one on simulations of these experiments. Figure 3.2 shows the performance of the average-reward TD model on a conditioning experiment originally designed by Montague et al. (1996) to demonstrate several common classes of phasic dopaminergic responses. The results are not visibly different from those reported using the undiscounted model.

The simplest situation in which the effects of long-term expectations ρ_t on the error signal might be visible is when unsignaled rewards are delivered at a changing Poisson rate. Unsignaled rewards are known to cause phasic dopamine activation (e.g. Schultz, 1998), but Figure 3.3 shows that in the average reward model, this activation occurs against a background level of firing determined by the rate of reward delivery. This background firing is gradually driven downward as the estimate ρ_t ramps up when the delivered reward rate increases abruptly. In the model, both the baseline between rewards and the peak phasic activation are affected by the reward rate. Of course, the specific dynamics of the transition depend on the method by which ρ_t is estimated, but the asymptotic behavior should be the same in any case.

I now consider the effects of *aversive* events, such as footshocks, on the modeled dopaminergic response. For this, I will assume that these are simply equivalent to negative reward. For this reason, the modeled dopaminergic response to such events will simply be an inverted version of the traces shown in Figure 3.3: in the model, aversive events phasically inhibit the dopamine system, but then *tonically disinhibit* it by contributing to a *decrease* in the average reward expectancy ρ_t, which returns to normal only slowly. This is illustrated in Figure 3.4.

Figure 3.5 illustrates other situations in which the effects of the average reward ρ_t on the error signal might be experimentally visible. First, consider adding reward signals to the Poisson random reward experiment of Figure 3.3. If we assume a cue precedes each reward by a constant interval (and that cue-reward pairs are delivered on a Poisson schedule except that a second cue cannot occur between a cue and its associated

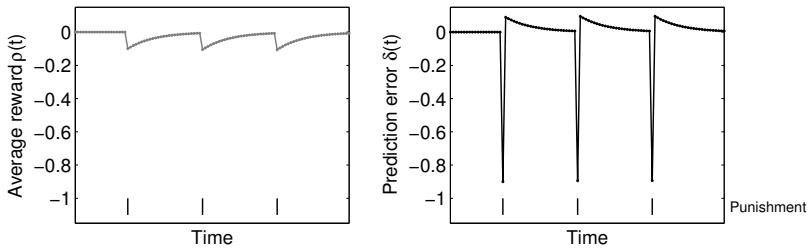

Figure 3.4: Effect of punishment on the average reward estimate ρ_t (left) and the average reward TD error (right). Phasic punishments (such as footshock) phasically inhibit the error signal, but they also *decrease* the average reward estimate, causing a subsequent jump in the error signal above baseline that only gradually returns back to normal. This corresponds to a predicted tonic excitation of dopamine neurons.

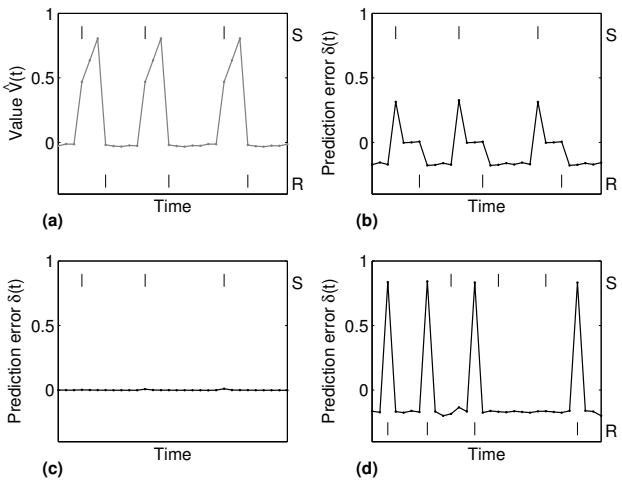

Figure 3.5: Behavior of the average reward TD model in a signaled reward experiment. (a) Asymptotic value function in a signaled reward experiment ramps between stimulus and reward. (b) The effect of this ramping is to cancel the tonic inhibition of the average reward TD error during the period between the stimulus and the reward, so that the predicted dopamine signal during this period is elevated relative to its between-trials baseline. (c) After extinction accomplished by repeated unrewarded presentation of the stimulus, the tonic inhibition disappears, and the error signal is zero everywhere. (d) If extinction is instead accomplished by presenting stimuli and rewards, unpaired, the error signal remains tonically depressed below zero and phasic activation is seen to the rewards.

reward), then the curious situation of Figure 3.5 (b) results. The effects of ρ_t are manifest during the period after each reward and before the next stimulus (pulling the error signal down below zero), but during the stimulus-reward interval, the baseline vanishes and the error signal is zero exactly (i.e. the modeled tonic baseline level increases). This is because the system is able to learn to cancel the $-\rho_t$ term in the average reward error signal by ramping up the value estimate $\widehat{\mathbf{V}}$ during the stimulus-reward interval (Figure 3.5(a)) so that $\widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} = \rho_t$ and the total error is zero. This is not possible outside the stimulus-reward interval since (because stimulus delivery is Poisson) $\widehat{\mathbf{V}}$ is the same at all times. Note that this is the same experiment simulated in Figure 3.2 — the difference is that the overall reward rate was lower in that earlier simulation, so that the effect of ρ_t on the simulated dopaminergic baseline was not appreciable.

Finally, consider the behavior of the average reward model when a stimulus-reward association of the type just described is *extinguished*. The traditional way to do this is to repeatedly present the stimulus and omit the reward. The basic TD model predicts phasic effects in this case: excitation to the stimulus, and inhibition at the time the reward had been expected, both of which should wane as the association weakens. The average reward model in addition predicts tonic effects, since extinction in this manner involves a reduction in the experienced reward rate. Thus the negative effect of ρ_t on the error signal should also wane during extinction, and so the tonic dopamine depression between trials (Figure 3.5(b)) would be expected to disappear (Figure 3.5(c)). An alternative way to perform extinction is to maintain the delivered reward rate, but decouple stimuli and rewards temporally, i.e. deliver both on independent Poisson schedules. In this case (Figure 3.5(d)), ρ_t should be preserved, so the tonic depression between trials should be maintained; instead the depression should spread into the period after the stimulus as well.

3.3.3 Discussion: Experimental evidence related to this model

Here I discuss how the simulations described in the previous section can help to explain several pieces of experimental evidence from dopaminergic recordings. Unfortunately, many of these issues have been studied only poorly, if at all, in electrophysiological recordings of dopaminergic spiking of the sort that we have mostly considered so far. Instead, many of the relevant data come from *neurochemical* recordings that measure the concentrations of dopamine or of chemical markers of its activity in areas targeted by dopamine neurons. Because neurochemical recordings are made at a slow timescale (for instance, in microdialysis experiments, chemical samples are usually taken once per ten minutes), they may be related to the tonic aspects of the prediction signal studied in the previous section. There is some (justified) confusion in the literature as to exactly how these results relate to those from recordings of neuronal spiking, though we can probably assume that higher extracellular concentrations are associated with enhanced spiking. One goal of the present research is to provide a framework for understanding both sets of results, through the distinction of tonic and phasic components to the modeled error signal. However, such a theoretical sketch is significantly simplified compared to the many complex processes (for instance, of dopaminergic release and reuptake) that influence the neurochemical recording results, so interpretational difficulties remain.

The average reward model predicts that baseline dopaminergic activity should be inversely proportional to the rate of reward delivery (Figure 3.3). This prediction has never been tested in electrophysiological recordings of neuronal spiking. But it is consonant with data from two labs that used voltammetry to measure dopaminergic concentrations in the striatum in behaving rats. A hallmark of these experiments is that, measured in terms of transmitter concentrations, the effects of dopaminergic activation *habituate* over time. In the data reproduced in Figure 3.6 (top), Kilpatrick et al. (2000) report phasic changes in dopamine concentrations in rat striatum in response to bouts of brain stimulation reward delivered in an unpredictable train. Though the rewards are delivered sporadically for about a minute, they eventually cease to have any effect on the dopaminergic concentration in striatum. Such a habituation of the response is easily explained (and not surprising) on mechanistic biological grounds — neurons' transmitter stores can deplete over time, and there are various regulatory devices that control release — but under the original TD model, this phenomenon has no clear *computational* interpretation. The model presented in the previous section, however, suggests that this habituation could result from the suppressive effects on the dopamine signal of the long-term average reward expectation ρ_t.

In evaluating this explanation, it is difficult to directly compare the data on neurotransmitter concentrations with the error signal simulated in Figure 3.3, since the concentration data reflect both cumulative

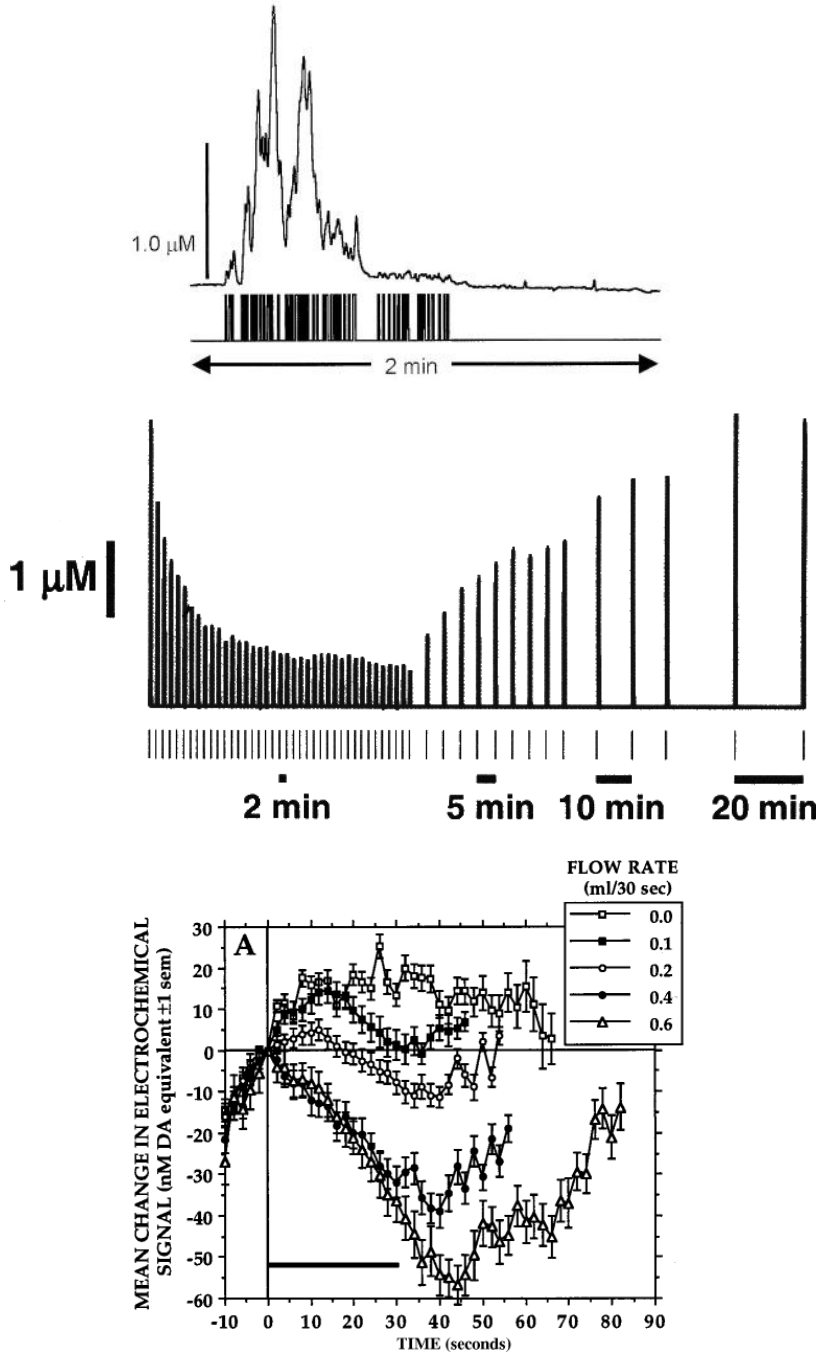

Figure 3.6: Data from three voltammetric recordings of dopaminergic activity in rat striatum. Top: Data from Kilpatrick et al. (2000), showing that, measured in terms of the dopamine concentration in striatum, the dopaminergic response to unpredicted bouts of rewarding medial forebrain stimulation habituates. Middle: Pre-publication data adapted from Montague et al. (2003), courtesy of S. McClure, showing that the level of habituation depends on the stimulation rate. Bottom: Data from Richardson and Gratton (1996), showing putative tonic dopamine levels in rats drinking milk are inversely proportional to the flow rate.

release and reuptake of dopamine. Also, the temporal resolution is not quite sufficient to discern the phasic change in concentration due to a single brain stimulation event. But there is clearly a slow decline in the effectiveness of the phasic stimulation events at releasing dopamine, which we might associate with the slow reduction in the peak height of the error signal during a train of rewards in the simulation. The corresponding reduction in the error signal baseline in Figure 3.3 does not seem to be similarly reflected in the baseline dopamine concentration. Also, where the modeled phasic response strength declines but remains positive, the effect of brain stimulation on striatal dopamine concentrations gradually wanes to essentially nothing. One way of reconciling these issues with the model is to imagine that there is a floor effect in the recordings (or in the complex scaling and release and reuptake dynamics relating neuronal spiking to striatal dopamine concentrations), which has the effect of truncating the response modeled in Figure 3.3.

Data forthcoming from the same laboratory (Montague et al., 2003; Figure 3.6, middle) show that the extent to which dopaminergic release habituates is controlled in a graded fashion by the rate of reward delivery, in accord with the model. These data are from an anesthetized animal, and the timescale of the plot is compressed compared to the figure above it. Each bar under the trace represents a 10-second bout of brain stimulation. In the data, the asymptotic level of habituation for a particular rate of stimulation is inversely related to the stimulation rate. The pattern of results is to be expected under the interpretation that the response habituation reflects the suppressive effect of the average reward estimate ρ_t on the error signal, though again the baseline concentration data do not reflect a baseline effect of ρ_t.

Similar results are seen in an earlier study by Richardson and Gratton (1996), who used a slower and more error-prone voltammetric technique to measure putative dopamine concentrations in striatum, in rats drinking milk delivered at a range of flow rates. These recordings, reproduced in Figure 3.6 (bottom), show a graded, inverse relationship between a slow-timescale concentration change and the rate of reward delivery (here, milk flow speed). It should be stressed how counterintuitive these results are — under previous TD models considering only short-term predictions, they are exactly backwards, since larger TD error would be expected when the animal receives larger amounts of reward (though it is not clear what the "phasic" dopaminergic signal would do during an extended period of continuous reward like those used here). The average reward model would instead suggest that these data reflect the tonic component $-\rho_t$ of the error signal, which would show a similar graded inverse relationship with the flow rate.[3] It is odd that these data would seem to reflect a tonic effect of ρ_t on baseline striatal concentrations while the other voltammetric recordings seemed to reflect only a change in phasic evoked release. However, there are significant methodological differences between the way the two groups recorded and analyzed their data, which could perhaps explain the differences. In fact, the differences might be explained *away* by the fact that the methods used to obtain Figure 3.6 (bottom) were seriously problematic (R. M. Wightman, personal communication, 2002). In particular, the recordings are prone to contamination from chemical activity unrelated to dopamine. Thus, these data, while interesting from the perspective of the theory presented here, must be treated cautiously.

Neurochemical recordings have revealed another dopaminergic response that is extremely puzzling under the original TD models with short-term returns: an apparent *increase* in dopamine activity in response to aversive events such as footshock (see Horvitz, 2000). Assuming these events are equivalent to negative reward, the original TD models would predict only phasic dopaminergic inhibition in these cases and provide no obvious explanation for excitation at any time course. The time course of the measured response is poorly studied but is presumably tonic, as most of the data concerning this point (e.g., Figure 3.7, top), come from microdialysis measurements of chemical markers for dopaminergic activity on an exceptionally slow timescale (one sample per ten minutes or slower). These data are neatly explained under the average reward model, however. As shown in Figure 3.4, the average reward model predicts that aversive events would evoke phasic dopaminergic inhibition followed by a long period of tonic dopaminergic excitation (or, more properly, disinhibition, due to the decrease in the average reward estimate ρ_t). Thus the model would explain the microdialysis results as reflecting the slower, tonic phase of this response.[4]

[3] One might wonder why, from the perspective of the model, dopaminergic concentrations *increase* in the data for milk flow rates of 0.0 and 0.1 ml/30 sec. This is evidently because the standard flow rate was 0.2, which would determine the average reward estimate ρ_t in effect prior to the probe trials using other reward amounts.

[4] Note that simply averaging the error signal over a long period would produce no response at all, since the deeper but quicker phasic depression would exactly cancel the longer-duration but lower-magnitude increase. One solution to this conundrum is that, as already mentioned, the dopamine response is at least somewhat positively rectified, and so is unlikely to reflect the

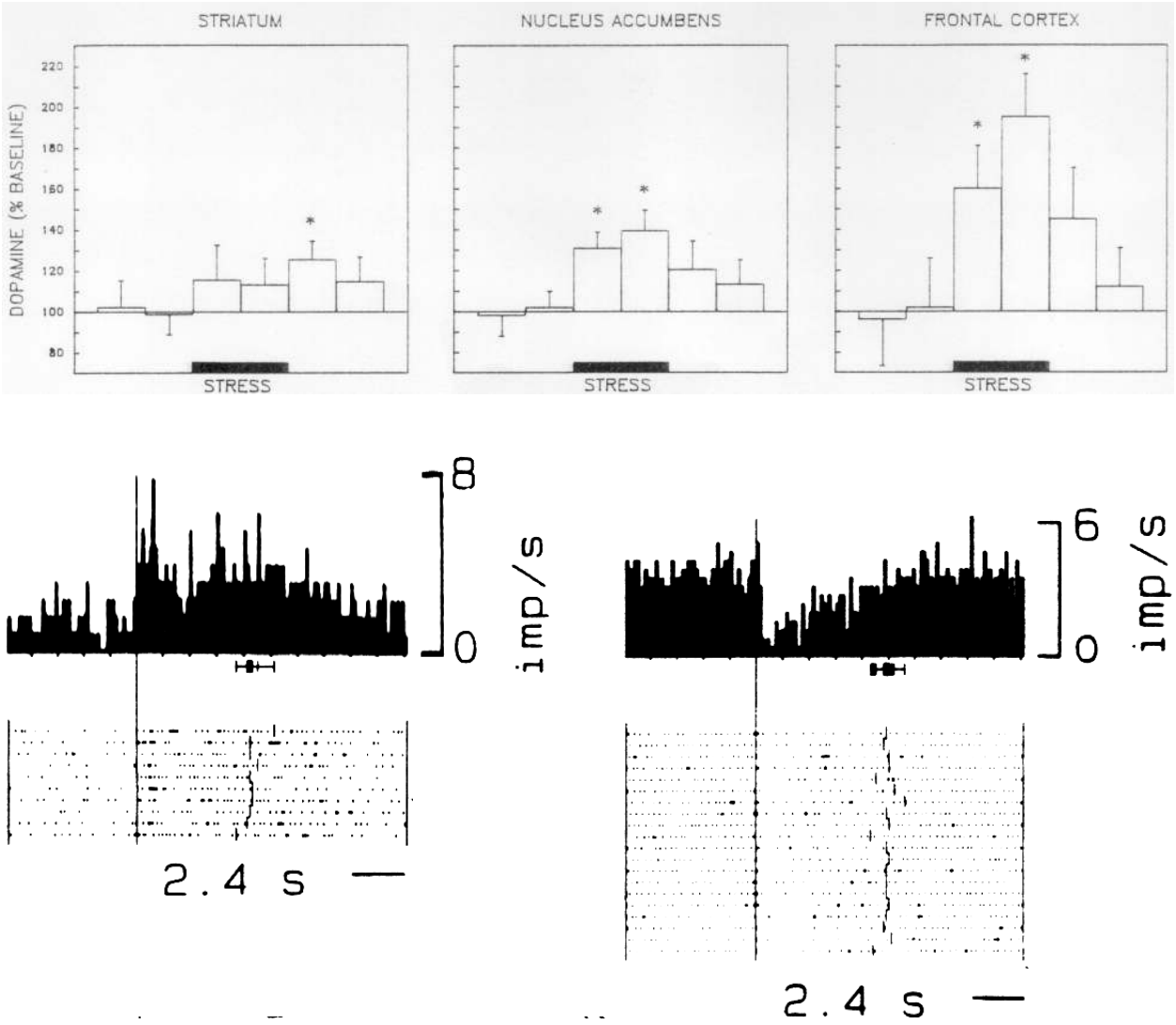

Figure 3.7: Results from two experiments studying the dopaminergic response to aversive stimulation. Top: Extracellular dopamine levels at three dopaminergic targets before, during, and after a 30-minute period of "stress" in which rats were repeatedly subject to tail shock, from Abercrombie et al. (1989). Each bar corresponds to a microdialysis sample collected over 15 minutes. Bottom: Recordings from two dopamine neurons in an anesthetized monkey, while the monkey's face was pinched for about 5 seconds, adapted from Schultz and Romo (1987). One of the neurons responds with a prolonged excitatory response while the other responds with transient inhibition. (Traces are aligned on pinch onset, and vertical bars appear in the rasters for each trial at pinch offset.)

One might hope that the assumptions this explanation makes about the time course of the response could be verified in recordings of dopaminergic spiking in aversive situations, but, as reviewed in Section 2.2.1, the few such recordings are all problematic and not mutually consistent. However, there is evidence both for phasic inhibition (Mirenowicz and Schultz, 1996) and for excitation with a somewhat slower time course (Schultz and Romo, 1987; Guarraci and Kapp, 1999). Schultz and Romo (1987) saw both excitation and inhibition (Figure 3.7, bottom), usually in different neurons. Consonant with the theory and with the illustrated examples, aggregate results they report suggest that the excitatory responses were somewhat longer than the inhibitory ones.

The remaining simulations in the previous sections (Figure 3.5) concerned the acquisition and extinction of a stimulus-reward association. While response to signaled rewards has been studied repeatedly in dopaminergic recordings (see e.g., Schultz, 1998), the predicted effects of ρ_t would not be appreciable. No effects were visible in the experimental data, but the same is true of the model when a slower reward rate is used (Figure 3.2). As far as I know, extinction (as opposed to occasional reward omission) has not been studied, but again, higher-than-usual rates of reward would be required to see the predicted effects.

Note also that there has recently appeared a single datum on slow-timescale dopamine responses in electrophysiological recordings (Fiorillo et al., 2003); the results are reproduced in 4.20 on page 109. This concerns ramping, elevated dopamine responding during the delay period between stimulus and reward in a partial reinforcement experiment (the stimulus predicts that reward will occur at a deterministic time, but only with some probability). Dopamine neurons seem to respond during this period with a ramping tonic elevation that is proportional to the uncertainty in predicted reward (e.g. with a form like $p \cdot (1 - p)$, where p is the chance of reward). There is no obvious relationship between this result and the predictions made here (note that partial reinforcement would actually decrease the magnitude of tonic effects like the one seen in Figure 3.5, top right, since it would reduce dopamine responding). However, this result does bear an interesting relationship to modeling presented in the next chapter about state uncertainty in the TD prediction; I will return to it there.

There remain a few open issues. One is the nature of the tonic signal that I have posited. It may be no coincidence that the data most consistent with this theory mostly derive from neurochemical recordings rather than electrophysiological ones. This may be because only in neurochemical recordings has slow-timescale activity been addressed, or it may be because the tonic part of the signal is computed by processes controlling transmitter release rather than neuronal firing. For the voltammetry experiments measuring dopamine release in response to brain stimulation reward, the experimenters evidently believe that the stimulation causes identical neuronal firing in all cases and the variation in the response is due to some sort of habituation at the level of transmitter release (Montague et al., 2003). (This assumption is based on the idea that the stimulation directly activates dopamine neurons, which may not in fact be the case: Murray and Shizgal, 1996a,b.) In any case, I have presented the predictions in this chapter as though the tonic effects occur at the level of neuronal spiking, though this need not be the case and the computational theory is obviously agnostic about its substrates. Average reward could be computed either internally to the dopamine neurons (through levels of available transmitter, for instance, or intracellular calcium levels) or externally (e.g., as the following section suggests, in the serotonergic system, which might then gate striatal dopamine release), without the effects being visible at the level of dopamine neuron spiking. Experiments that combined electrophysiological recording of dopamine neurons with voltammetry at their targets would be most illuminating in this regard.

3.4 Opponency and serotonin

Here I discuss how the computational considerations discussed thus far intersect with some psychological theory and data concerning the opponency between signals representing different timescales of prediction, and between appetitive and aversive motivational systems. I also discuss neurophysiological data concerning the representation of negative error in the dopamine signal. In light of these computational, psychological, and neurophysiological concerns, I propose a family of models in which the dorsal raphe serotonin signal

full magnitude of any phasic inhibition. Further, the nonlinearities of dopamine release and reuptake may favor the excitatory response over the inhibitory, and the dialysis method may also be selectively sensitive to the slower timescale of the excitation.

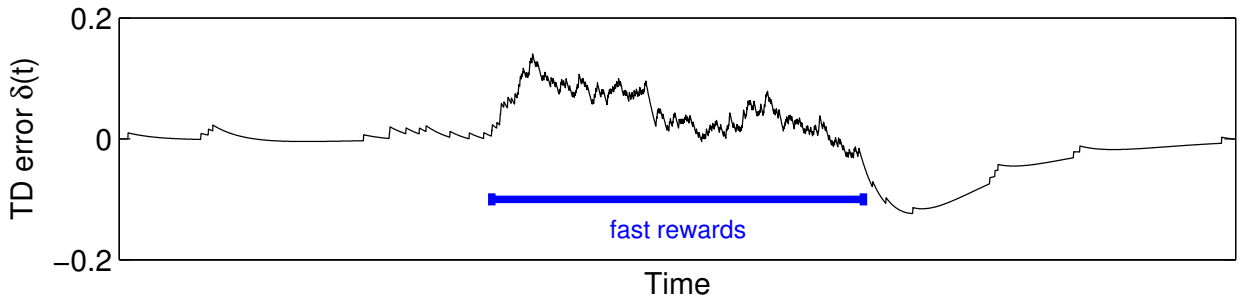

Figure 3.8: The prediction error trace from Figure 3.3 (top), low-pass filtered by convolution with a geometric kernel to simulate the effect of indirect measurement of the error signal through some correlated variable such as heart rate. The period when rewards are delivered at an increased rate is marked with a bar. Like the opponent motivational response modeled by Solomon and Corbit (1974) (Figure 2.5 on page 31), this signal responds to the onset and rebounds at the offset of the high rate reward delivery period, with habituation following each response.

serves as an opponent to dopamine, and explore the predicted behavior of the two systems under this theory. I end with discussions of several different empirical issues raised by the theory.

3.4.1 Opponency between timescales in the model of Solomon and Corbit (1974)

The predictions of the average reward model, as shown in Figure 3.3, relate to an influential psychological theory, the opponent process theory of motivation of Solomon and Corbit (1974), and to the behavioral data underlying it. That article, reviewed more fully in Section 2.3.2, argues that affective responses to motivationally significant events measured many different ways and in many different situations follow a canonical pattern involving response, habituation, rebound, and rehabituation. For instance, the events might be a series of juice squirts delivered to a thirsty monkey, and the measured affective response could be the monkey's heart rate before, during, and after the train of rewards. In Solomon and Corbit's model, illustrated in Figure 2.5 on page 31, the observed response dynamics result from the competition between two opponent representations of the reward rate that adapt at different timescales. Thus the model proposes that animal behavior reflects an opponency between different *timescales* of prediction or representation.

If we (crudely) take δ_t as determining the affective response, then the average reward estimate ρ_t can be viewed as playing a role similar to slow-timescale opponent in Solomon and Corbit's model. (Their fast-timescale opponent corresponds to the rewards r_t themselves.) Because of the slow changes in ρ_t, the overall error signal shows compensation and rebound dynamics similar to the Solomon and Corbit model (and to the data it simulates). The main difference between the original model (Figure 2.5 on page 31) and the average reward version of Figure 3.3 is that, in the TD model, δ_t contains positive impulses for each reward. This feature would seem to better reflect the pulsatile nature of the rewards than the smooth envelope assumed by Solomon and Corbit. Of course, the measured motivational response, such as heart rate, might not be sufficiently dynamic to track these quick changes in the underlying error signal; in this case it could resemble a low-pass filtered version of δ_t (Figure 3.8), which has essentially the same features as Solomon and Corbit's version.

Thus, in the average reward TD model, Solomon and Corbit-style opponency between timescales of prediction takes place in the computation of the error signal δ_t, through the subtraction of the long-timescale expected reward ρ_t from the immediate observed reward r_t. This model also raises a question about the neural substrates for this opponency: what neural system is responsible for tracking and reporting the average reward signal ρ_t as part of the overall computation of the error signal? In a discounted model, future average reward information is encoded in the value estimates, so this information may just be implicitly reported by the same systems that code for $\widehat{\mathbf{V}}$. In the average reward model it is a separate signal. I will shortly lay out a model in which the dorsal raphe serotonin system fulfills this role.

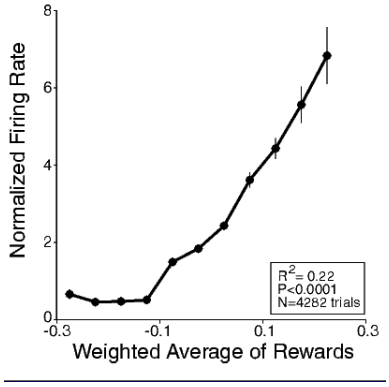

Figure 3.9: Average firing rates in response to reward in 11 dopamine neurons, as a function of the estimated prediction error engendered by the reward. Prediction error and neuronal firing rate are proportional, but negative error is rectified below some threshhold. Unpublished data from Bayer and Glimcher (2002), courtesy of H. M. Bayer.

3.4.2 Negative error and opponency between appetitive and aversive systems

A separate thread of psychological data and theory about opponency concerns that between appetitive and aversive motivational systems. Data from conditioning experiments involving combinations of both appetitive and aversive events, discussed in Section 2.3.2 and reviewed more thoroughly by Dickinson and Balleine (2002), suggest that conditioned responding can reasonably be viewed as arising from the interaction between a single aversive and a single appetitive channel, rather than one combined system, or multiple systems of each type. Behavioral data from many experimental paradigms (involving for instance appetitive/aversive interactions, conditioned inhibition, and extinction) are well explained by conditioning models that incorporate this sort of motivational opponency (notably, Grossberg, 1984, 1988, 2000; Grossberg and Schmajuk, 1987).

In the context of the dopaminergic models discussed in this chapter, these psychological considerations raise the question of whether the dopamine system is involved in processing aversive as well as appetitive events. A related question is whether (or to what extent) the firing of dopamine neurons represents *negative* prediction error, which in the computational theory can arise both in aversive situations, as when an unexpected shock occurs, and in appetitive situations, e.g. when a received reward is smaller than expected or an expected reward is omitted altogether.

There are two pieces of physiological data that bear on this question. First, dopamine neurons show a well-timed pause in their background firing at the time of omitted reward (see, e.g., Figure 2.2 on page 16), suggesting that negative error is represented by firing rate excursions below baseline. However, as the baseline firing rates are already quite low, it is unlikely that the signal has sufficient dynamic range to effectively represent the full range of possible negative errors. This issue was studied directly by Bayer and Glimcher (2002) who repeatedly delivered squirts of different volumes of juice to a thirsty monkey, and also included trials in which juice was altogether omitted. Using a TD-like model, the experimenters computed the prediction error that the monkey's putative TD prediction systems should have experienced with each bout of juice (a function of the juice amount compared to the amounts delivered on previous trials), and compared this error to the firing rates of dopamine neurons responding to the rewards. As shown in Figure 3.9, the data show that firing rate is proportional to prediction error over a range of errors (the diagonal line in the figure), but that the negative error below some threshhold is rectified, as seen in the horizontal portion of the plot.

Thus the Bayer and Glimcher (2002) data suggest that a separate neural signal is needed to represent the missing negative portion of the error signal, and psychological considerations similarly suggest that the dopamine signal should have a motivational opponent. In the following section I lay out a model in which the dorsal raphe serotonin system plays this role.

3.4.3 Model of serotonin as a dopaminergic opponent

The preceding sections discussed two sorts of opponency that arise in the computation of the prediction error δ_t. The average-reward estimate ρ_t is subtracted from the error signal and represents a slow-timescale opponent to the immediate report r_t of received rewards. A motivational opponency also exists between rewards and punishments, which are both reported through the term r_t. (In this section it will be useful to introduce new notations to refer to rewards r_t^+ and punishments r_t^- separately, with $r_t = r_t^+ - r_t^-$, and similarly for average rewards ρ_t^+ and average punishments ρ_t^-.) Thus the subtracted punishments r_t^- are an aversive opponent to the appetitive reward signal r_t^+. Another source of negative prediction error is omitted reward, which enters the signal through a negative difference $\widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t}$. This section, following modeling originally published by Daw et al. (2002b), proposes a family of models in which the dorsal raphe serotonin system is responsible for reporting some or all of these sources of negative error, and thus plays the role of an opponent to the dopaminergic prediction system. The motivation for this idea is that the dorsal raphe serotonin system seems, on the basis of a number of sorts of physiological evidence discussed in Section 2.2.3, to act as an opponent to dopamine.

I consider the family of models in which the complete error signal is reported as the difference between a dopaminergic error signal δ_t^{DA} and a serotonergic error signal δ_t^{5HT}:

$$\begin{aligned} \delta_t^{DA} - \delta_t^{5HT} &= \delta_t \\ &= (r_t^+ - r_t^-) - (\rho_t^+ - \rho_t^-) + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \end{aligned}$$

This general scheme can be instantiated in many different ways, depending on how the full error is shared between the two opponents. In this section I lay out two specific schemes for how the error signal might be divided up, one in which the serotonergic error signal contains only tonic information, and another in which it contains phasic information as well. Each signal might have its own scaling factor, and be delivered against its own arbitrary background firing rate, two effects I omit from all equations. For simplicity, I will also assume that the channels are separate rather than interacting, i.e. that the channels independently report their portions of the error signal to dopaminergic targets.

A simple, initial model would envision that the serotonergic channel carries only tonic information. Thus we can take:

$$\delta_t^{DA} = (r_t^+ - r_t^-) + \rho_t^- + \widehat{\mathbf{V}}_{s_{t+1}} - \widehat{\mathbf{V}}_{s_t} \tag{3.5}$$

$$\delta_t^{5HT} = \rho_t^+ \tag{3.6}$$

That is, serotonin reports the tonic average reward signal, while the dopaminergic error signal reports the remaining terms of δ_t. In this model, serotonin serves as a slow-timescale opponent to dopamine, implementing the notion of timescale opponency from the model of Solomon and Corbit (1974).

A problem with this simple model is that it does not account for the finding of Bayer and Glimcher (2002) that the phasic dopamine signal is rectified, since δ_t^{DA} carries the entire phasic portion of the error. Under the present scheme, serotonin can also be granted responsibility for some portion of the negative phasic error, so that it serves as both a slow-timescale opponent (Solomon and Corbit, 1974) *and* a motivational opponent to dopamine. This is an even more speculative (but more interesting) hypothesis than the previous one. There are of course a variety of schemes by which the phasic error could be divided up. As a simple starting point, I assume a perfectly symmetrical division. This model constructs both opponent signals by blending together both the positively rectified error signal $[\delta_t]_+$ and the negatively rectified signal $[\delta_t]_-$, differently scaled:

$$\delta_t^{DA} = \alpha[\delta_t]_+ - (1-\alpha)[\delta_t]_- \tag{3.7}$$

$$\delta_t^{5HT} = \alpha[\delta_t]_- - (1-\alpha)[\delta_t]_+ \tag{3.8}$$

This model assumes that a small background firing rate is added to each channel, so inhibitory signals would appear as phasic inhibition. The parameter α controls the degree to which both negatively and positively rectified information occur in each signal. For $\alpha = 1$, dopamine and serotonin are just the positively and negatively rectified signals, but when $\alpha < 1$, the dopamine channel contains some of the negatively rectified

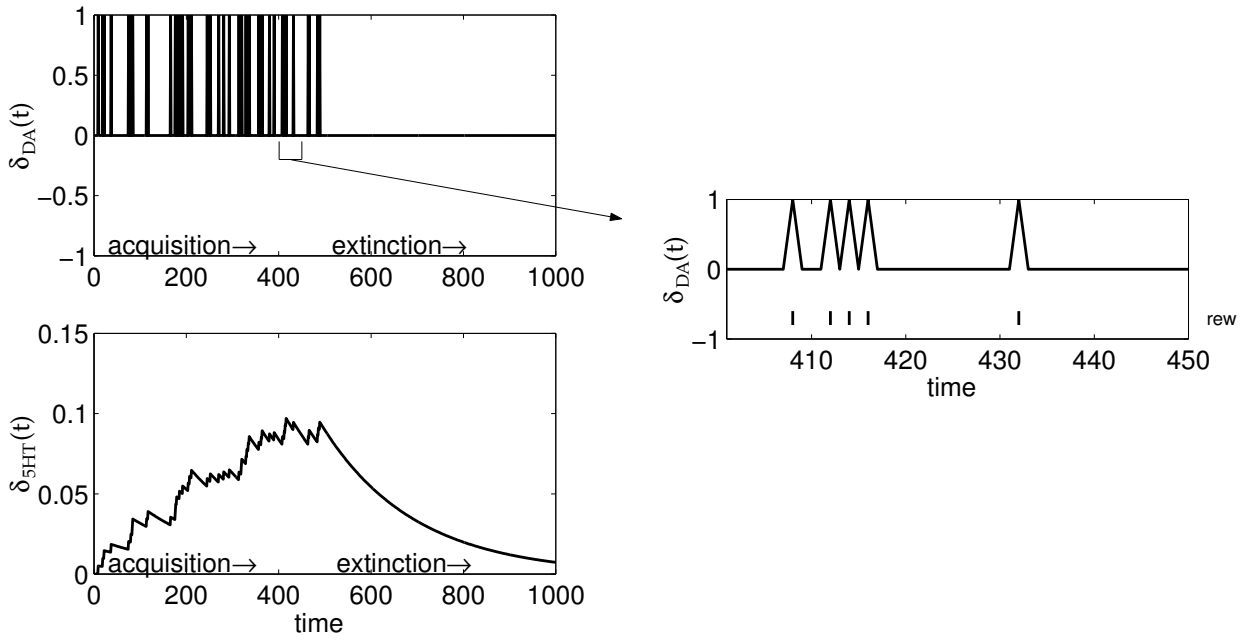

Figure 3.10: Modeled dopaminergic and serotonergic signals during delivery of unsignaled rewards on a Poisson schedule. This is the same situation as modeled in Figure 3.3, but serotonin is assumed to carry the tonic average reward portion of the signal.

signal as well, scaled by $1-\alpha$ (and similarly for positive error and serotonin). So in this model, both negative and positive information can occur in both channels, but differently scaled.

3.4.4 Results of simulations of the model

Here I present some simulation results to demonstrate how the dopaminergic and serotonergic signals would be expected to behave under the models described in the previous section. In particular, I show results for the model in which serotonin carries only tonic information, as well as for the full model in which serotonin carries both tonic and phasic information.

Figures 3.10 and 3.11 show the predicted behavior of the dopaminergic and serotonergic channels in two experiments (unsignaled and signaled rewards) in the model of Equations 3.5 and 3.6, in which serotonin carries only tonic information. These are similar to the results reported previously for the average reward model of dopamine, but the tonic effects of the average reward estimate have been moved into the serotonergic channel and inverted.

I turn now to the model of Equations 3.7 and 3.8, in which the serotonin signal carries both positive and negative error. Figure 3.12 demonstrates how the two error signals are produced by blending of scaled positive and negative error, in a situation where a stimulus signals a reward, which is then omitted. In the model, reward omission causes negative phasic error, which activates the serotonin system phasically. As we know, reward omission also inhibits dopamine neurons. This would not be seen in a strictly rectified model (Figure 3.12, left, in which $\alpha = 1$ and thus no negative error is included in the dopamine signal), but does occur for $\alpha < 1$, as shown on the right side of the figure.

Since the two channels are perfectly symmetric in this model, the same behavior illustrated in the figure would be seen in an aversive conditioning experiment, but with the channels' roles reversed. Thus a signal predicting future shock would be expected to phasically excite the serotonergic channel; when the punishment failed to occur, dopamine would be excited. This is an instance of the theory's more general prediction that dorsal raphe serotonin neurons in aversive conditioning should respond analogously to the well-known behavior of dopamine neurons in appetitive conditioning experiments (Figure 2.2 on page 16) — with a phasic excitation to unexpected shocks that would transfer through conditioning to stimuli that

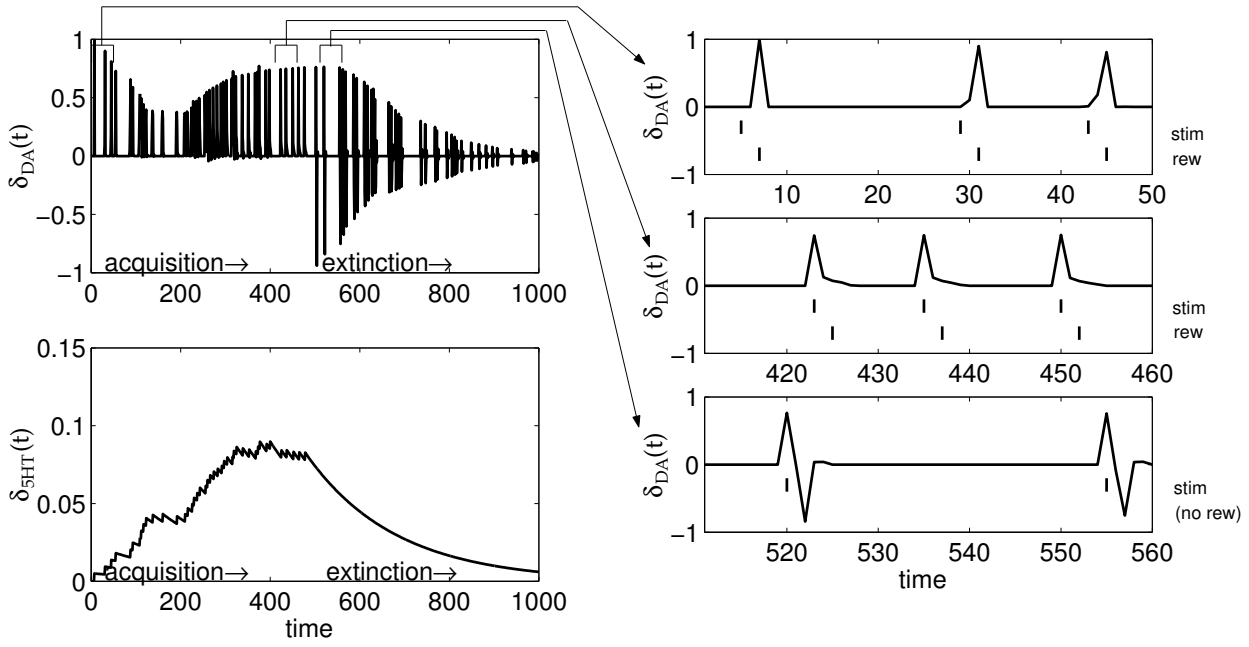

Figure 3.11: Modeled dopaminergic and serotonergic signals during an experiment in which rewards are signaled by a prior stimulus. The dopaminergic system shows the conventional phasic response transfer, but the tonic part of the signal (putatively carried by serotonin) is the same as in the unsignaled reward situation of Figure 3.10.

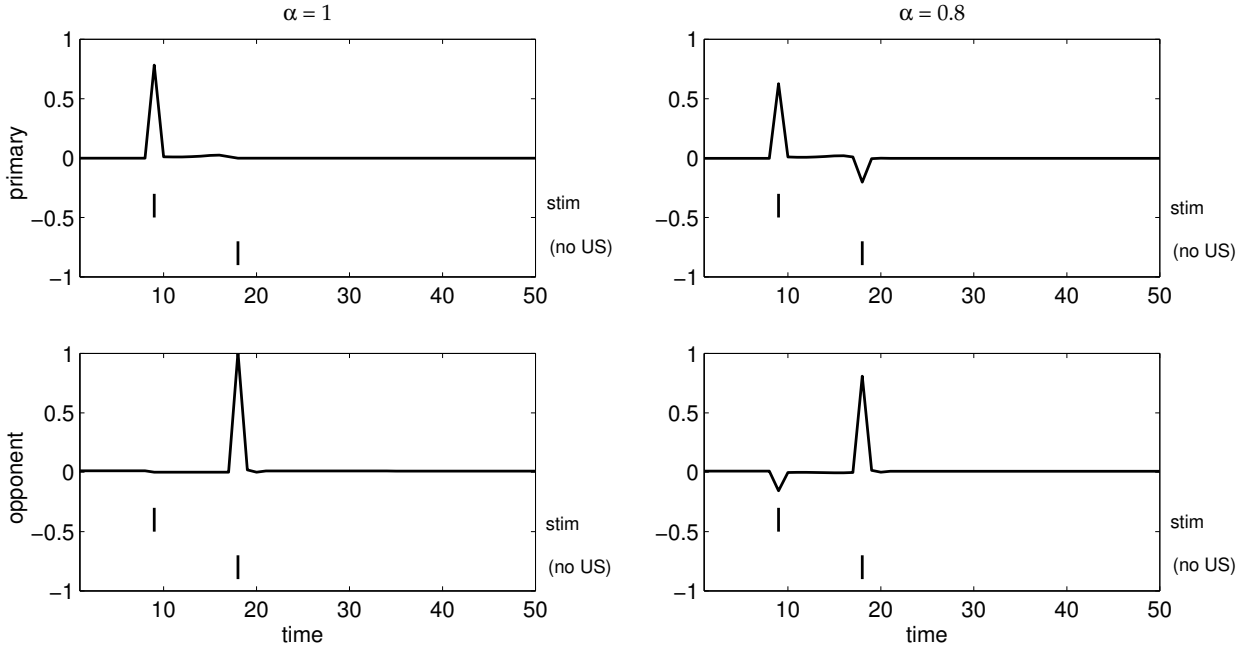

Figure 3.12: Blending of positive and negative error in both dopaminergic and serotonergic systems demonstrated in a situation where a signaled reward is omitted. For appetitive conditioning, the traces labeled "Primary" are dopamine and the traces labeled "Opponent" are serotonin; these roles are reversed in aversive conditioning. If the signals are fully rectified ($\alpha = 1$, left), then dopaminergic pausing is not seen to missed reward. If some negative error is blended into the positive channel ($\alpha = 0.8$, right), then the problem is solved.

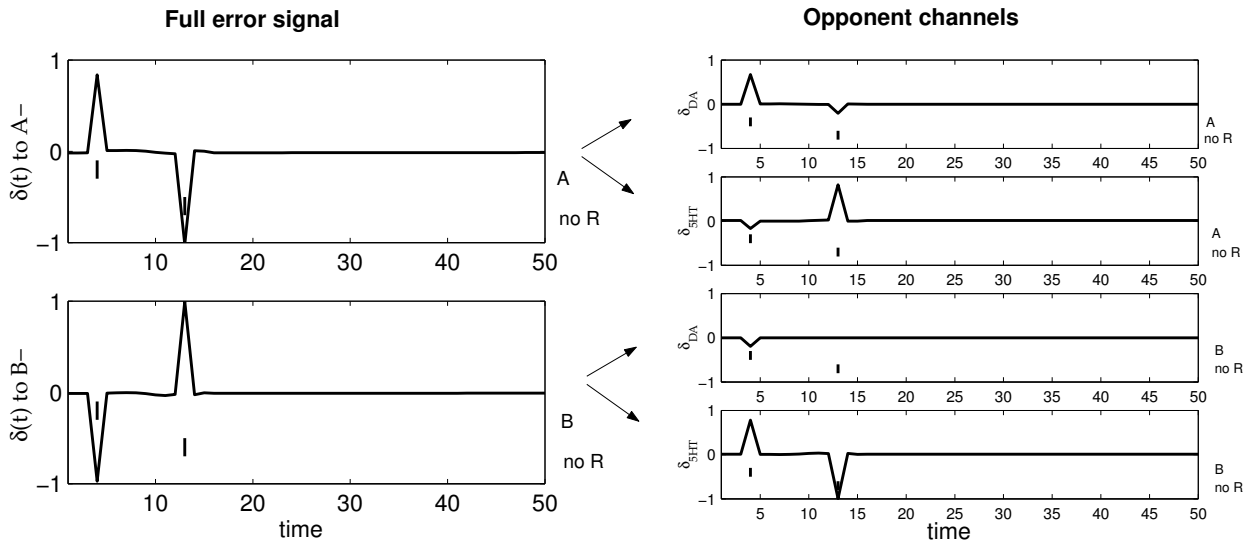

Figure 3.13: Modeled dopaminergic and serotonergic responses during a conditioned inhibition stimulus. We consider probe trials in which either the conditioned excitor A or inhibitor B are delivered alone and unrewarded. The left side plots show the full error signal in these two cases; on the right, these signals are decomposed according to Equations 3.7 and 3.8 with $\alpha = 0.8$, and modified for consistency with the experimental results of Tobler et al. (2001).

predict them.

An experiment that would be expected to exercise both positive and negative opponent channels is conditioned inhibition (Rescorla, 1969). In this, a standard experiment in classical conditioning, an animal learns that while some stimulus A (a "conditioned excitor") predicts reward, it does not do so when it is presented together with some other stimulus B (the "conditioned inhibitor"). The inhibitory properties of B can be tested in a number of ways, for instance through a summation test in which B is presented together with a second excitor C, assessing whether this combination reduces the reward expectation normally accompanying C. The left column of Figure 3.13 depicts TD prediction error in a simulation of two probe trials that might be performed after training on a conditioned inhibition task. The top trace illustrates the presentation of the excitatory stimulus A alone but unrewarded, and is just the standard TD error for reward omission we have repeatedly seen previously: A induces positive error, followed by negative error at the time the reward should have arrived. The lower trace shows TD error when the conditioned inhibitor B is presented alone: it is exactly the negative image of the excitor. This is because the (negative) value prediction carried by the conditioned inhibitor is exactly opposite that carried by the conditioned excitor, since they must cancel to zero when presented together, assuming as we do an additive model for aggregate predictions. So the presentation of a conditioned inhibitor alone should produce negative TD error (because it signals a reduction in future reward availability), but followed by *positive* TD error. This error occurs at the time a reward would normally have been omitted, had B been paired with A: not missing a reward is a pleasant surprise.

The right column of the figure shows the two error traces divided between dopaminergic and serotonergic channels. For this, I have used an ad-hoc modification to the division predicted by Equations 3.7 and 3.8. In particular, I have assumed that the positive error at the time a reward would have been omitted (after presentation of B) is reported by serotonergic inhibition rather than dopaminergic excitation. This is for consistency with the results of dopamine recordings by Tobler et al. (2001), which did not reveal the expected dopaminergic excitation in this condition. Of course, the general family of opponent models described in the previous section accommodates such rearrangement of the phasic error between the two channels. I will further discuss this point in the following section.

3.4.5 Discussion: Experimental evidence on serotonergic firing

The general argument for this model is a bit unusual and indirect. The idea is to combine our relatively strong empirical and theoretical understanding of dopaminergic physiology with data about how dopamine and serotonin interact and about some broad functional effects of serotonin, in order to make inferences about the computational role and the detailed behavior of serotonergic neurons — issues about which there is little direct evidence. Thus, particularly with regard to the firing of serotonergic neurons, the suggestions given above go far beyond the available data.

In fact, on a superficial reading, the literature on serotonergic recordings would appear to directly contradict one of the key predictions of the model: the hypothesized phasic responses to aversive USs. Jacobs and Fornal (1997, 1999) review their laboratory's early recordings in cats, and report no significant serotonergic response to aversive events. Mason and collaborators reached a similar conclusion (Mason, 1997; Gao et al., 1997, 1998) from their more recent recordings, which were in the descending spinal cord-directed serotonergic system in rats — a separate group of neurons from the system described in the present theory. These negative reports are not conclusive, however, since both groups appeared to be searching for *tonic* changes in firing rates. In general, this was accomplished by examining neuronal firing during a single trial, often using firing rates smoothed with a coarse temporal average rather than raw spike trains. The subtler phasic responses predicted here might very well have escaped such analysis — they are envisioned as being similar to phasic dopaminergic responses, which emerge in the average over many trials from weak responses that, in a single trial, consist of only perhaps an extra spike or two. Mason and collaborators (Mason, 1997; Gao et al., 1997, 1998) in fact report in passing that subtle, transient responses are seen to aversive events such as tail heat or foot pinch in 25-50% of the serotonergic neurons they record. They base their bottom-line conclusion that the serotonergic neurons are *unresponsive* to aversive stimulation on the fact that *non*serotonergic neurons recorded in the same nucleus were much more responsive, showing very strong tonic excitation or inhibition. However, weak, transient responses are precisely what the present modeling predicts for serotonergic neurons.

Because of the poverty of available recordings of serotonergic neurons, a conclusive test of the theory's predictions with regard to phasic serotonergic responding awaits a more systematic study of the behavior of the neurons over a series of controlled aversive conditioning trials, patterned after the studies of dopamine neurons in appetitive conditioning. More generally, the modeling in this section strongly calls for a series of matched recording experiments in both serotonergic and dopaminergic systems, studying both tonic and phasic activity in both systems over a range of appetitive and aversive experiments designed to induce both positive and negative prediction error.

The work of Gao et al. (1998) on the effects of morphine on the responses recorded in the descending serotonin system suggests a further possible explanation for why the serotonergic responses predicted in the present model have not been observed. While those authors review a range of data that morphine analgesia is mediated by serotonergic projections from that system into the spinal cord, they found that morphine affected the firing rates of *non*serotonergic neurons in the nucleus, rather than of serotonin neurons. This pattern of data suggests that serotonin release at the targets might thus be *gated* by the activity of the nonserotonergic neurons, and that the serotonergic signal (in this descending system) is thus dissociated from the firing of the serotonin neurons themselves. If such a mechanism also exists in the ascending system of the dorsal raphe, then the "serotonergic" responses hypothesized by the present theory might instead be seen in nonserotonergic neurons that perform the gating function.

Finally, the lack of data about phasic responses in serotonin neurons was the motivation for the less radical model of Equations 3.5 and 3.6, in which the serotonergic signal carries only tonic information. Though the hypothesis of this model that serotonin neurons should be tonically modulated by the average reward signal has also never been tested, it seems plausible given that these neurons are known to be tonically modulated by similar factors such as arousal (Jacobs and Fornal, 1997, 1999).

3.4.6 Discussion: The hypothesized dopamine response and rectification of negative error

The dopamine response in the present model is roughly similar to that modeled in Section 3.3, and so the experimental considerations discussed there mostly carry over to this model. The two-channel model in this

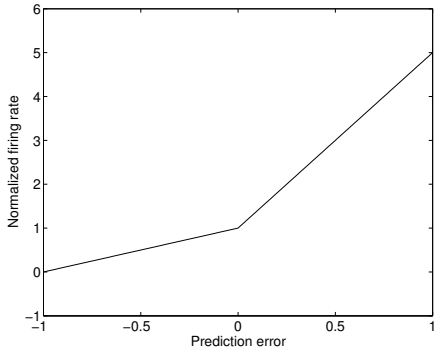

Figure 3.14: Plot of firing rate against prediction error in the blended model of Equations 3.7 and 3.8, with $\alpha = 0.8$. Unlike the data reported by Bayer and Glimcher (2002), the slope of the line is different for positive and negative prediction errors, with an inflection at zero prediction error.

section expands and improves on the basic average reward model by including an account of the scaling and rectification of negative phasic error, which I discuss at length below.

I will first say a few words about why the other apparent difference between the models is less than meets the eye. In the simple model of Equations 3.5 and 3.6, the average reward estimate ρ_t^+ is supposed to excite the serotonin signal rather than inhibit the dopamine signal. If this is the case, an increase in ρ_t^+ could not explain the habituation of dopamine release, as proposed before. (Note that in this model, the tonic *punishment* signal ρ_t^- is still included in the dopaminergic channel, so the explanation for the dopaminergic response to aversive events holds up. Also, this is not an issue in the version of the model from Equations 3.7 and 3.8, where the signals are built up by blending both rectified signals.) This difference between the theories is really just a symptom of an oversimplification in the model of dopamine/serotonin interactions assumed here. It is not in actuality the case that the two channels are strictly separate and their opponency is expressed only through opposing effects on target structures, as I have assumed. Instead, there is evidence that serotonin inhibits the dopamine neurons themselves (Kapur and Remington, 1996). Thus if ρ_t^+ excites serotonin neurons, this presumably also has the effect of inhibiting dopamine neurons. A more realistic model based on Equations 3.5 and 3.6 but including this interaction would make all the same predictions about tonic dopamine as the model from Section 3.

I turn now to the account of error rectification in the model of Equations 3.7 and 3.8, and in particular to its relationship with the experimental results of Bayer and Glimcher (2002), which are reproduced in Figure 3.9. Those data (and more general concerns about the low baseline firing rates of dopamine neurons) raise an important question of how the brain represents information about negative TD error that is evidently missing from the dopamine signal. The present theory answers this question by hypothesizing that the dorsal raphe serotonin signal plays this role. While this general hypothesis is viable, some details of the rectification and scaling scheme I have presented do not match the observations of Bayer and Glimcher (2002). Here I analyze these issues and discuss alternative schemes for division of the signal and their implications for understanding dopaminergic responses. Important to this discussion will be issues of how firing rates are measured by averaging over trial-by-trial spike counts, the details of which will also play an important role in Chapter 4.

The error rectification and scaling scheme of Equations 3.7 and 3.8 is the one proposed in Daw et al. (2002b). To relate the error signals δ^{DA} and δ^{5HT} to dopaminergic and serotonergic firing rates, we can assume that the firing rates are proportional to the error signals, added to a baseline firing rate b. In principle, since δ^{DA} and δ^{5HT} can have arbitrarily large negative magnitude while firing rates are bounded below by zero, this process could involve truncating errors less than $-b$. However, the intent of the scaling scheme was that α should be chosen so that no such truncation would occur over the relevant range of errors. (It is unclear why the negative error should be rescaled, except to match the available dynamic range of the signal.) Figure 3.14 shows dopaminergic firing rate as a function of prediction error on this scheme, presented in the same form as Bayer and Glimcher's (2002) data. The difference in scaling between negative

and positive errors is visible here as a change in the slope of the line relating prediction error and firing rate, at $\delta = 0$. For more sophisticated, nonlinear gain control schemes (such as sigmoidal scaling), the function would curve rather than bend abruptly.

Comparing the Bayer and Glimcher (2002) (Figure 3.9), we see no such change in the slope of the function for negative errors. Instead, firing rates appear proportional to prediction error, with a common slope for positive and negative errors, but only down to a (slightly negative) error threshhold corresponding to a near-zero firing rate, where the function flattens out abruptly. This suggests that the negative and positive errors are scaled the same, but the function is truncated when the firing rate reaches zero.

These data suggest that the blending scheme should be replaced with a model in which dopamine reports the full positive and negative error down to some threshhold $-b$, which is slightly below zero: $\delta_t^{DA} = [\delta_t + b]_+$. In this case, the serotonergic signal would report the remainder of the negative error signal, $\delta_t^{5HT} = [\delta_t + b]_-$. An alternative is $\delta_t^{5HT} = [\delta_t - b]_-$, in which case the opponent signals would be symmetric but overlapping in their negative portions (requiring relaxation of the model constraint that the two signals sum to δ_t). In this case, all of the simulation figures in the previous section would appear qualitatively the same.

There is some further insight to be gained by considering a lower-level model, which describes how trial-by-trial spike counts (measured during some short period after the reward) rather than mean firing rates might covary with prediction error. Reanalysis of the Bayer and Glimcher (2002) data in terms of median rather than mean firing rates showed that zero error corresponds to very near zero median firing rate (H. M. Bayer, personal communication, 2002). One way to explain this result is to assume that in a given trial, the error δ_t is corrupted by zero-mean Gaussian noise prior to rectification. If we rectify this noisy error with a threshhold of zero, dividing the positive and negative portions between dopaminergic and serotonergic channels, we have:

$$\delta_t^{DA} = [\delta_t + \nu_t]_+$$
$$\delta_t^{5HT} = [\delta_t + \nu_t]_-$$
$$\nu_t \sim N(0, s)$$

When $\delta_t = 0$, δ_t^{DA} is distributed as a truncated half-Gaussian (with elevated probability mass at $\delta_t^{DA} = 0$ accounting for the portion of the Gaussian that lies in negative territory). I assume that this error distribution is discretized to produce a distribution over per-trial spike counts. Averaged over many such trials, dopaminergic neurons would show a positive mean firing rate (Figure 3.15, top). Moreover, at times of negative error, less of the Gaussian lies in positive territory, so the mean firing rate declines (Figure 3.15, middle). But to reiterate, in this model, there is no constant baseline term explicitly added to the dopamine signal; instead, zero firing corresponds to zero error, and the elevated background firing rate and the pauses beneath it are an artifact of the averaging over trials of asymmetrically rectified noise.

This account suggests a new explanation for the dopaminergic response to novelty. If a novel event increased uncertainty in the value prediction V_t , and this uncertainty was reflected as increased variance in the prediction error, then the mean dopaminergic (and serotonergic) firing rates to the novel event would increase, even though the true prediction error remained zero (Figure 3.15, bottom). Unlike previous accounts of this phenomenon (Kakade and Dayan, 2001a, 2002b; Suri and Schultz, 1999), this scheme does not require that a novelty signal be separately added as a special case to the dopamine signal. However, unlike Kakade and Dayan's (2002b; 2000) model, the new account does not explain their (to my knowledge, anecdotal) observation that dopaminergic excitation to novelty is often followed by inhibition. Moreover, implementing this new account would require producing a system that tracked uncertainty in the value estimate V_t, which is technically difficult because of the function's recursive construction. A simpler approach, which might still be capable of instantiating the idea presented here, could use machinery introduced in Chapter 4 to track uncertainty in the state estimate s_t, which can cause uncertainty in the value estimate and, through it, the error signal.

3.4.7 Discussion: Conditioned inhibition and opponent value representation

Finally, I discuss several considerations about the model related to the conditioned inhibition experiment simulated in Figure 3.13, and in light of the dopaminergic recordings in this task reported by Tobler et al. (2001).

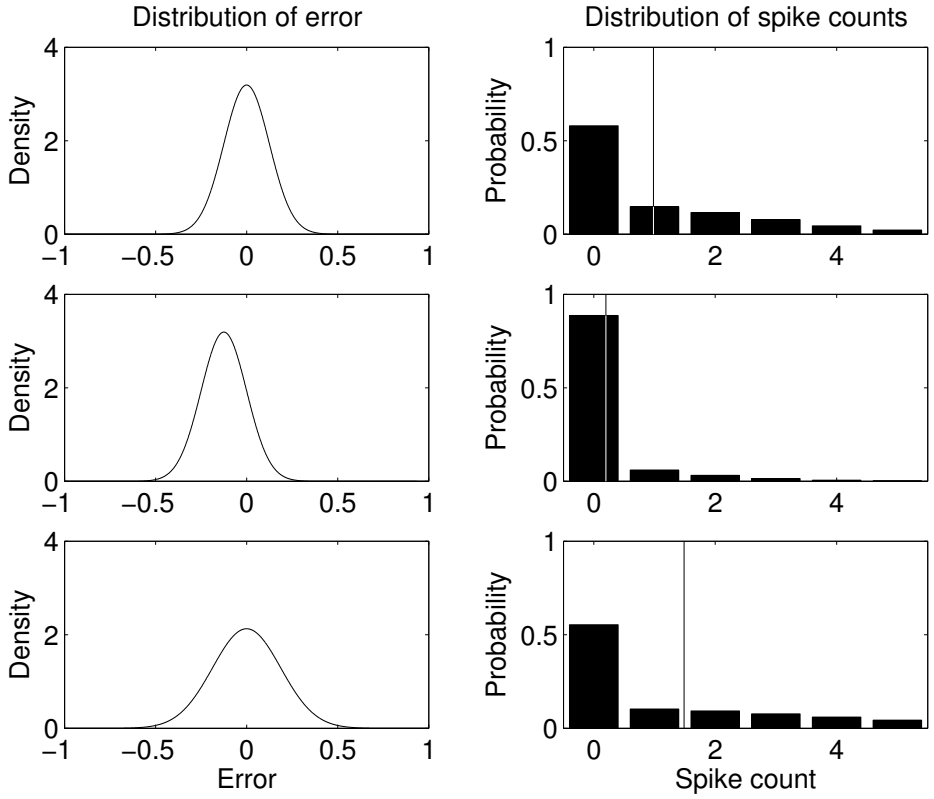

Figure 3.15: Illustration of a model in which the true TD error is affected by zero-mean Gaussian noise prior to rectification. In each row, the left trace shows the distribution of TD errors plus noise in some situation and the right barplot the corresponding distribution of dopaminergic spike counts. A vertical line indicates the mean spike count. Top: Distributions of noisy TD error and dopaminergic spike counts in trials where the true TD error equals zero. Though the mean TD error is zero, the mean spike count is positive — producing a nonzero background firing rate. Middle: Distributions of noisy TD error and dopaminergic spike counts in trials where the true TD error equals -1/8. Because greater probability mass is occupied by negative errors, the mean spike count is depressed here (compared to the top plot where $\delta_t = 0$), producing inhibition of the background firing rate. Bottom: Distributions of noisy TD error and dopaminergic spike counts in a situation where the true TD error is zero but the variance in its noise is elevated, demonstrating a possible explanation for the dopaminergic response to novel stimuli. Compared to the normal zero-error situation illustrated in the top plot, the average spike count is elevated in this case, producing phasic neuronal excitation.

The key condition of interest is the presentation of the conditioned inhibitor B, alone and unrewarded. The modeled TD error in this situation is illustrated in the bottom left trace of Figure 3.13, which shows negative phasic TD error to B followed by positive error at the time a reward should have been omitted. According to the model of Equations 3.7 and 3.8, these TD errors should be reflected in dopaminergic firing rates by phasic inhibition (to B) followed by later phasic excitation (at the time a reward should be omitted). In the recordings of Tobler et al. (2001), the response of dopamine neurons to B is indeed often inhibitory. (Often they are also initially excited, so that the full response follows a burst-pause profile. Further experiments showed the excitatory phase was due to generalization between B and the conditioned excitor A, a phenomenon discussed at some length in Section 2.2.1). But at the time a reward should have been omitted, no subsequent dopaminergic excitation was seen, contradicting the model.[5]

In the opponent scheme, positive error can be reported either by excitation of the appetitive, dopaminergic channel, or by inhibition of the aversive, serotonergic channel. Thus, to account for these data, in the simulations of Figure 3.13 (bottom right), I have assumed that the missing dopaminergic excitation is replaced by serotonergic inhibition. This hypothesis offers a clear prediction that could be tested with serotonergic recordings. However, the assumption is entirely ad hoc: I have no explanation *why* the error signal should be divided up this way. That said, there are not strong normative foundations for the precise division described by Equations 3.7 and 3.8 either, beyond the general justification of using rectification and opponency to represent both positive and negative errors with firing rates.

There is another question raised by this ad hoc assumption, which has an easier answer and relates to some richer models that could be explored in future work. The question is *how* the system could arrange for the positive error in that particular situation to be reported by the serotonergic channel, when all other positive error, regardless of origin, is accompanied by dopaminergic excitation. How can this situation be special? The question can be answered in the context of a common extension of the opponent representational framework that I have no so far discussed. In many opponent schemes (e.g. those of Grossberg and collaborators; Grossberg, 1984, 1988, 2000; Grossberg and Schmajuk, 1987), there is a separation not just between appetitive and aversive error signals, but also between appetitive and aversive predictions, which are each represented with their own set of weights. In the TD model, we can represent that value function using two sets of weights, $\mathbf{w}^+$ and $\mathbf{w}^-$. These weights can each be associated more or less exclusively with one of the opponent channels, perhaps trained primarily by its error signal. The weights represent opponent value predictions, $\widehat{\mathbf{V}}_t^+ = \mathbf{w}_t^+ \cdot \mathbf{s}_t$ and $\widehat{\mathbf{V}}_t^- = \mathbf{w}_t^- \cdot \mathbf{s}_t$, whose difference is the complete value estimate: $\widehat{\mathbf{V}}_t = \widehat{\mathbf{V}}_t^+ - \widehat{\mathbf{V}}_t^-$. In the present case, such a dual representation would allow information about predictions derived from conditioned excitors and inhibitors to be separated between the two sets of weights. Such a richer representation allows the two sorts of information to be treated differently in terms of error signal computation; in particular, it would be possible to arrange for the positive error following presentation of a conditioned inhibitor to be carried in the serotonergic error signal.

This sort of representational flexibility offers a number of further interesting possibilities for behavioral modeling that have been particularly explored by Grossberg and collaborators. Many of these insights could be accommodated in an opponent TD model that separated appetitive and aversive values. One example is so-called *active extinction,* in which the extinction of an excitatory appetitive association is accomplished not by the usual method of reducing the positive weights in $\mathbf{w}^+$ that represent it, but instead by increasing corresponding negative weights in $\mathbf{w}^-$ to counteract it, so that after extinction the CS excites both channels in a balanced manner. This device has been used to explain various behavioral phenomena surrounding extinction such as the spontaneous recovery of previously extinguished associations (Grossberg, 1984), which on this scheme can occur when any of a number of events upsets the balance of excitation between the channels. Conditioned inhibition experiments also suggest that predictions are represented with separate appetitive and aversive associations; in particular there is evidence that conditioned inhibitors carry both positive and negative associations simultaneously. (The positive associations are thought to be due to second-order conditioning from B's repeated pairings with the excitor A.) Attempts to extinguish a conditioned inhibitor by presenting it alone can actually *strengthen* its inhibitory properties, presumably by

[5]There are anecdotal experimental reports (P. Dayan, personal communication, 2003) that if B is followed by a reward (an event which is particularly unexpected after a conditioned inhibitor), the dopaminergic response to that reward will be enhanced compared to a totally unsignaled reward. In the model, extra positive prediction error is expected in this case since $V(t+1) - V(t)$ is positive when the reward is delivered (reflecting that reward is usually omitted there), but this is also exactly why excitation is predicted when B is presented but unrewarded; it is unclear why these two situations should differ empirically.

extinguishing its positive associations (Williams and Overmier, 1988). This phenomenon would be difficult to capture in a standard TD model, but would pose no problem for an opponent version with separate appetitive and aversive weights.

3.5 Behavioral choice experiments

There is a large behavioral literature that may bear on the question of what return animals are learning to predict. To the extent that the dynamic programming idea behind the TD models is correct — i.e. that animals are learning to predict a return in order to select actions that *optimize* it — then animals' choices on decision tasks should help to reveal what return they are using. Such a view is at best incomplete (see Section 2.3.3 and the work of Dayan, 2002; Dayan and Balleine, 2002), in that animal behavior is influenced by many factors such as arousal by Pavlovian cues, which are extrinsic to an optimal-choice framework and might, conceivably, corrupt the results of this sort of optimal-choice analysis. In this section, I will set aside these important caveats in order to pursue more directly the question of whether the choices animals make on decision tasks can be reconciled with the idea that they are maximizing a return of a sort that could be learned with TD. In particular, I consider how the decision results reviewed in Section 2.3.4 could arise in a TD system. For this, I must address two basic challenges in the data. First, how can the form of animals' temporal discounting functions (which are evidently hyperbolic in contrast to the exponential functions used in RL algorithms) be reconciled with a TD approach? Second, how can animal sensitivity to variability in reward amounts and delays be reconciled with an algorithm that is supposed to maximize expected value? I will touch only inconclusively on the separate, normative question of *why* animals should favor any particular return. The latter question is the subject of intensive theoretical work (see Kacelnik and Bateson, 1996, for a review) but very much unresolved.

In the rest of this section, I first lay out some background about previous theoretical approaches to animal discounting in order to motivate my approach. Next, I introduce an exponentially discounted return similar to the predictor valuation model of Montague and Berns (2002; Montague and Baldwin, 2003), which can be learned through TD methods, and discuss empirical evidence regarding the settings for its parameters. Next, I demonstrate how the return fits animal choice data regarding discounting and risk sensitivity, and close with a discussion of several open points.

3.5.1 Theoretical approaches to temporal discounting

Section 2.3.4 reviews evidence from choice experiments that is commonly thought to demonstrate that animals discount the value of future rewards hyperbolically (Equation 2.16 on page 35) rather than exponentially (Equation 2.15) in their delays. The reason these data are interesting from the perspective of this thesis is that the hyperbolically discounted return that has primarily been used to model these data has no recursive Bellman formulation, and thus cannot be learned directly using TD-style methods. However, compared to the returns at issue in this thesis, the choice models of Equations 2.15 and 2.16 are oversimplified for two reasons, both of which strongly affect the interpretation of the data. First, as was pointed out in this context by Kacelnik (1997), and in accordance with the focus of this chapter, Equations 2.15 and 2.16 are single-trial returns, ignoring any effects of rewards expected in future trials. Second, the models fail to account for the well-studied phenomenon (Gibbon, 1977) of variability in animals' estimates of time intervals. This topic will be a major theme of the next chapter, and an important factor in the choice model I present here.

As reviewed at length in Section 2.3.4, Kacelnik (1997) explained hyperbolic discounting as resulting from animals maximizing their long-run average reward expectancy, but this model has a serious quantitative deficiency in that in order to fit the data, it must be assumed that animals ignore the intervals between trials when computing the reward rates available at each alternative. Kacelnik's (1997) choice model can be directly implemented using an actor/critic model incorporating the average reward TD algorithm presented in this chapter (Daw and Touretzky, 2000). However, the model's assumption that the inter-trial intervals do not affect the learned values seems particularly ad hoc when made explicit in the context of a TD implementation. For this reason, I develop another approach below.

Another normative justification for hyperbolic discounting is that it can emerge from exponential discounting in a single-trial return when there is uncertainty as to the discounting factor γ (see Kacelnik, 1997,

who attributes the idea to a 1992 personal communication from Y. Iwasa). That is, if we believe there is some interest rate (or loss probability) but do not know what it is, then the proper approach to computing the expected value of a future reinforcer is to choose some distribution over the discounting factors and take an expectation of Equation 2.15 over that distribution. If we assume γ is drawn from a uniform distribution on the range $[0, 1]$, then the expected discounted value is

$$\int_0^1 \gamma^d r \cdot d\gamma = r/(1+d) \tag{3.9}$$

— exactly hyperbolic discounting. Such values could be computed neurally by averaging over a family of exponentially discounted value functions learned using TD methods in parallel by different groups of neurons under different hypotheses about γ. This represents an indirect method of learning a hyperbolically discounted value function using TD methods; as already noted, because of the lack of a recursive form of Equation 2.16 on page 35, a direct TD method cannot be derived in the usual fashion.[6]

Here I pursue a different version of the same insight. The same story can be told in terms of uncertainty about the delay d rather than the discounting factor γ because there is a duality between these two variables. That is, with a change of variable, Equation 3.9 can be rewritten as an integral over the delay d, assuming a fixed γ. This manipulation would impose an exponential distribution over d with a rate parameter λ proportional to $1/d$. This would be a strange choice of distribution to represent an animal's uncertainty in d (we might at least expect a distribution that peaks near d). However, in practice, more reasonable distributions for uncertainty in d can be used to similar effect. Montague and Berns (2002) mention that this effect produces something like hyperbolic discounting in their predictor valuation model (Equation 3.2), in which d has a Gaussian-like but skewed distribution. While it is easy to show that such an account will give rise to qualitative features of animal decision-making, to my knowledge, this sort of theory has never been systematically compared to quantitative choice data such as that from Mazur (1987) and Kacelnik (1997), in order to determine if the data could be explained under reasonable assumptions about the discounting and temporal uncertainty factors involved. In the following sections I specify a version of this account in which temporal uncertainty enters due to noise in animals' timing processes (Gibbon, 1977), and then demonstrate that its behavior closely resembles quantitative animal choice data.

3.5.2 A model of choice

Here I specify a model of choice that exploits the fact — key to the next chapter — that animals' interval timing processes are noisy, and discuss experimental constraints on its parameters. Like the predictor valuation model of Montague and Berns (2002; Montague and Baldwin, 2003), this model assumes that animals choose between alternatives based on their expected exponentially discounted values, but that they must integrate out uncertainty about the timing of rewards. For this discussion, I will assume (counter to the general focus of this chapter) a parameter regime in which γ is small enough that rewards outside the current trial do not materially contribute to animals' valuations. (The intervals between trials in the choice experiments are in fact fairly long.) In this regime, I am justified in using a single-trial return, which is a good approximation to an infinite horizon return for small γ.

Animals famously suffer from substantial noise in their interval timing processes (Gibbon, 1977), and so a deterministic choice-reward delay will appear to the subject as variable. A TD model encountering this sort of variability will use samples of delays to compute the expected discounted value of the reward with respect to its noisy delay. There is an issue as to how the timing mechanisms used in TD models can incorporate noise that has the scalar form measured in behavioral experiments. I will address this in the following chapter; here, I determine choices by directly computing the values (estimating the integrals numerically) rather than learning them with a TD system.

Following Gibbon (1977), I assume that animals' measurements d' of some true interval d are distributed as a Gaussian whose mean is d and whose standard deviation scales as a constant fraction of the mean. The level of uncertainty in this model is controlled by a single parameter, the coefficient of variation c. Putting

[6] Another indirect approach to learning a finite-horizon hyperbolic return might be to sample the discount factor γ in effect during each learning episode.

it all together, the choice model holds that when animals must choose between a reward of magnitude r_1 after true delay d_1 versus a reward of magnitude r_2 after delay d_2,they choose the first reward if

$$\int_0^\infty N(d'_1 - d_1, c \cdot d_1)\gamma^{d'_1} r_1 dd'_1 > \int_0^\infty N(d'_2 - d_2, c \cdot d_2)\gamma^{d'_2} r_2 dd'_2 \tag{3.10}$$

and the second reward otherwise. Here $N(x - \mu, \sigma)$ denotes a Gaussian evaluated at x. The differences between this model and the predictor valuation model of Montague and Berns (2002) are that the delay variability is represented by a true Gaussian, rather than a skewed one, and that this distribution's standard deviation, rather than its variance, scales with the mean.

The same approach can be used to model experiments in which reward timing is subject to programmed variation. In this case, the expected discounted value is computed with respect to both the programmed variability and the measurement variability (the two are presumably indistinguishable from the perspective of the animal). Assume the delay for one alternative is chosen from n_1 possibilities $d_{1,1} \ldots d_{1,n_1}$ with associated probability $p_{1,n_1} \ldots p_{1,n_1}$; likewise for $d_{2,1} \ldots d_{2,n_2}$ and $p_{2,1} \ldots p_{2,n_2}$. Then the first alternative will be preferred if:

$$\sum_{i=1}^{n_1} p_{1,i} \int_0^\infty N(d'_{1,i} - d_{1,i}, c \cdot d_{1,i})\gamma^{d'_{1,i}} r_1 dd'_{1,i} > \sum_{i=1}^{n_2} p_{2,i} \int_0^\infty N(d'_{2,i} - d_{2,i}, c \cdot d_{2,i})\gamma^{d'_{2,i}} r_2 dd'_2 \tag{3.11}$$

The parameters γ and c of this model represent rather fundamental behavioral variables that can be constrained by experiments other than the choice experiments I am attempting to fit here. In order to set them, I first determine the ranges of parameters that have been reported from these other experiments, and then select the values from this range that provide the best fit to the choice data. Noise in animal timing can be estimated from the temporal distribution of timed animal behaviors in a variety of experimental paradigms. Gibbon (1977) reports that a wide range of behavioral data are consistent with a coefficient of variation c between 0.35 and 0.5. There have been attempts to estimate the discount factor γ from neural rather than choice data, by fitting neural recordings to TD models under various (and in all cases oversimplified) assumptions. Suri and Schultz (1999) estimated that the discounting factor was about 0.8/second (i.e. rewards lose 20% of their value for each second they are delayed), evidently by comparing the spike rates of dopamine neurons to rewards and to stimuli predicting rewards at different delays. Estimates of the same parameter have also been made by fitting exponentials to the ramping anticipatory firing seen in striatal or orbitofrontal cortical neurons, which are sometimes thought to represent value function estimates. These estimates range from around 0.5/second by Montague and Baldwin (2003) to 0.6-0.9/second by Suri and Schultz (2001). I should note that this entire range of estimated discounting factors seems vastly too small to be normatively justifiable, and that for the numbers reported here I have rescaled the (dimensionless) factors γ reported in the various original papers to derive commensurable measures of the amount of value lost per second delay.

To choose parameter values from these ranges, I simply note that the behavioral data are well fit by the single-trial hyperbolically discounted return, and I thus choose parameter values for which Equation 3.10 is as close as possible to the hyperbolic return. Asymptotically (for large ds), the valuation integrals in Equation 3.10 approach proportionality to r_1/d_1 and r_2/d_2, and thus approach hyperbolic discounting. This can be verified using an asymptotic series expansion of the error function that appears in the analytic integral solution. The integrals approach their asymptotic hyperbolic behavior faster as $-log(\gamma) \cdot c$ grows: that is, more quickly for larger timing variation factors c and smaller discounting factors γ. For the simulations reported here, I have accordingly chosen parameter values that best satisfy these extremes: $c = 0.5$ and $\gamma = 0.5$/second.

3.5.3 Results and discussion: discounting behavior

Mazur (1987) quantitatively tested discounting behavior in pigeons by studying choice between large and small rewards. The data (reproduced in Figure 3.16) reveal the animals' *indifference functions*, the pairs of delays for which subjects are equally likely to choose either reward, and thus presumably value them as equivalent. So, for instance, the subjects value a 20 second delay to a small reward as equivalent to a large

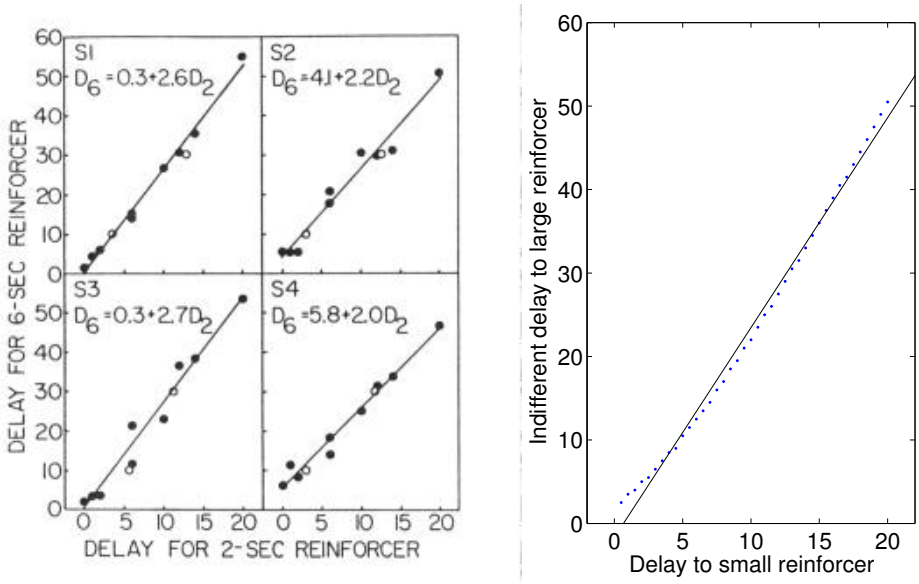

Figure 3.16: Empirical and modeled indifference functions in a discounting experiment. Points represent pairs of delays at which an animal (or algorithm) is indifferent between a reward of magnitude 2 and a reward of magnitude 6. Regression lines are superimposed. Left: Data for four pigeons, from Mazur (1987). Right: The indifference points predicted by the model of Equation 3.10. Indifference points are shown as dots, and the line of best fit is also shown.

reward after a delay of almost 60 seconds. If the delay to the larger reward is increased, the animal will tend to prefer the smaller reward, and vice versa.

Figure 3.16 shows that the decision model of Equation 3.10 produces choices similar to those measured by Mazur (1987). The figure was plotted by solving the integrals numerically in Matlab, and then finding indifference points to the nearest half second.[7] As noted before, for large ds, this model approximates hyperbolic discounting. For small ds the model instead behaves like exponential discounting. This progression can be observed in the slightly curvilinear appearance of the indifference function in Figure 3.16, right, and also causes a slightly negative y-intercept for the line of best fit. Such curvature is not obvious in the original data (Figure 3.16, left), and lines fit to those data have slightly positive intercepts. Notwithstanding these departures from aggregate features of the data, the individual indifference points predicted by the model lie within the ranges of those chosen by Mazur's subjects. Incidentally, some attempts have been made to detect curvature in pigeons' indifference functions (Mazur, 1986a, 1987), but without discovering anything appreciable. Curvature has been detected in human discounting experiments (Green et al., 1994), but using a methodology so dissimilar that it is outside the scope of the present theory.

3.5.4 Results and discussion: Risk sensitivity

Choice experiments testing animals' risk sensitivity raise a further potential obstacle to TD accounts of decision-making. Animals are often sensitive to variability in reward delays and amounts. Specifically, as reviewed in Section 2.3.4, they are almost always *risk-prone* for delays and tend to be *risk-averse* for amounts. In contrast, many value functions for reinforcement learning are risk-neutral (though the *learning algorithms* for estimating these values may not be; Bateson and Kacelnik, 1996; Niv et al., 2002). Here I show that detailed form of animal sensitivity to delay variability emerges from the choice model of Equation 3.11, and discuss a number of strategies for incorporating sensitivity to amount variability in a TD framework.

[7] Note that these integrals are not strictly expectations since the Gaussian "distributions" are truncated at zero delay and have not been renormalized. This has no consequence since, because of their scaling, the Gaussians for all delays are missing the same amount of area.

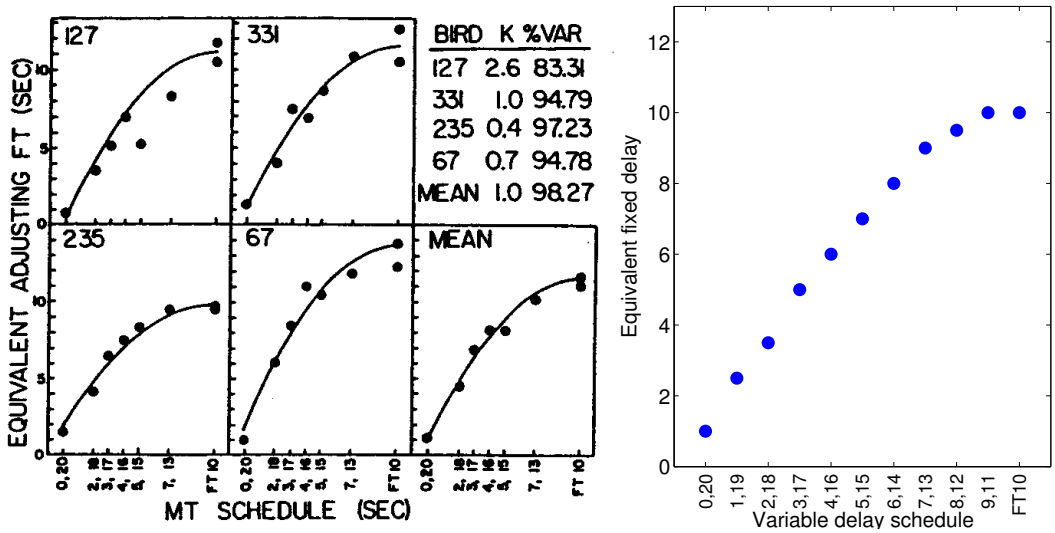

Figure 3.17: Fixed delays to reinforcement judged equivalent to reinforcers offered after a range of different variable delays. Left: Data from four pigeons and their mean, reproduced from Mazur (1984). Variable rewards were delivered after either of two delays (shown as pairs of numbers on x axis labels), with equal probabilities. Expected time to reward on the variable option was always 10 sec. The y axis shows the corresponding fixed delay to reinforcement at which the subjects were indifferent between the alternatives; the fact that these indifferent delays are less than ten seconds demonstrates that the animals were risk-prone. The point labeled "FT 10" is a fixed delay probe. Measured data (circles) are fit with curves derived from the EOR model of Equation 2.17 on page 36. Right: Data predicted by the model of Equation 3.11.

Mazur (1984) studied the specific form of pigeons' risk sensitivity for delays by offering them choices between rewards with fixed and variable delays. Figure 3.17, left, reproduces his results. For each condition, the variable delay was drawn with equal probability from a pair of delays (shown as pairs of numbers on the x axis), each with mean 10 seconds. For each pair, the fixed delay to the alternative was titrated to determine the indifference point. That the animals were risk-prone is demonstrated by the fact that the indifferent fixed delays are less than 10 seconds. The overall pattern of results is well characterized by the Expectation of Ratios (EOR) model of Equation 2.17 on page 36, which produced the fit curves.

How do TD value functions account for this data? Any discounted value function is risk-prone for delay, since by Jensen's inequality, the expected discounted reward when timing varies will be larger than the discounted expected reward. But this is only a qualitative claim. As shown in Figure, 3.17, right, the preferences of the choice model of Equation 3.11 closely resemble those of Mazur's subjects. This is essentially because, to whatever extent the integrals in Equation 3.10 approximate hyperbolic discounting, the sums over these integrals in Equation 3.11 approximate EOR.

Equation 3.11's predictions are within the range of the reported results, though the model's indifferent delays to the fixed reinforcer are consistently a bit lower than the mean over the subjects. This is evidently because the subjects' choices tended to be biased toward the fixed-delay alternative, boosting its value. This can be seen from the points marked "FT 10" in the graphs, which are probe conditions in which a fixed delay of 10 seconds was matched to another fixed delay by titration. The fact that the titrated delays tend to be longer than 10 seconds signals bias in the choices. (Note that this makes sense, since the titration procedure itself introduces delay variability on the adjusting alternative that I did not incorporate in reasoning about its fixed point.) In any case, the model closely matches the data from subject 235, whose choices showed no such bias.

I conclude by discussing risk sensitivity to reward amounts. In general, the data on this point are not nearly as quantitative or consistent as those for delay, and as discussed in Section 2.3.4, the measured effects are strongly driven by the subjects' energy budgets — an issue which is far beyond the scope of the present

modeling. That said, animals are most often risk-averse with respect to variability in reward amount.

The decision model of Equations 3.10 and 3.11 is risk-neutral with respect to reward amount. However, in TD models (as in windowed averages: see Section 2.3.4; Bateson and Kacelnik, 1996; Niv et al., 2002) the stochastic value estimation process itself produces risk aversion in amounts. A TD algorithm for the present choice model would thus produce risk-averse choices. On top of this, TD models can incorporate any arbitrary utility curve controlling how subjective values scale with reward amount, which is how Montague et al., 1995, explained risk aversion. The effect of energy budget manipulation might be captured by allowing the energy budget to change the utility curve, perhaps in conjunction with a model that incorporates effects of animals' motivational states on their behavior (Dayan, 2002; Dayan and Balleine, 2002).

3.6 Discussion and open issues

In this chapter we began with the general question of what return is the dopamine signal reporting, and this led to a series of richer models that address several other issues that wouldn't otherwise have seemed related. The simple change behind most of these results was the elimination of the artificial trial-boundary horizon on the value function. This raised the question of the effect of the long term predictions on the modeled dopamine signal. The average reward model was introduced to explore this; it indicates that long-timescale predictions would produce slow, tonic effects on the dopamine signal. Most of these experimental predictions remain untested, but several also match well with the otherwise confusing results of slow-timescale neurochemical recordings of dopamine levels in target structures. Particularly noteworthy is that the model would predict a slow-timescale excitatory dopamine response to aversive stimuli, which could explain microdialysis results that had previously been thought to challenge the TD theories. These considerations led to a more general exploration of the representation of aversive predictions and negative error, and to the (speculative) idea that the serotonergic system might be performing these tasks. On a more psychological level, such a model connects with a large body of theory on *opponency*: it provides a computational counterpart to the more phenomenological model of opponency between timescales of Solomon and Corbit (1974), and it unifies this explanation with a previously rather separate set of ideas about motivational opponency that underly classic theories about a variety of Pavlovian conditioning experiments, e.g. extinction, conditioned inhibition, and blocking between appetitively and aversively associated predictors. Finally we connected the theory to experimental and theoretical work from psychology and ethology about how to understand animals' choices as driven by an optimization process.

The most critical gap in the present theory is one of *timescale*. The model presented here has two timescales of response: a "phasic" timescale controlled by the size of the discrete timesteps in the Markov process, and a "tonic" one controlled by the speed σ at which the average reward estimates change. But surely this is an oversimplification: doubtless there are many timescales at which the dopaminergic signal might have important, and potentially different, response profiles and physiological effects. Moreover, the same general motivational response patterns described by Solomon and Corbit (1974) are observed at timescales ranging from seconds (classical conditioning experiments) to months (drug abuse and withdrawal), while my simulations of them are locked to the specific timescales of the model parameters. This problem, and others like it, is a major motivation for the modeling in the next chapter. There I will introduce semi-Markov models that are agnostic to timescale. There is a closely related question, which is how the learning rates for the predictions should be set. In general, this depends on one's beliefs about the speed of change in the underlying aspects of the world that the algorithm is trying to track (Kakade and Dayan, 2000, 2002a), but of course, there are presumably multiple timescales of change operating as well (Yu and Dayan, 2003).

A related issue of timescale is the tension between the physiological modeling in this chapter, which stressed the importance of long-term reward predictions to explain such phenomena as the dopaminergic response to aversion, and the behavioral choice data, which seem to be best explained in a TD setting by assuming a return that is (effectively, if not literally) truncated at a single trial. Exponential discounting in an infinite horizon return is in principle compatible with both approaches, depending on the steepness of discounting relative to the length of delays in the various tasks, but it is hard to determine whether there is a single timescale of discounting that could unite all the modeling in this chapter. In the sorts of physiological experiments envisioned in Section 3.3.2, the effect of strong exponential discounting would be to distort the predicted effects of average reward on the dopamine signal. In particular, rather than

being constant, the inhibitory effects of predicted future reward rate would be modulated by the temporal proximity of reward. As that modeling was fairly speculative and the data it connects with very qualitative, it is not really possible to estimate from data what sorts of discounting levels this model would accommodate. Meanwhile, in modeling the choice tasks, I assumed a single-trial return, but given a steep enough level of discounting and the long intervals between trials, an infinite horizon return would behave approximately the same, since rewards in future trials would contribute little to present valuations. However, it is unclear how the temporal uncertainty that model relied on would affect this assumption, since it has the effect of somewhat decreasing the effective discounting factor at larger delays, and so of somewhat increasing the extent to which future rewards might affect present choices. (In the extreme, if the approach perfectly replicated hyperbolic discounting, future rewards would contribute so much that the infinite horizon return would diverge, as it does with hyperbolic discounting.) Because there is no analytic solution to integrals of the form of those in Equation 3.10, it is difficult to determine whether extending the return to an infinite horizon one would have affected the model's behavior.

At any rate, to the extent that the tension between the two sets of models in this chapter is a real one, the most likely explanation is that my account of the choice tasks rests on far too naive a model of how the TD system controls behavior. As discussed in Sections 2.3.3 and 2.3.4, a more convincing TD account of the choice data could probably be based on a model incorporating several idiosyncratic influences on behavior, particularly conditioned reinforcement and its cousin, Pavlovian-instrumental transfer (Dayan, 2002; Dayan and Balleine, 2002). As discussed in the literature review, it is indeed likely that these processes have significant effects on the experimental results, which could cast doubt on the conclusions of the present, simple modeling. In particular, choices are sensitive to variations in the duration of the cues that bridge the gap between decisions and rewards, without changing the actual times or magnitudes of the rewards themselves (Mazur, 1997). It is tempting to conclude that choices in these experiments simply turn on subjects' undiscounted estimates of the reward rates co-occurring with each cue, which would straightforwardly explain hyperbolic discounting and the neglect of the intertrial interval. Such a rate estimate (while a poor basis for decision-making) should not require any computation of integrated future value expectancy and thus we might hope it would depend on some simpler neural system for learning correlations, bypassing the dopaminergic TD system altogether. But actually, dopamine is strongly implicated in conditioned reinforcement (e.g. Robbins et al., 1983), and pharmacological manipulations of dopamine seem to affect preferences on discounting tasks specifically through a cue-driven, conditioned reinforcement mechanism (Cardinal et al., 2000). Thus, under this account, we are left with two putative dopaminergic valuation systems with starkly different characteristics, a simple correlational one suggested by behavior and a more complex future value integration system suggested by physiology. Again, this puzzle would best be addressed in the context of TD models that explicitly incorporate phenomena such as conditioned reinforcement (Dayan, 2002; Dayan and Balleine, 2002). The issue is similar to an open puzzle in models of classical conditioning: animals clearly learn about the timing of USs, as shown by the timing of their CRs, yet data on how quickly they acquire a response seems to be best explained by models that assume responding is controlled by the co-occurrence rates of the CS and US, ignoring information on their relative timing (Gallistel and Gibbon, 2000; Kakade and Dayan, 2000, 2002a). (The key point is that CS duration would not affect acquisition speed at all in the coincidence detection model described on page 40 if animals were only trying to detect correlation between a timed event like the CS *offset* and the US.) I will return to this latter puzzle in the following chapter.

Chapter 4

Uncertainty, timescale, and state representation

Animals face two sorts of uncertainty that have have received only glancing treatment in the models I have discussed thus far. First, they may be uncertain about the relative timing of events — either because of internal noise in their timing processes, or because of external (e.g. experimenter-programmed) variability in event timing. Also, they may be uncertain as to the state of the world. That is, animals' immediate observations typically do not correspond to the states of a Markov process. It is rarely the case — as the Markov property requires — that at any instant, animals can immediately observe all of the information necessary to completely specify the probability distributions over future states and rewards.

Though these types of uncertainty are distinct, there is also a great deal of interplay between them, and I shall thus discuss them together. For one thing, standard dopamine models, including the ones discussed in the previous chapter, track the passage of time by introducing a series of marker states. On this account, uncertainty about timing can be viewed as a special case of state uncertainty. (This is not, in fact, how the standard models treat the phenomenon, since they contain no explicit notion of state uncertainty either.)

One example in which these issues play out together is in trace conditioning experiments, which will be a major focus of this chapter. Many of the experiments on dopamine neurons can be idealized (ignoring some instrumental contingencies) as trace conditioning. In a trace conditioning experiment, a punctate stimulus occurs, followed by a pause and then by reward delivery. Such an experiment is not Markov in immediate sensory observations, since nothing observable distinguishes the pause between stimulus and reward from any other empty interval that does not precede reward. While the animal's immediate sensory observations in this situation do not satisfy the Markov property, the task events can be perfectly described by a partially observable Markov process (also known as a hidden Markov model) — a Markov process whose state is not directly observable. If we assume the animal understands the task contingencies, and is capable of perfect timing, it can in principle infer the hidden process state with certainty from current and previous sensory observations. Say that we introduce variability in the interval between stimulus and reward, and also that we occasionally omit the reward altogether. The animal is now uncertain about the reward timing; moreover, the reward timing uncertainty induces state uncertainty. Imagine that there is a state, or a set of states, corresponding to the interval when the reward is pending, and a separate set of states for the interval between trials, when the reward is no longer pending. If the stimulus occurs and the reward has not arrived, the animal cannot know for sure, due to temporal uncertainty, whether the world is still in a state in which reward is pending, or the reward was omitted and the world has already transitioned into a state where reward is no longer available.

In this chapter, I will introduce separate but intertwined formal devices for dealing with each sort of uncertainty in TD models of dopamine and behavior. I address temporal uncertainty using a semi-Markov process, a formalism that incorporates explicit variability in the intervals between events. State uncertainty is formally handled by assuming that the semi-Markov process is also partially observable, that the animal's sensory observations are a potentially noisy or ambiguous function of the hidden semi-Markov state. The dopamine model I propose, based on a TD algorithm for this formalism, provides a more faithful account

than previous models of the behavior of dopamine neurons in situations involving temporal uncertainty. It also sheds light on many normative issues animals face in reasoning, learning, and predicting in the face of such uncertainty, issues that were often obscured in previous models.

This principled approach to uncertainty sheds light on several other issues surrounding dopamine system modeling — issues involving representation, timescale, and the functional organization of brain systems. The theory provides a formal framework for understanding the dopaminergic value prediction system as one of an interacting set of brain systems involved in learning and prediction. I address partial observability by positing a cortical representational system that learns a statistical *model* of the world's hidden processes and uses it together with sensory observations to *infer* the underlying value of the world's hidden state. The dopamine system is then responsible for learning, much as before, to map these inferred states to long-run predictions of future reward.

Here I focus on the value estimation and state inference portions of this framework — which are the pieces important for understanding asymptotic dopamine neuron responding of the sort usually studied in recording experiments — and leave for future work a fully worked-out theory of cortical model learning. But simply providing a formal specification for the problem of learning representations to support dopaminergic prediction is itself an advance. In a partially observable context, appropriate representations to support value prediction are task-dependent and must be *learned*, but previous dopamine system models use static representations and provide no normative guidance for how this learning should proceed. Here this problem is recast as that of learning a model of a partially observable process, which is well-specified and well-studied on the algorithmic level. This functional framework also suggests connections with neurophysiological theories that view representations for early sensory processing as acquired by statistical model learning (Lewicki, 2002; Lewicki and Olshausen, 1999) and with purely behavioral models of conditioning that view animal responses in conditioning experiments as also resulting from world modeling (Courville and Touretzky, 2001; Kakade and Dayan, 2002a). Finally, because value learning in a semi-Markov TD model is timescale invariant, the present framework is better suited than previous models for understanding several timescale invariance properties of animal learning and behavior (Gallistel and Gibbon, 2000).

The chapter is organized as follows. In Section 4.1, I review physiological and behavioral data relevant to issues of timing, timescale, and representation. In Section 4.2, I discuss how previous TD models of the dopamine system treated these issues, focusing on how the models might account for the experimental data. In Section 4.3, I introduce a new TD model of the dopamine system, based on partially observable semi-Markov processes, and also derive several reduced forms of the model under various limits that are useful in studying its behavior and in relating it to other TD models including the one studied in the previous chapter. The bulk of the chapter's results are presented in Section 4.4, where simulations of the model's behavior are compared to dopaminergic responses recorded in a range of tasks involving variability in event timing. Section 4.5 discusses how the model and the semi-Markov framework relate to data involving different sorts of timescale invariance properties in animal behavior. Finally, Section 4.6 discusses of a number of remaining issues. Some of the material in this chapter overlaps and expands on modeling originally presented by Daw et al. (2003).

4.1 Relevant data

In this section, I review a number of pieces of physiological and behavioral data relevant to issues of timing and temporal variability in dopamine system models. This will lay the groundwork for me to subsequently review in detail how previous dopamine models handled timing and representations of the passage of time, focusing on how they might cope with these data.

4.1.1 Physiological data

Several experiments have studied the behavior of dopamine neurons in situations when the timing of events varies. The results are summarized in Figures 4.1 and 4.2. Schultz et al. (1993) studied neuronal behavior in two tasks in which reward followed a series of two sequential stimuli, labeled "instruction" and "trigger" on the graphs. When the relative event timings were deterministic (separated by one second), dopamine neurons fired only to the first stimulus. However, when the timing of the second stimulus relative to the

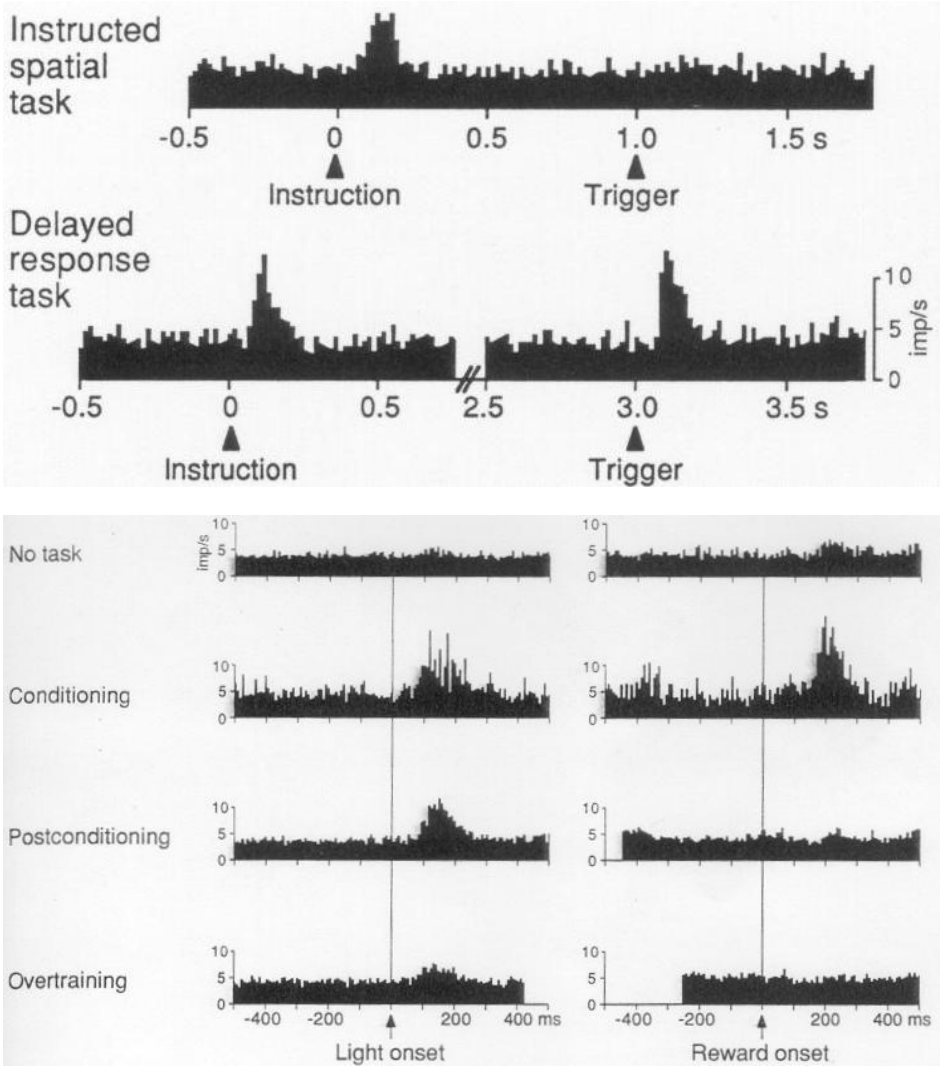

Figure 4.1: Data from two experiments studying whether dopamine neurons respond to events whose timing is variable. Top: Data from Schultz et al. (1993), comparing dopamine population responses in two tasks in which two successive stimuli ("instruction" and "trigger") predict reward. Dopamine neurons respond to the later stimulus only in the case where its timing with respect to the earlier stimulus is unpredictable. Bottom: Data from Ljungberg et al. (1992), pooled over a population of recorded neurons, demonstrating that dopamine neurons in overtrained monkeys cease responding to a stimulus that is predictive of reward. In this case, stimulus onsets occurred every 6-7 seconds, though the trial-to-trial variability is unclear. In the trace marked "no task" animals were given juice drops every 2.5-3.5 seconds; again, it is unclear how this variability was scheduled over trials. The neurons do not respond to the rewards.

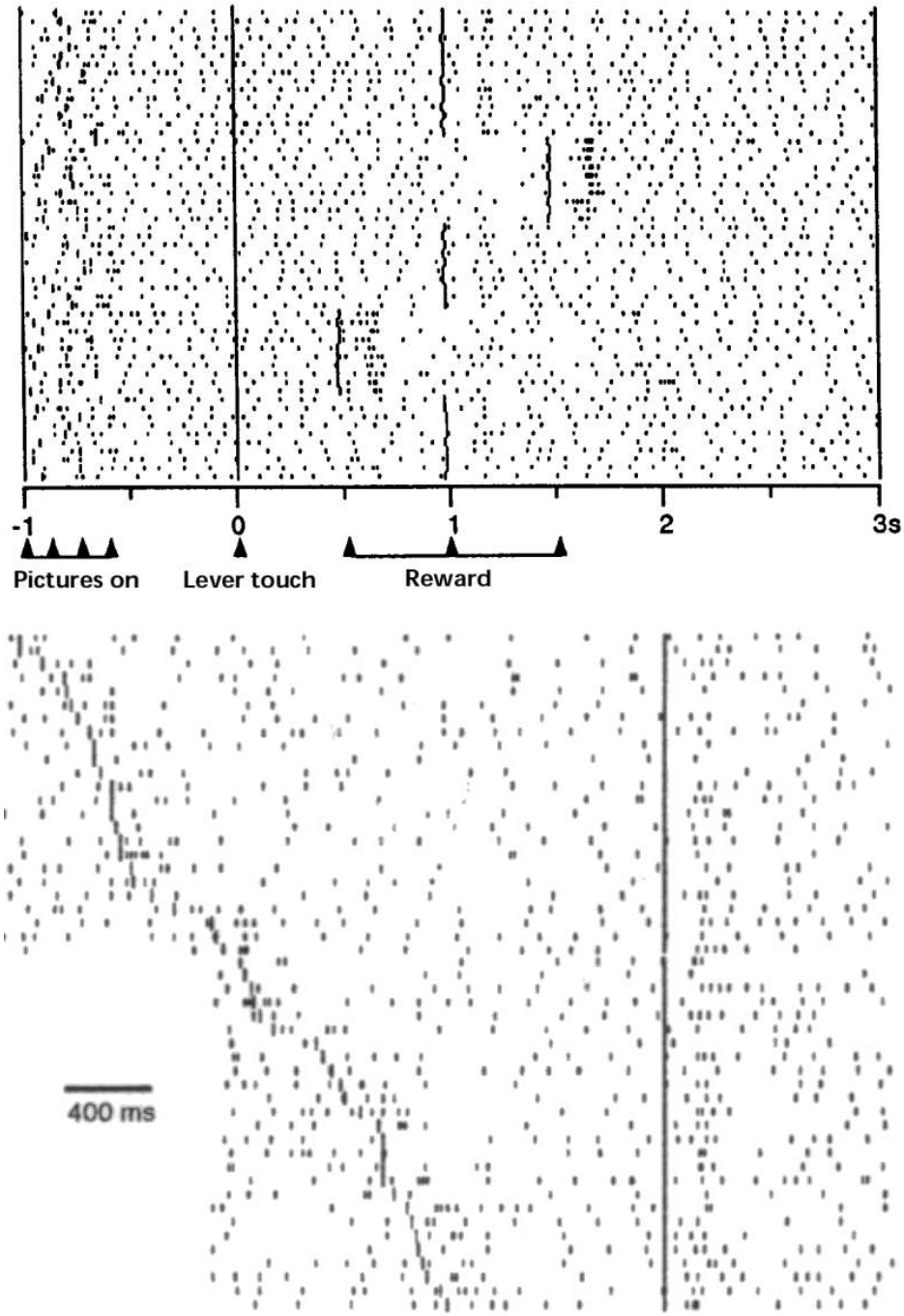

Figure 4.2: Data from two experiments studying the behavior of dopamine neurons when event timing can vary. Top: Data from Hollerman and Schultz (1998), showing a dopamine neuron recorded when an animal expects reward one second after a lever touch. Responding to the reward is seen in probe trials when it is delivered a half-second early or late; following early reward, there is no subsequent inhibition at the time reward had been expected. Bottom: Unpublished data from Fiorillo and Schultz (2001), courtesy of C. Fiorillo, showing the response of a dopamine neuron when the cue-reward interval varies uniformly over a two-second range. Rasters are sorted by cue-reward delay. More firing is seen to the cue after smaller delays.

first was varied randomly between 2.5 and 3.5 seconds, neurons responded to both cues (Figure 4.1, top). This result suggests that dopamine neurons respond to reward-related events only when their occurrence or precise timing is unpredictable on the basis of previous events.

Such responding occurs when the variability in event timing exceeds some threshhold (200-500ms; personal communication, Wolfram Schultz, 2003). Dopamine neurons can tolerate lower-level jitter in event timing; that is, dopamine neurons do not respond to reward-related events whose occurrence is predicted but whose timing has such jitter. This is to be expected because animals' interval estimation processes are demonstrated behaviorally to have substantial noise, so the system must cope with at least this source of variability even in situations (like Figure 2.2 on page 16) where the programmed intervals are constant and deterministic. However, the parameters of neurons' variability tolerance have not been systematically explored in experiments using experimenter-introduced variability (tolerance is demonstrated to variability around 100 ms; personal communication, Wolfram Schultz, 2003). Two examples of variability tolerance may be visible in the results of an early experiment of Ljungberg et al. (1992), though the paper predates the modern understanding of what factors are important to dopamine firing and the experimental methods described are correspondingly vague on the key points (and the experimenters were unable to recall the details in a recent personal communication). The data are reproduced in Figure 4.1, bottom. Shown are stimulus and reward responses pooled over populations of dopamine neurons recorded during different phases of monkeys being trained on a cued reward task. In the traces marked "no task," dopamine neurons are shown not to respond to uncued rewards delivered rapidly in sequence, and in the traces marked "overtraining," dopamine neurons are additionally shown not to respond to a stimulus that predicts a later reward. In the "no task" case rewards are described as being delivered "regularly timed" but also "every 2.5–3.5 seconds" (so it is unclear whether the variability is between-trial or between-session); similarly, the stimuli in the "overtraining" case occur every 6–7 seconds but there is no indication as to the source or trial-to-trial distribution of this variability, or whether the stimulus' exact timing might have been predictable exactly on the basis of some prior event.

A related finding, about which no data have been published, is that in conditioning experiments in which the deterministic stimulus-reward delay exceeds some animal-dependent threshhold (in the range of 3–5 seconds), dopaminergic responses to reward never completely disappear despite excessive training (C. Fiorillo, personal communication, 2002). We can understand this as another case of the responding when variability exceeds some threshhold by recalling that the variability of animals' timing noise scales with the interval being timed. Thus, in terms of the animal's noisy measurements, longer delays also appear to be more variable.

Taken together, all of these data show that whether dopamine neurons respond to events depends on how predictable is their timing relative to other events. Relative timing does not need to be *perfectly* deterministic for responding to disappear, but variability in relative timing beyond some threshhold produces responding.

Two further experiments have probed dopamine responding to time-varying events. Hollerman and Schultz (1998) trained animals to expect reward one second after a cued leverpress, and studied the responding of neurons in occasional probe trials when rewards were delivered a half second early or late. Their data are reproduced in Figure 4.2, top. When a reward is delayed (as when it is omitted altogether, see Figure 2.2 on page 16), dopamine neurons show a pause in their background firing at the time reward should have occurred. When the late reward finally arrives, dopamine neurons show a burst of activity. Early rewards do not show quite the opposite pattern. An early reward causes a dopaminergic burst, but there is no corresponding pause at the time the reward would have been received.

Fiorillo and Schultz (2001) conducted another conditioning experiment in which cue-reward delays were varied. In this case, cue onset was followed 1-3 seconds later (chosen uniformly) by reward. Note the contrast to the Hollerman and Schultz (1998) experiment, in which reward timing was normally deterministic and varied only in occasional probe trials. Here, reward timing is randomized throughout the experiment. The general finding is that in this case, dopamine neurons are on average excited by the reward, with more excitation for earlier rewards. Responses for a single neuron demonstrating this pattern are reproduced in Figure 4.2, bottom. The spike rasters are sorted by delay to reward. In these traces, a graded excitatory response to reward is seen after shorter cue-reward intervals, and the neuron also seems somewhat inhibited after the excitatory phase of the response. (A burst-pause responding pattern of this sort is well demonstrated to be common in some situations, e.g. by Schultz and Romo, 1990, and has been reported more sporadically

in others.) At longer delays, the burst portion of the response wanes to around baseline but the inhibitory phase seems slightly more persistent. Fiorillo and Schultz (2001) also show group averages of 20 neurons, with mean reward responses over trials and neurons, grouped in ten temporal bins ranging from early reward delivery to late. For these plots, spike counts were evidently taken during only the early, burst portion of the response. The mean response was excitatory in all cases, more so for earlier rewards, and waning to near baseline for later rewards. These data will be important to the model in this chapter, which makes specific predictions about the dopaminergic response to events covarying with the delay preceding them.

I should note here that the Fiorillo and Schultz (2001) task was delay conditioning (i.e. the cue persisted during the delay before reward delivery), while the other experiments described here were all trace conditioning. The distinction is not particularly meaningful from the perspective of timing, since the only useful event for predicting the reward timing is the cue onset. (Later in this chapter I will discuss one case in which the two situations may differ fundamentally.)

4.1.2 Behavioral data

Here I quickly recap several distinct phenomena thought to reflect timescale invariance properties in animal learning and behavior (Gallistel and Gibbon, 2000). A full discussion of these issues occurs in Section 2.3.5.

As demonstrated in Figure 2.6 on page 38 and studied extensively by Gibbon (1977), the profile of timed animal behaviors on tasks such as the peak procedure is proportional to the task's scheduled delay. In the figure, lever press rates are shown for fixed interval responding on an operant schedule, with the interval between reinforcement set at 30, 300, and 3,000 seconds. Plotted as a function of the inter-reinforcement interval, the response rates superimpose.

Similar scalar superimposition is seen for *distributions* of timed responses over trials — for instance, the distribution over trials of the latency to peak responding in the peak procedure. (Recall that, in the peak procedure, animals are trained to expect that an operant response will be rewarded at some fixed delay after a cue, which is varied systematically. The profile of responding is studied in unrewarded probe trials across these different delay conditions.) Two specific results follow from this general property. First, the *means* of these distributions scale with the delay being timed — e.g. if the mean latency to peak responding is 9 seconds when reinforcement is expected at 10 seconds, then the mean peak latency for a 100-second reward delay will be 90 seconds. Second, the variability in animal responding also scales with the delay; specifically, the *standard deviations* of the response distributions are proportional to the schedule delay. Thus, equivalently, the standard deviations of timed animal responses are proportional to their means, so that the response distributions have a constant coefficient of variation. I refer to this property as *scalar timing noise*.

These are reflections of timescale invariance in *behavior*. There are also data that suggest that animal *learning* processes reflect a timescale invariance property. Specifically, the number of trials an animal takes to acquire a response in classical conditioning is *invariant* to contractions or dilations of the timescale of events; this number is roughly proportional to the ratio of the inter-stimulus (CS-US) interval to the inter-trial (US-CS) interval (Gallistel and Gibbon, 2000; recall that CS, for conditioned stimulus, is the reward-predicting cue, and the unconditioned stimulus or US is the reward itself). While both intervals vary during timescale contractions or dilations, their ratio is constant; varying one interval while holding the other constant can wildly affect trials to acquisition. This property bears some resemblance to the timescale invariance properties in timed behaviors discussed above, but it is a distinct phenomenon. I will refer to it as *timescale invariant acquisition*.

4.2 Previous models

In this section I will review how three previous TD models of the dopamine system handled issues of representation and timing, focusing particularly on the data discussed above. Because the models fail to capture most of these data, I will try to diagnose the problems and speculate about what changes would be required to improve the accounts.

The key issue here is how models represent the time elapsed since some past event, which is necessary for predicting reward timing in both trace and delay conditioning. (In trace conditioning experiments, there is a

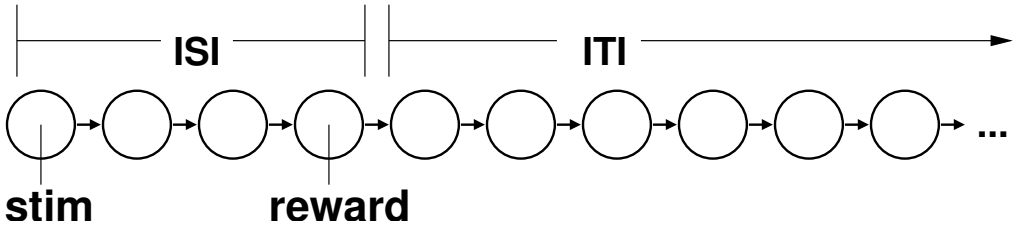

Figure 4.3: The state space for a tapped delay line model of a trace conditioning experiment. The stimulus initiates a cascade of states that mark time relative to it; the reward falls in one such state. ISI: inter-stimulus interval; ITI: inter-trial interval.

gap between the CS and US; in delay conditioning the CS persists until reward delivery.) Dopamine neurons clearly have access to such a representation, since they show a well-timed pause when reward is omitted. In the case of trace conditioning with TD(0), a representation of past events is also necessary to allow value to accrue to the CS over the temporal gap between CS and US. A secondary and separable issue about these experiments is how the algorithms *make use* of the temporal representation in constructing a prediction of future reward.

In evaluating the models' temporal representations, I will assume that these representations underlie not just value prediction, but timed behaviors in general, in order to compare their noise and scaling properties against behaviors of the sort studied by Gibbon (1977), even if these behaviors might not directly result from reward expectancy or from TD learning per se. For the purpose of evaluating the models' relationship to the Gallistel and Gibbon (2000) result on timescale invariant acquisition in classical conditioning, I will assume that response acquisition in these tasks corresponds to value learning in the models. That is, I assume that a conditioned response occurs to some stimulus if the value $\widehat{\mathbf{V}}$ predicted by that stimulus exceeds some threshhold level, and that the timing of responses is controlled by whatever mechanism each model uses to track the passage of time. An important caveat is that Gallistel and Gibbon (2000) base their conclusions about the rate of acquisition on delay conditioning experiments. It is not certain whether their results would generalize to the case of trace conditioning, which has been the primary focus of dopaminergic models. Here I suppress this distinction, assuming provisionally either that timescale invariance applies to trace conditioning as well, or at least that the trace/delay distinction is lost at the level of representational abstraction used by the models reviewed here. In fact, the dopamine system models I'm reviewing do not explicitly distinguish between trace and delay conditioning, and they make no obvious provision for representing temporally extended stimuli. And while experiments on dopamine neurons by Schultz's group have used both delay and trace conditioning (with comparable results), models invariably treat them as trace conditioning.

4.2.1 The model of Houk et al. (1995)

The model of Houk et al. (1995), which was rather abstract, had only vague notions about state representation and contained no timing mechanism whatsoever. Only immediate sensory observations were represented (as patterns of cortical activity); however, when stimuli disappeared their representation persisted as decaying eligibility traces (see Section 2.1.5), which allowed discounted value to accrue over the trace delay.

Because it lacks a timing mechanism, the model is silent on essentially all of the temporal issues considered here. It does, however, violate timescale invariance in acquisition because eligibility traces decay at some fixed timescale; longer trace intervals will result in slower value learning.

4.2.2 The model of Montague et al. (1996; Schultz et al., 1997): overview

Montague et al. (1996; Schultz et al., 1997) used a tapped-delay line timing mechanism to time out the trace delay. (The same timing mechanism was used in the model of Chapter 3.) This device, illustrated for the example of a trace conditioning experiment in Figure 4.3, introduces a cascade of marker states triggered by the CS, through which the system progresses one per Markov timestep following the stimulus presentation.

Value from the reward backs up to the CS state by way of the intermediate states. Apart from discretization error, the timing is noiseless.

4.2.3 The model of Montague et al. (1996; Schultz et al., 1997): physiology

I first consider how this model deals with the dopaminergic response data from Figures 4.1 and 4.2. It fails on everything apart from the first experiment.

As Montague et al. (1996) show in their original paper, this device correctly handles the Schultz et al. (1993) result (Figure 4.1, top) that a reward-predicting stimulus excites dopamine neurons only when its timing relative to a previous stimulus is variable. The reason for this is easier to see in the context of a simpler hypothetical experiment involving only a single stimulus followed by reward. If the stimulus-reward timing is deterministic, then the reward invariably falls in the same state in the stimulus' delay line, and the asymptotic TD error to reward (which is related to the probability of reward in any particular state) is zero. When stimulus-reward timing varies, the reward occurs in different delay line states in different trials, and its occurrence in any particular state is thus only partially expected. Positive asymptotic TD error results to the reward. Analogously, in the model of the Schultz et al. (1993) result, the trigger stimulus does not evoke prediction error if it (and the reward it signals) always fall in the same delay line states, but if the trigger stimulus occurs in different states in different trials, it evokes an increase in predicted value that causes positive prediction error.

For the same reason, however, the model has difficulty explaining why rewards and reward-predicting events can fail to evoke dopaminergic responding even when their timing is at least somewhat variable (Figure 4.1, bottom, may be an example of this phenomenon). In the tapped delay line model, if a reward (or a reward-predicting stimulus) occurs in a state in which it is not 100% expected, positive TD error to that event results asymptotically. The only obvious hope for explaining the lack of response in the model is to assume that the Markov timestep is so coarse (i.e. the tapped delay line is progressing from state to state so slowly) that events that vary in continuous real time nonetheless always fall in the same discretized state. The maximum duration of the Markov timestep should be constrained by the durations and latencies of phasic dopaminergic events — which are in the range of 100 ms. This might be barely enough to explain the rather sparse data on dopamine neurons' tolerance of programmed variability (which is also demonstrated in the range of 100 ms; Wolfram Schultz, personal communication, 2003), but is insufficient to cope with additional variability in animals' time measurement processes (which has a standard deviation of 300–500ms for a measurement of one second; Gibbon, 1977).

Assuming the time representations that drive dopamine responding show behaviorally plausible timing variability, even fully predicted, deterministically timed rewards would result in asymptotic TD error — so long as the stimulus-reward delay is long enough that the timing noise would cause rewards to fall in different states from trial to trial. Given plausible levels of noise, determined behaviorally, and of state discretization (discussed above), the maximum delay should be in the range of a few hundred milliseconds. But in experiments, the dopamine response to predicted rewards is routinely trained entirely away even if the stimulus-reward delay is 1–2 seconds. As noted in Section 4.1, for delays longer than 3–5 seconds, the reward response cannot be trained away. I had previously suggested that noisy timing could account for this result, but in the tapped delay line model it clearly cannot. Rather, the model's best chance for explaining these data would seem to be to assume unrealistically noiseless timing and a limit of a couple of seconds on the maximum duration that can be timed at all (i.e. a limit to the number of states in the delay line). If the delay lines are assumed to control timed behaviors as well as dopamine neuron responses, this is an unreasonable assumption, since animals can display timed behaviors at intervals of minutes or even hours (e.g. Gibbon, 1977).

The model also mispredicts the result of the Hollerman and Schultz (1998) experiment, in which a reward was delivered earlier or later than expected. The model rightly predicts that if reward does not occur in a state in which it is expected (because the reward is late or omitted), then negative TD error will result, corresponding to a timed pause in dopaminergic background firing. However, the model predicts that this same pause will occur even if the reward had already been delivered early, which was not observed by Hollerman and Schultz (see Figure 4.2, top).

At the simplest level, this misprediction might seem to occur because reward delivery is not registered

in the model's representational system at all (e.g. reward does not have its own tapped delay line), so the early reward has no effect on subsequent predictions. But simply adding a second tapped delay line to represent reward delivery does not solve the problem, for reasons both deep and shallow. The shallow reason has to do not with the stimulus representation but with the way the algorithm makes use of it in constructing a value estimate. Specifically, the model assumes that multiple active delay lines sum linearly to produce an aggregate value estimate. Under this assumption, it is not possible to assign weights to the stimulus and reward delay line states in such a way as to mimic the observed dopaminergic behavior in both the cases when the reward occurs early and at its normal time: the necessary value functions for these two situations are not linear in the state representations. (This is analogous to the famous XOR problem with linear perceptrons.) The deeper issue is that even if the model could learn the proper value functions — for instance, even if a nonlinear function approximator were used — it would not have the opportunity to do so. In the experiment, the early rewards were tested only in occasional probe trials — the value functions were presumably determined entirely by experience with the reward occurring at its normal time. Thus the experiment does not seem to probe the system's ability to *learn* that an early reward should cancel a later reward expectation, but instead to probe the default behavior of the system experiencing such an event more or less for the first time.

Finally, and for a similar reason, the model fails to capture the finding of Fiorillo and Schultz (2001) that when cue-reward timing is varied uniformly over a range, more dopaminergic excitation is seen to early rewards than to late ones (see Figure 4.2, bottom). In this case, because of the uniform distribution of reward timing, the probability of reward occurring in any delay line state is equal, and so the dopaminergic response in any of these situations is also equal. Again, the problem would seem to have something to do with the fact that early reward occurrence does not do anything to change the model's expectations about subsequent reward.

4.2.4 The model of Montague et al. (1996; Schultz et al., 1997): behavior

Here I discuss how the Montague et al. (1996; Schultz et al., 1997) model fares in light of the behavioral results on various sorts of timescale invariance discussed in Section 4.1.2. The bulk of this section concerns reviewing various possibilities for incorporating scalar timing noise into the tapped delay line device.

One important regularity is *timescale invariant acquisition*: the number of trials it takes animals to acquire a conditioned response is invariant to contractions or dilations in the speed of events (Gallistel and Gibbon, 2000). TD(0) value learning with the tapped-delay line representation manifestly violates this property. Halving the speed of events, for example, will double the number of marker states necessary to time the interval, slowing the process by which value backs up to the CS.

The other relevant results about timescale invariance concern the scaling properties of animals' timing noise. Recall that the standard deviations of animals' timed behaviors scale linearly in their means, a property I refer to as scalar noise (Gibbon, 1977). The model of Montague et al. (1996; Schultz et al., 1997) incorporates no timing variability (apart from discretization noise, which does not scale), so an appropriate question is whether the model could be retrofitted with scalar variability. Here I consider the tapped-delay line representation apart from its involvement in a TD model of dopamine, putting aside the fact (mentioned above) that timing noise would disrupt the broader model's account for dopaminergic behaviors.

The tapped delay line mechanism closely resembles the pacemaker/accumulator timers analyzed in Section 2.3.5; in both, timing is accomplished by counting periodic, discrete events (state transitions, in the present case). As we know from the analysis of pacemaker/accumulator timers, the simplest proposal for introducing timer noise — independent, identically distributed (i.i.d.) variability in the intervals between state transitions — will not in itself produce variability with the proper scalar form, due to the central limit theorem. In Section 2.3.5, I argued that attempts to correct this shortcoming by introducing correlations to the pacemaker noise are frustrated by the fact that explaining different experiments — or even different parts of the same experiment — requires assuming different timescales of noise correlation, i.e. that timer noise is correlated within, but not between, trials. Moreover, since the present model has no separate concept of storage or retrieval of observed intervals (instead, interval expectations are implicit in the pattern of weights that represent the value function), it cannot accommodate the assumption from Scalar Expectancy Theory that timing variability is dominated by scalar noise introduced in those stages (Gibbon, 1992; Gibbon and

Church, 1984). And because the model contains no concept of uncertainty in any of its learned parameters, it cannot incorporate yet another proposed explanation for scalar timing variability: asymptotic uncertainty in an animal's estimate of a parameter in the world (in this case, a temporal interval) owing to an assumption that the underlying parameter may be changing from measurement to measurement (Kakade and Dayan, 2000, 2002a).

The best remaining possibility for a version of tapped delay lines with scalar variability seems to be introducing noise at the representational level: in pacemaker/accumulator terms, it can be introduced in the "accumulator," which is the delay lines themselves. That is, since the position of the active element in a delay line represents a count of how many pacemaker ticks have occurred since the stimulus became active, all that is needed is to arrange that these counts degrade as they become larger (cf. the spectral timing model of Grossberg and Schmajuk, 1989). One simple way to do this is, whenever a delay line becomes active, to randomly choose some scale factor near unity, and multiply the accumulating count by that amount. (That is, multiplicatively accelerate or decelerate the progression of the active delay line element. This is not the same thing as *smearing out* the delay line representation by activating multiple elements.) This is really a variation on the proposal for correlated pacemaker noise (e.g. Gibbon, 1992) that I criticized above. The difference is that because the scale factors are introduced locally (in individual delay lines) rather than globally (in the pacemaker), there is no problem choosing an appropriate timescale for their action. Each time a stimulus recurs, its delay line reactivates and there is an opportunity to take new samples of the intervals between the stimulus and subsequent events. Thus, the speed of delay line reactivation can control the speed with which its scale factors are refreshed.

Thus, although the original model did not include timing noise, scalar timing noise can be added at the representational level. But there does not seem to be any way of addressing the learning algorithm's violation of another timescale invariance property: timescale invariant acquisition in classical conditioning.

4.2.5 The model of Suri and Schultz (1998, 1999): overview

Suri and Schultz (1998, 1999) introduced a dopamine system model that resembles that of Montague et al. (1996; Schultz et al., 1997) in many respects, but uses a rather different temporal representation. Another important disadvantage of the tapped delay line mechanism is that the proliferation of marker states is an inefficient representation of the (fairly simple) true value structure of a task like trace conditioning; since the model of Montague et al. (1996; Schultz et al., 1997) is undiscounted, the states between stimulus and reward all have exactly the same value but must be learned independently. This can be seen more clearly by considering another view of the model's state representation: how it builds the value function by adding together "basis functions" that produce different pieces of it (Figure 4.4, left). The temporally extended value function is built up as the sum of many little basis functions that each persist for only one timestep, one corresponding to each marker state from Figure 4.3. A more parsimonious alternative is to use a single temporally extended basis function, as in the model of Suri and Schultz (1998, 1999) (Figure 4.4, right). In this model, stimulus onset immediately activates a large number of representational elements, each of which stays active for a different interval, ramping up exponentially and then shutting off. As the model uses exponentially discounted TD, each of these basis functions exactly corresponds to the full, temporally extended value associated with a single reward expected at some delay. The learning problem simply comes down to selecting the correct basis function and giving it the correct weighting. The authors use a variation on the usual delta-rule weight update (Equation 3.4 on page 45) that selects the correct basis function by making only stimulus elements that have just shut off eligible for learning.

4.2.6 The model of Suri and Schultz (1998, 1999): physiology

Apart from the use of temporally extended basis functions, the Suri and Schultz (1998, 1999) model closely resembles that of Montague et al. (1996; Schultz et al., 1997). It would make precisely the same (mostly wrong) predictions about the experiments from Figures 4.1 and 4.2, except that the model includes a rather simple fix intended to correct some of them. However, I will argue that this approach is not a general solution to the underlying problems, and as a result it does not plausibly generalize to other situations.

Like the Montague et al. (1996) model, this one predicts that dopamine neurons will respond asymptotically to any reward or reward-predicting stimulus whose timing varies, and not to any whose timing

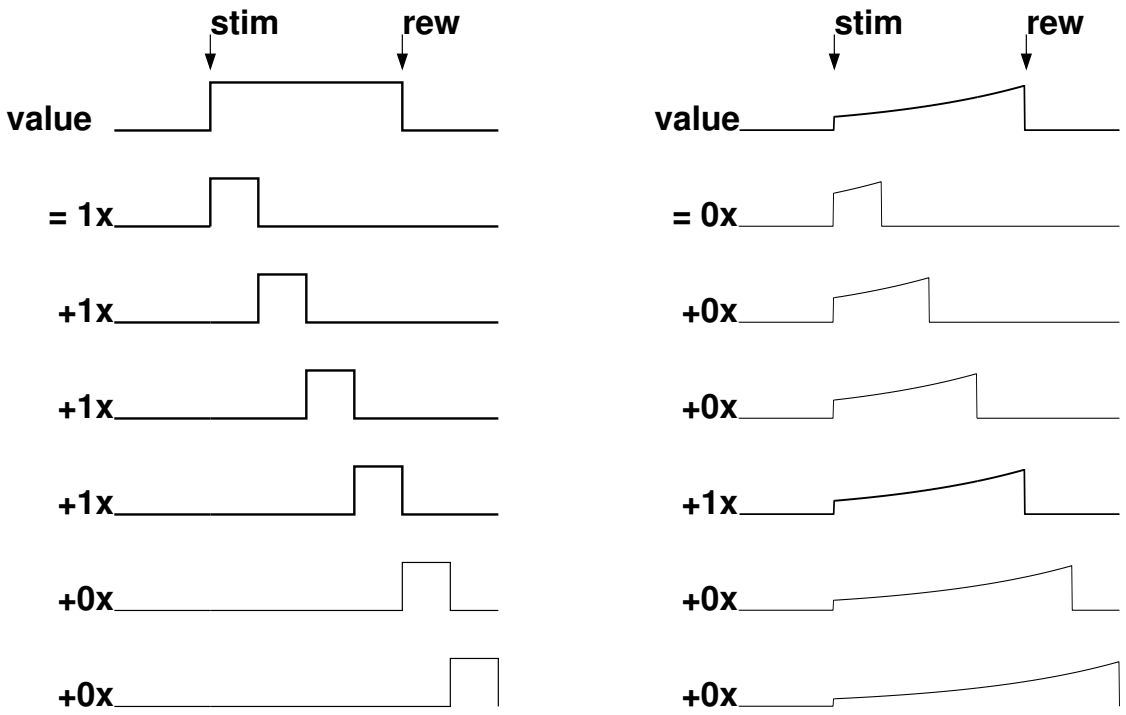

Figure 4.4: Schematic comparing how the models of Montague et al. (1996; Schultz et al., 1997) (left) and Suri and Schultz (1998; 1999) (right) construct the desired value function as a weighted sum of basis functions.

does not.[1] Thus it correctly reflects the results of Schultz et al. (1993) (Figure 4.1, top) but not those of Ljungberg et al. (1992) (Figure 4.1, bottom). The discussion of the effects of timing noise on the previous model's responses is equally applicable here.

The difference between the models comes out in their accounts of the experiment of Hollerman and Schultz (1998), for which Suri and Schultz (1998, 1999) included a targeted fix to correct the prediction of spurious pause after reward delivery. The basic notion is that reward delivery resets the representational system, clearing all pending predictions and thus avoiding negative TD error when they fail to play out. (The same trick is also used in the dopamine model of Brown et al., 1999.) In the Suri and Schultz (1998, 1999) model, the prediction cancellation is actually implemented by having a second stimulus cascade for reward occurrence, which, instead of combining linearly with the stimulus' predictions, overrides them. Thus the stimulus representation is, strictly speaking, still active but it has no effect on value prediction.

This simple device also corrects the previous model's account of the Fiorillo and Schultz (2001) result (Figure 4.2, bottom). Because reward occurrence resets the stimulus representation, the basis functions for longer stimulus-reward intervals are subject to learning only on trials when the reward hasn't already occurred. The result of this is that the TD error to a reward occurring at some delay after the stimulus varies with the probability of reward at that delay *conditional* on the reward having not already occurred. As time passes unrewarded, the conditional probability of reward at any particular instant increases, and the TD error to reward correspondingly decreases. Thus the model accords with the experimental finding of greater dopaminergic excitation for shorter delays.

While this simple device brings the model into line with these specific experiments, it does not generalize properly to other situations. For instance, we know from behavioral experiments that animals can learn that a single stimulus predicts two or more sequential rewards, for instance in a paradigm called downwards unblocking. But (assuming that the dopamine system underlies this behavior, or at least that it has equivalent

[1]Strictly speaking, both this model and Montague et al.'s respond to a CS that begins a trial no matter what its timing, since they do not carry predictions over trial boundaries. But this is easily corrected as in the previous chapter.

predictive capabilities), this ability is violated by the Suri and Schultz (1998, 1999) model, since the stimulus representation is rendered ineffective after the first reward. It's also not possible to "chain" the prediction of a second reward off the representation of the first, since the two rewards are indistinguishable in the model, so the second reward would improperly predict a third reward. The general point of this example is that an appropriate task representation is task-dependent and must be *learned*. Which past events are relevant to the present value predictions, and for how long, varies from task to task. Montague et al. (1996; Schultz et al., 1997) fail to predict the results of several experiments because they assume that reward delivery is not relevant to subsequent value prediction; Suri and Schultz (1998, 1999) correct this particular problem by assuming that reward delivery *is* relevant, but making the equally dubious and inflexible assumption that, after the reward occurs, no prior stimuli are predictively relevant. In the case of trace conditioning, the stimulus is predictively useless once its associated reward is received, but in downwards unblocking, it maintains predictive importance even after the first reward is received. This line of reasoning rules out simple but inflexible reset rules like the one used by Suri and Schultz (1998, 1999). The model I present later in this chapter offers a similar explanation for dopamine behavior under early reward delivery, but it uses a more flexible and general representation in which the extent to which a stimulus remains predictively useful after intervening time or events (the "reset policy," as it were) can in principle be learned.

Thus the model of Suri and Schultz (1998, 1999) does not address one of the problems with its predecessor (the oversensitivity in the error signal to stimuli whose timing is not perfectly deterministic), and it offers only a partial fix to the other problem, of conditioning value estimates on the appropriate set of past events.

4.2.7 The model of Suri and Schultz (1998, 1999): behavior

Here I discuss the issues of timescale invariance in the model of Suri and Schultz (1998, 1999).

The main difference between this model and its predecessor has to do with timescale invariant acquisition. It is actually fairly easy to modify the Suri and Schultz (1998, 1999) value learning scheme to produce a model that predicts timescale invariant acquisition. However, the resulting model does not reflect more detailed regularities about acquisition times that are demonstrated experimentally. (Recall that the number of trials to acquisition in conditioning is roughly proportional to the interstimulus interval divided by intertrial interval, and that timescale invariance in this context is a result of this more specific pattern.)

In order to achieve timescale invariant acquisition we must eliminate two sources of timescale dependence from the model as originally published. One is exponential discounting, and the other is an assumption that the representation decays (exponentially, on top of the additional decay due to discounting), so that the basis functions corresponding to longer delays peak lower than those corresponding to shorter delays. Instead, assume an undiscounted, finite-horizon model with equal basis function scaling. In this model, the speed of acquisition (that is, the learning curve of the value $\widehat{\mathbf{V}}$ attributed to the stimulus at its presentation) is independent of the delay between the CS and US: whatever basis function spans the delay acquires value at the same rate, and that value is immediately attributed to the CS. So this version of the model captures the invariance of acquisition time to alterations in the overall speed of events, but not the more specific dependence of acquisition time on the ratio of inter-stimulus to inter-trial intervals.

On the issue of scalar variability in timed behaviors, the Suri and Schultz (1998, 1999) representation is basically equivalent to the tapped-delay line model of Montague et al. (1996; Schultz et al., 1997). Like their predecessors, Suri and Schultz (1998, 1999) do not include any timing variability, but the various proposals for adding it that I discussed in the context of the earlier model would play out similarly in the newer model.

4.3 A new TD model of the dopamine signal

Here I will develop a model of the dopamine system based on TD in a partially-observable semi-Markov process. First I lay out the general functional framework into which the model fits. I next present a simplified model based on a fully observable semi-Markov process, and then build up the full model by adding partial observability to the mix. The simpler model helps motivate many of the considerations involved in the more complicated model. After discussing some further details about how to relate the model equations to dopaminergic firing, I end with a discussion of how the model reduces to a number of different alternative models in different limits.

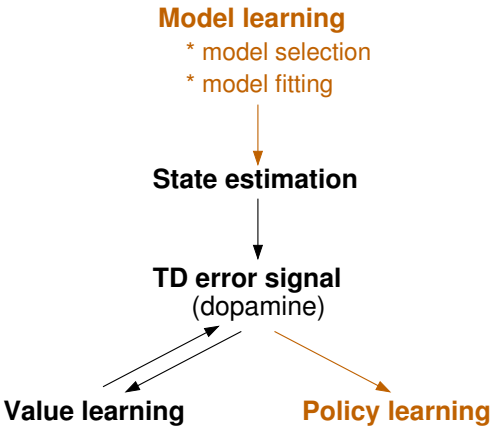

Figure 4.5: Schematic of the interacting learning functions suggested by the present model. Those in black are treated in detail in this chapter.

In this chapter, I will discuss only the problem of value prediction in semi-Markov processes, leaving aside the problem of action selection and suppressing the actions from all equations and models. This simplification is *not* inconsequential; to the contrary, it eliminates some difficult but important complications that I will discuss further below and in the chapter conclusion.

I wish at the outset to call attention to infelicitous vocabulary: three distinct uses of the word "model" collide in this section. A model such as a hidden Markov model is a generative statistical formalism that is supposed to capture the probability distributions over different sequences of events. Where confusion may result, I refer to this as a "formal model." Such formal models are the settings for various algorithms for prediction, action selection, etc. When I associate the behavior of the dopamine system with some signal in such an algorithm, I refer to this as a model of the dopamine system. I sometimes substitute the word "theory" for clarity. Finally, these algorithms sometimes make use of knowledge about the generative contingencies such as the state transition function. Such algorithms are known as "model-based" because the internal representation of the contingencies is known as a model of the process. Where necessary, I use terms such as "process model," "inference model," or "internal model" for this.

4.3.1 Value learning in a broader functional context

The theory presented here envisions the dopaminergic value learning system as part of more extensive framework of interacting learning systems than had been previously considered in dopaminergic models, and I wish here to lay out the general plan before describing any of the details. Below I describe algorithms for some of the pieces of this broad framework (state estimation and value learning). In this thesis, I do not provide detailed implementations for some of the other pieces (policy and model learning), as they are not essential for my primary goal of studying asymptotic dopamine responding. Some of these pieces *are* relevant to studying the model's predictions about acquisition times in classical conditioning experiments, and to that end I present some discussion and speculation about what the data suggest the final form of these components should be.

Figure 4.5 lays out the various portions of the model; the pieces implemented in this chapter are shown in black.

The basic idea of the system is to address prediction in the face of partial observability by using inference together with a statistical model of the world's contingencies to *infer* a probability distribution over the world's (unobservable) state, and then to use this inferred representation as a basis for learning to predict values using a TD algorithm. Thus we have:

- A **state estimation** system that infers the world's state (and related latent variables) using sensory observations and a world model. This is a *representational* system and might correspond to cortical

sensory processing systems.

- A **value learning** system that uses a TD error signal to learn to map this inferred state representation to a prediction of future reward. This portion of the system works similarly to previous TD models of the dopamine system.

The state estimation system requires access to a statistical model of the world's contingencies in order to infer its state. Obviously, such a model must be learned. A key idea of the present model is to recast the (fuzzy) problem of learning a representation for reinforcement learning into the well-defined problem of modeling the world's hidden processes. (This approach had not previously been considered in dopaminergic modeling, though it appears in various forms in the computational literature on reinforcement learning, e.g., Kaelbling et al., 1998). Thus we require:

- A **model learning** system that learns a one-step forward model of state transitions, state dwell times, and how states give rise to observable events. Learning such a model might involve a couple of distinct (or, likely, intertwined) stages:
  - A **model selection** stage to determine, for instance, how many states the model has and how they are connected.
  - A **model fitting** stage that optimizes the parameters of the appropriate model based on the animal's sensory observations.

I do not implement these functions here, though the problems are well-specified and algorithmically well understood. (How these algorithms scale with the timescale of events in the semi-Markov domain is *not* well understood, which is a subject of some speculation in this chapter and an important topic for future work.) In general, the goal of model fitting is to find model parameters that (locally) maximize the likelihood of observed data, which can be done using the expectation-maximization (EM) algorithm (Dempster et al., 1977); in the context of partially observable Markov models this is known as the Baum-Welch algorithm (Baum et al., 1970). The semi-Markov version is presented by Guedon and Cocozza-Thivent (1990). Here we would require an online version of these methods (e.g. Courville and Touretzky, 2001). Similarly, if we are uncertain about the model structure (e.g. the number of states), model selection algorithms can be used to select from a set of candidate models the one that best fits the observations (i.e. maximizes their likelihood, with some penalty for model complexity). Various approximations, sampling methods, and heuristics can be used to estimate the maximum likelihood model without fitting all of the candidates. (These are maximum likelihood approaches to model-selection and parameter fitting; a full Bayesian treatment would involve inferring distributions over model structure and parameters and integrating them out.)

Finally, the goal of all of this value prediction is to aid in selecting actions. Thus we need:

- A **policy-learning** system that uses information from the TD error (or potentially the value function and the model) to learn an optimal action selection policy.

Here I do not treat the function of policy learning; indeed I have eliminated actions altogether from the simplified formal model. There are significant complexities to action selection in a partially observable setting (described further in the chapter discussion), and coping with them is an important goal for future work.

This completes the description of the model components. As for the overall structure, this theory differs importantly from most previous dopaminergic theories in that it assumes that the system learns and makes use of an internal model of the process. Most previous dopamine theories learned value and policy functions directly from data without any intermediate world modeling. Here, an internal model is used to address representation in a partially observable domain. Of course, internal models are also directly useful for value estimation and action selection (Suri, 2001; Dayan, 2002), and, in principle, an internal model could be solved directly for the values (or for an action selection policy). It is important to note that just having an internal model does not itself solve the reward prediction or action selection problems without a great deal more work, for the same reason that knowing the rules of chess does not in itself tell you how to select winning moves or predict the victor from a board position. I return at the end of the chapter to further discussion about the rationale for the present approach.

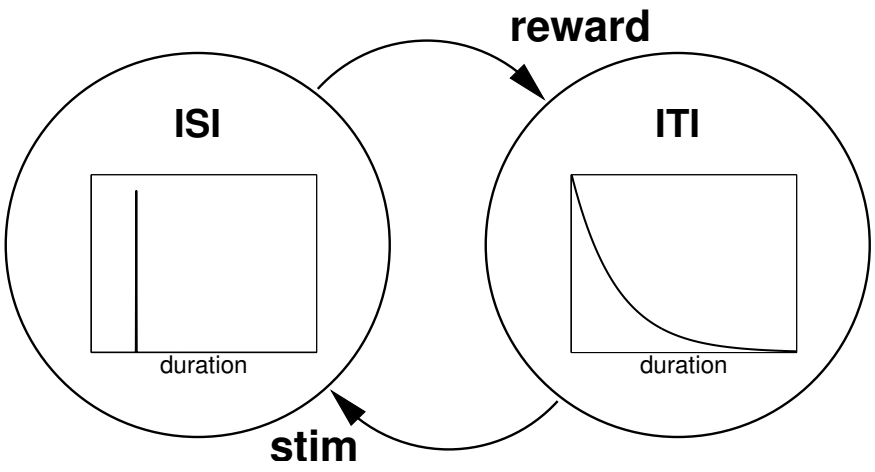

Figure 4.6: Semi-Markov structure of a trace conditioning experiment. States model intervals of time between events, which vary according to the distributions sketched in insets. Stimulus and reward are delivered on state transitions. ISI: Inter-stimulus interval; ITI: inter-trial interval.

4.3.2 A fully observable semi-Markov model of the dopamine response

Here I discuss a dopamine model based on a TD algorithm for learning in a fully observable semi-Markov process, as a way of motivating many of the issues involved in the more complicated partially observable semi-Markov model, which I present next. Because here I am assuming that the world's state is fully observable, the portions of the above framework dealing with world modeling and state inference are irrelevant, and I present only a TD algorithm for value learning.

Section 2.1.5 describes TD for a semi-Markov process. To recap, the key difference between the Markov processes we have used thus far and a semi-Markov process is that state transitions in a semi-Markov process can occur irregularly: the dwell time for each state visit is randomly drawn from a distribution $\mathbf{D}$ associated with the state. The time spent in some state s_k is d_k, drawn with probability $\mathbf{D}_{s,d} = P(d_k = d|s_k = s)$, where k indexes the irregular state transitions. (As described in detail in Section 2.1.5, we can also interchangeably subscript variables using a more traditional continuous time variable t.)

Every time a state transition occurs, the semi-Markov TD algorithm updates the value of the predecessor state based on the time dwelt in the previous state, the identity of the successor state, and the reward received. The TD error is Equation 2.14 on page 13. For the present purposes a slight difference in bookkeeping will be useful: we will not count the reward received on entering state s toward the value of that state, i.e. we omit the first term from the sum that defines the value function (Equation 2.13 on page 12). Additionally, in this chapter, we will use an average reward rather than an exponentially discounted return. Then the error signal is:

$$\delta_k = r_{k+1} - \rho_k \cdot d_k + \widehat{\mathbf{V}}_{s_{k+1}} - \widehat{\mathbf{V}}_{s_k} \tag{4.1}$$

where d_k is the time spent in state s_k, drawn on each visit from a distribution associated with the state, and ρ_k is the average reward estimate.

In the semi-Markov context, it makes sense to estimate ρ_k in a timescale-invariant, event-driven manner. The simplest such estimator is the average reward per timestep over the last n state transitions, i.e.

$$\rho_k = \sum_{k'=k-n+1}^{k+1} r_{k'} / \sum_{k'=k-n}^{k} d_{k'} \tag{4.2}$$

.

To connect this equation with dopaminergic responses, we must only define the state s in terms of the sensory events experienced by the animal. For this model I assume the state is fully observable (and also unitary), so that each state s_n uniquely corresponds to some event o_n that the animal observes. I will take

the events as instantaneous (punctate stimuli and rewards) and assume that o_n occurs on the transition into s_n. From this perspective, a trace conditioning experiment has a very simple structure, consisting of two states that correspond to the intervals of time between events (Figure 4.6; compare Figure 4.3). The CS is delivered on entry into the state labeled "ISI" (for inter-stimulus interval), while the reward is delivered on entering the "ITI" (inter-trial interval) state. This formalism is convenient for reasoning about situations where inter-event time can vary, since such variability is built into the model. Late or early rewards just correspond to longer or shorter times spent in the ISI state.

In this model, prediction error and learning updates are triggered entirely by external events (rather than ticks of clock time, as in the tapped delay line models). As I will discuss later in the chapter, this feature gives rise to timescale invariant acquisition in this fully observable model. Though the model assumes a timing mechanism (in order to measure the elapsed interval d_k between events used in the update equation), the passage of time, by itself, does not have any effect on learning or the progression of the state representation — only external events do. This is the main gap in the model from the perspective of the data considered here: The scheme cannot account for a dopaminergic pause when reward is omitted. It will instead sit in the ISI state with zero TD error until another event occurs, triggering a state transition and learning.

In the following section, I introduce a version of the model which corrects this flaw by incorporating *partial observability* (Section 2.1.6), which decouples states from events.

4.3.3 A partially observable semi-Markov model of the dopamine response

Here I extend the model described in the previous section to include partial observability. I specify the formal model, and discuss algorithms for state inference and value learning (see Figure 4.5).

Formally, partial observability results from relaxing the one-to-one correspondence between states and observations that was assumed above. The state evolves as before, but the agent's observations are now only a probabilistic function of the state, and not necessarily uniquely related to it. For our purposes here, we can continue to view each sensory observation as corresponding to some unitary punctate event such as a reward or a cue light, drawn from a set $\mathcal{O}$ of possible observations.[2] Each state specifies a probability distribution over members of $\mathcal{O}$. If we enter state s at time t, then the observation o_t takes the value $o \in \mathcal{O}$ with probability $P(o_t = o|s_t = s)$, which we abbreviate $\mathbf{O}_{s,o}$. One observation in $\mathcal{O}$ is distinguished as the null observation, $\emptyset$. If no state transition occurs at time t, then $o_t = \emptyset$. That is, non-empty observations can only occur on state transitions. Crucially, the converse is not true: *state transitions can occur silently,* with $o_t = \emptyset$. This is how the case of omitted reward in the trace conditioning experiment is modeled.

Section 2.1.6 discusses one method for value prediction in a partially observable Markov process: TD in the belief state MDP. The idea is that if the agent learns (or is given) a model of the Markov process — that is, the probabilistic functions controlling state transition, observations, and reward delivery — then, given a sequence of observations, it can recursively and continually estimate a probability distribution over the states, $\mathbf{B}_{s,t} = P(s_t = s|o_1...o_t)$, using Bayes' theorem. It's easy to verify (see Section 2.1.6) that these belief states $\mathbf{B}$ themselves form a Markov process, though one with a continuous state space. Thus standard TD can be used for value prediction simply by substituting the belief state $\mathbf{B}_{s,t}$ for the unavailable observed state s_t and using some function approximation method to tile the continuous state space.

Here I extend this general approach to the semi-Markov case: using a process model for state estimation, and TD methods to learn state valuations. In this section I present algorithms for state estimation and value learning assuming the model is known. These pieces are sufficient for exploring asymptotic dopamine neuron behavior. However, as already discussed, these pieces are envisioned in the context of a broader functional framework incorporating processes for learning the internal model

Fully observable semi-Markov models assume continuous time. When semi-Markov states are hidden (e.g. Guedon and Cocozza-Thivent, 1990), it is normal to instead assume that the time variable is discrete. This is to enable state estimation using a forward-backward state estimation scheme that recurses over timesteps.[3]

[2] It is straightforward to extend the formalism to include more structured observations that are vector-valued (e.g. to represent the simultaneous presentation of multiple punctate stimuli) and/or real-valued (to handle rewards of different magnitudes). For simplicity, I elide these features here.

[3] For this purpose, it would actually probably be sufficient to retain continuous time with the weaker assumption that state dwell times are bounded below by some minimum, to allow recursion over blocks of time containing at most one state transition. I do not pursue the extra complexity of such a scheme here.

Here I follow this convention, assuming that the system is clocked by some particular small timestep Δt. I further assume temporal units where $\Delta t = 1$, allowing me ti index the time variable discretely as t, $t+1$, etc.

A model of a partially observable semi-Markov process consists of three functions: the state transition function $\mathbf{T}$, the observation model $\mathbf{O}$, and the dwell time distributions $\mathbf{D}$. Given a sequence of observations experienced by an animal, the first step in the algorithm is to use this model for state estimation. In place of a probability distribution $\mathbf{B}$ over which state is *active* at some time given the observations, in hidden semi-Markov models, a useful quantity is the chance that the process *transitioned out* of each state at that time. The chance that the system left state s at time t is:

$$\beta_{s,t} = P(s_t = s, \phi_t = 1 | o_1 \dots o_{t+1})$$

In this equation, ϕ_t is a binary indicator which takes the value one if the state transitioned between times t and $t+1$ (self-transitions count), and zero otherwise. Note that this definition is conditioned on observations made through time $t+1$; this is chosen to parallel the one-timestep backup in the TD algorithm. β can be tracked using a version of the standard forward-backward recursions for hidden Markov models; the equations and their derivation are given in this chapter's appendix, Section 4.7.

Given the model, the values of the states could in principle be computed offline using value iteration. (Since the hidden process is just a normal semi-Markov process, partial observability does not affect the solution.) I will present an online TD algorithm for learning the same values.

First, a few caveats about what these values represent, and how they relate to expected future reward when we are uncertain about the state. Value in a semi-Markov process is defined as the expected discounted future reward *at the moment the state is entered* (see Equation 2.13 on page 12). As the process dwells in a state we do not attempt to learn rediscounted values based on the expected remaining dwell time (though these would be easy to compute using the model). Because of this somewhat peculiar definition, if we take the expectation of the values $\widehat{\mathbf{V}}$ with respect to a belief distribution $\mathbf{B}$ over the states, $\sum_{s \in \mathcal{S}} \mathbf{B}_{s,t} \cdot \widehat{\mathbf{V}}_s$, at some time when a state transition is not known to have just taken place, then the result does not precisely measure expected discounted future reward. This is because the values $\widehat{\mathbf{V}}_s$ being averaged do not themselves reflect expected discounted future reward, except when the process has just entered the state s. Given that a state's value is defined to remain static throughout a stay there, the weighted average of these values seems a reasonable measure of value expectancy when the state is uncertain. In any case, the expression exactly corresponds to discounted future reward when there is a non-empty observation (signaling a state transition) at t. This last fact is a result of the simplification of our formal model to not include actions. In contrast, in a partially observable Markov *decision* process, expected future reward is not a linear function of the belief state. This is due to the fact that the action selection policy can itself depend in a continuous manner on the belief state. In the present, simplified setting, we need not concern ourselves with more sophisticated nonlinear function approximation methods for the values.

Now, how do we adapt the TD rule of Equation 4.1 to learn a state's value when we are uncertain which state is active? Actually, our uncertainty is worse than just that. In standard semi-Markov TD, a TD update step is triggered by a state transition, but with partial observability, we may not know when those state transitions are taking place (o_t can be empty even when $\phi_{t-1} = 1$). Since β measures our belief that a state transition has taken place, it is reasonable to perform a TD update at every timestep, weighted by β. In particular, we compute a TD error for each state s and time t under the hypothesis that the system left state s at time t, weighted by $\beta_{s,t}$, the chance it actually did:

$$\delta_{s,t} = \beta_{s,t}(r_{t+1} - \rho_t \cdot E[d_t] + E[\widehat{\mathbf{V}}_{s_{t+1}}] - \widehat{\mathbf{V}}_s) \tag{4.3}$$

In these equations, the expected dwell time $E[d_t]$, and expected successor state value $E[\widehat{\mathbf{V}}_{s_{t+1}}]$ are computed from the observations using the model, conditioned on the hypothesis that the system left state s at time t:[4]

$$E[d_t] = \sum_{d=1}^{d_{max}} d \cdot P(d_t = d | s_t = s, \phi_t = 1, o_1 ... o_{t+1})$$

[4]Note that d_t is not defined when there has not been a state transition, i.e. when $\phi_t = 0$, but this is not a problem since $E[d_t]$ is computed under the hypothesis that there *has* been a state transition.

$$E[\widehat{\mathbf{V}}_{s_{t+1}}] = \sum_{s' \in \mathcal{S}} \widehat{\mathbf{V}}_{s'} P(s_{t+1} = s' | s_t = s, \phi_t = 1, o_{t+1})$$

where d_{max} is the maximum number of timesteps that could have been spent in s (which is the time since the last non-null observation, where a transition must have taken place). Expressions for computing these quantities in terms of the model parameters and the observations are given in this chapter's appendix, Section 4.7.

It's fairly easy to show that, assuming the inference model is correct (i.e. that it accurately captures the process generating the samples), this algorithm is exact in that it has the same fixed point as value iteration in the model. The proof is sketched in the appendix, Section 4.7.

In order to understand the algorithm, it's instructive to consider how the error signal, Equation 4.3, behaves in two special case conditions. When states and transitions are known with certainty, the learning rule correctly reduces to the standard semi-Markov TD rule of Equation 4.1. That is, if $\beta_{s,t} = 1$ for some known predecessor state s and the dwell time d and successor state s' are also known with certainty, then Equation 4.3 reduces to $\delta_{s,t} = (r_{t+1} - \rho_t d + \widehat{\mathbf{V}}_{s'}) - \widehat{\mathbf{V}}_s$ and $\delta_{s'',t} = 0$ for all $s'' \neq s$. Another interesting boundary case occurs when $\beta_{s,t} = 1$ and s and d are known, but the observation o_{t+1} reveals nothing about the successor state s'. In this case, the expected successor value $E[\widehat{\mathbf{V}}_{s'}]$ is computed entirely from the inference model using the transition probabilities $\mathbf{T}$. The resultant update is similar to a value iteration update of $\widehat{\mathbf{V}}_s$. (Unlike a value iteration update, shown in batch form in Equation 2.3, the observed sample of r_{t+1} is used in place of the expected reward from the model, and the predecessor value $\widehat{\mathbf{V}}_s$ is only updated by a small step toward the estimate.) Thus, depending on the level of state observability, this rule smoothly varies between sample-based TD updates and model-based updates along the lines of value iteration.

In this section I have described the state inference process and the TD error signal for learning values in a partially observable semi-Markov process. It may be useful to review how the computation actually proceeds. At each timestep, the system receives a (possibly empty) observation or reward, and the representational system uses this to update its estimate of the state departure distribution β and other latent quantities. The TD learning system uses these estimates to compute the TD error δ, which is reported by the dopamine system. Stored value estimates $\widehat{\mathbf{V}}$ are updated to reduce the error, and the cycle repeats.

4.3.4 Modeling dopamine responses: inference models

In this and the following sections, I discuss four further empirical and computational issues about using the error signal from Equation 4.3 to simulate the dopamine response. First I specify the characteristics of the inference models I am assuming. The remaining issues are how the dopamine system could transmit a vector error signal, how the present model is impacted by the partial rectification of negative error, and how we can model the effect of time measurement noise on the dopamine signal.

In the big-picture framework (Figure 4.5), the internal model for state inference would be learned from the observations. As already noted, I do not explicitly provide algorithms for model learning in this chapter but rather assume the model is well learned and then study asymptotic dopamine responding. Here I describe the characteristics of the inference models I am assuming.

The models I use are based on the actual generative models that produced the task events; for instance, for trace conditioning experiments, the model is based on Figure 4.6. However, each model is systematically altered to reflect an important assumption of the present work: The inference models incorporate *asymptotic uncertainty* about the world's contingencies. Such uncertainty can result from an assumption that the world is nonstationary (Kakade and Dayan, 2000, 2002a), and also from sensor and measurement noise. In the modified model, even a deterministic inter-stimulus interval (e.g. the ISI duration in Figure 4.6) will be represented as a distribution with nonzero variance (specifically, a Gaussian with scalar noise). Similarly, I assume that the model's representation of observation and transition contingencies is not entirely deterministic. This is achieved by requiring that the probability of anomalous observations and state transitions — for instance, the chance of receiving no reward on leaving the ISI state, or the chance that a state makes a self-transition — is bounded slightly away from zero. The effects of these modifications on the semi-Markov model of trace conditioning are illustrated in Figure 4.7. Compared to 4.6, the ISI stay duration is modified, and small chances of anomalous observations are shown. (There is a similarly small chance of self-transition, which I have left off the diagram for simplicity.)

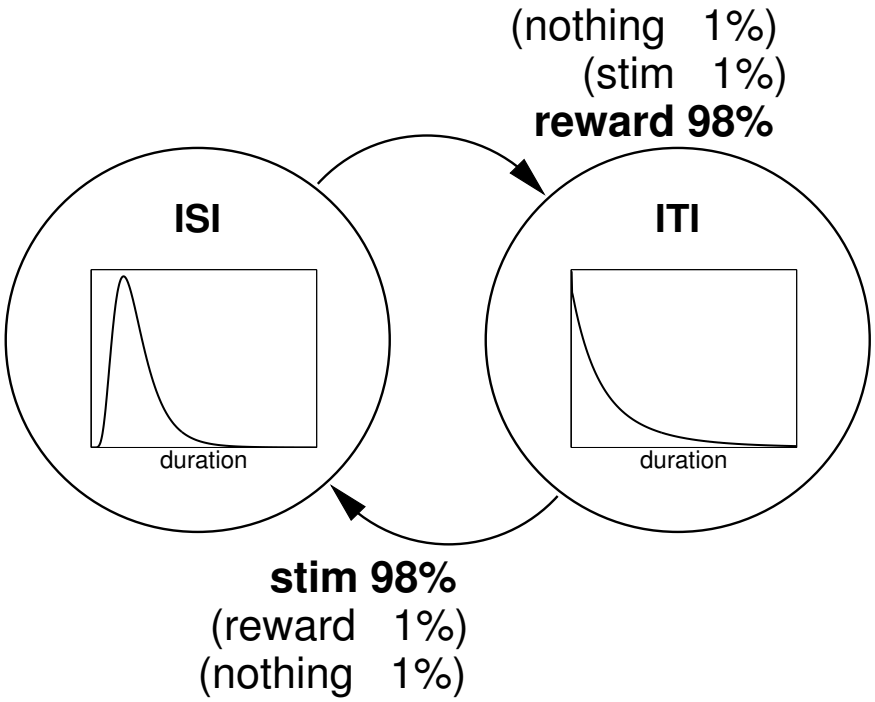

Figure 4.7: Semi-Markov model of trace conditioning experiment, with asymptotic uncertainty as to dwell time duration and observations.

4.3.5 Modeling dopamine responses: vector versus scalar signaling

Equation 4.3 has one important difference from previous TD models: it is a vector rather than a scalar signal. That is, since the current state of the process is uncertain, error is computed for every state the process might be in and apportioned according to β. Experimental reports have noted striking similarity between the responses of most dopaminergic neurons (Schultz, 1998), which has been taken to support the idea that the system carries a scalar signal.

However, there is some variability between neurons; notably, only a subset (55-75%) of neurons displays any particular sort of phasic response (Schultz, 1998). Several studies report a different sort of variability, in which different groups of neurons (each representing sizeable fractions of the population) displayed either excitatory or inhibitory responses (or both) to the same situation (Schultz and Romo, 1990; Mirenowicz and Schultz, 1996; Tobler et al., 2001; Waelti et al., 2001). There are also anecdotal observations that neurons differ in terms of their differential sensitivity to graded variations in reward probability (Fiorillo et al., 2003).

It might be possible to specify a model in which response variability between dopamine neurons is due to the neurons coding a vector error signal like Equation 4.3 in some distributed manner. As none of the above-mentioned experiments was designed to study vector coding in a partially observable context, it is difficult to determine whether the observed variability was sufficient or of the right sort to signal a vector error. Also, it is not clear whether the dopaminergic projections (which are rather diffuse) are organized to support such vector coding.

With those caveats in mind, let us for concreteness specify a proposed vector signaling scheme. It cannot be the case that each dopamine neuron signals $\delta_{s,t}$ for some particular candidate state s — in this case, since $\delta_{s,t}$ is weighted by $\beta_{s,t}$, the neuron would be silent most of the time (when s was not thought to be active), contradicting the finding that a large subset of dopamine neurons responds any particular event that induces TD error, regardless of the likely world state. We can instead imagine that any particular dopamine neuron signals the cumulative error signals over a particular random subset $\mathcal{S}'$ of states, $\sum_{s \in \mathcal{S}'} \delta_{s,t}$. If $\mathcal{S}'$ represents a sizeable fraction of all states $\mathcal{S}$ (e.g. 55-75%), then this scheme would seem to be consistent with the data on response variability reviewed above. However, in addition to the other caveats mentioned

above, the problem of reconstructing the error for any particular state from this signal at dopamine targets would be extremely difficult, and it probably could not realistically be solved exactly.

In any case, for this scheme, the average signal over a population of dopamine neurons would have the form:

$$\delta_t = \sum_{s \in \mathcal{S}} \delta_{s,t} \tag{4.4}$$

i.e. the cumulative TD error over all states. (Each term in the sum is weighted implicitly by $\beta_{s,t}$, since this factor occurs in the expression for $\delta_{s,t}$.) This is also a reasonable expression for what a "typical" dopamine neuron's response would look like on this scheme.

An alternative is to assume that the dopamine neurons signal a scalar quantity that can be used to reconstruct approximately the vector signal at targets. In particular, if every dopamine neuron carried the cumulative signal from Equation 4.4, then learning could be apportioned at targets in proportion to $\beta_{s,t}$; i.e. $\Delta\widehat{\mathbf{V}}_s \propto \beta_{s,t} \cdot \delta_t / \sum_{s' \in \mathcal{S}} \beta_{s',t}$. This approximation behaves well in situations where the posteriors over β, d and state occupancy are sharply peaked, since it is equivalent to the exact algorithm in the limit of full observability. In practice, this approximation performed well (and the same method could also be adapted to provide better approximate reconstruction using the vector signaling scheme above).

Thus, in this chapter I use the scalar signal from Equation 4.4 (together with rectification of negative error, described next) to simulate the response of a typical dopamine neuron or population of neurons, without having to definitively decide between vector and scalar coding.

4.3.6 Modeling dopamine responses: rectification of negative error

There is another issue, having to do with the rectification of negative error. As I shall demonstrate shortly, in the present model, the TD error from Equation 4.4 in response to some event like reward is often highly variable from trial to trial, but with a mean error (averaged over trials) of zero. Recall (from Figure 3.9) that dopaminergic firing rates are only proportional to TD error down to an error threshhold that is slightly negative. This effect would skew the measured average firing rate in situations like the ones just described, when the TD error varies between negative and positive values from trial to trial. In order to simulate the result of averaging rectified dopamine firing rates over a series of trials, I plot the simulated dopamine response as proportional to the mean TD error over trials, with error on each trial truncated at a slightly negative threshhold. That is, let $\delta_{t,T}$ be the error trace from Equation 4.4 for trial T (a function of time since the beginning of the trial) of N, and $[\delta_{t,T}]_{-\psi}$ be the same trace with all errors less than $-\psi$ truncated to equal exactly $-\psi$. Then the mean rectified dopamine signal is simulated as $\sum_{T=1}^{N} [\delta_{t,T}]_{-\psi} / N$. (This is a simpler scheme than the one described in the previous chapter involving assigning trial-by-trial spike counts and averaging over those.)

4.3.7 Modeling dopamine responses: the effect of timing noise

I will in some cases wish to consider the effect of time measurement noise on the dopamine signal. In order to understand how such noise can be added, it is necessary to consider how the system times intervals. The update rule for the *fully observable* semi-Markov TD algorithm (Equation 4.1) incorporates a sort-of black box measurement d_k of the time dwelt in a state, to which appropriate noise can be added directly. In the complete *partially observable* model (Equation 4.3), the expected dwell time $E[d_t]$ plays the same role. For the normal case of a dwell in a state framed by two events, this is essentially computed by counting the number of discrete timesteps Δt that occur between the events (the equations are given in the Appendix). Thus the timestep Δt between learning and inference steps controls the system's timing, similar to the tapped delay line model discussed in Section 4.2.4. Scalar variability in the elapsing counts can be accomplished using correlated noise in the timesteps. For this, assume the timestep Δt fluctuates around some small fundamental timestep (say, 100 ms), with a new value chosen randomly whenever a non-empty observation o occurs. To select a new Δt, we draw a timer scale factor from some distribution (say, a lognormal or truncated normal, with a mean of 1 and a standard deviation between 0.3 and 0.5, based on the measurements of Gibbon, 1977). The adjusted timer speed remains in effect until a new non-empty

	Semi-Markov	**Markov**
Partially observable	Equation 4.3	Figure 4.8
Fully observable	Equation 4.1	Equation 3.3 (Chapter 3)

Table 4.1: Matrix of different possible TD models related to the present one.

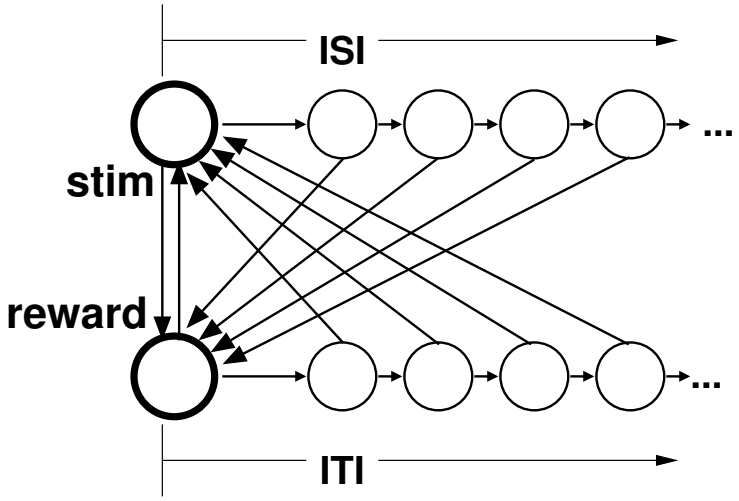

Figure 4.8: Hidden Markov model equivalent to the semi-Markov model of a trace conditioning experiment from Figure 4.6. Here, ISI and ITI states are broken up into a series of substates that mark the passage of time. Stimuli and rewards occur only on transitions from one group of states into another (transitions into the two states denoted with larger and darker circles); the distributions over the semi-Markov states' dwell time durations is encoded in the relative likelihoods of the two transitions coming out of each marker state.

observation occurs. Then, assuming o is always followed by o' after some constant amount of real clock time t, the inter-event counts measured during that period will be distributed as (discretized) scalar transforms of the timer scale factor distribution, that is, as t multiplied by the distribution. This is another variation on the idea (e.g. Gibbon, 1992) of correlated timer noise; here the timescale of events is used to set the timescale of correlation in the the noise.

There is one further wrinkle: Studying average responding with timing noise requires averaging over error signal traces that, due to the simulated animal's clock running faster or slower, are sampled at different discrete rates. This problem is trivially solved in the fully observable limit, since that error signal is defined to be continuously zero between events. When I encounter the problem in the context of the partially observable model's error signal (which occurs only in Figure 4.18), I solve it provisionally by upsampling the error signal traces to a common discretization timestep using Matlab's spline interpolation prior to averaging. A more realistic and elegant solution might be had by averaging firing rates produced by a model that translated the discrete errors into spike trains.

4.3.8 Limiting cases of the partially observable semi-Markov TD model

In this section, I place the models discussed in this section in the context of a broader family of TD models, by considering how alternative TD schemes follow from the model of Equation 4.3 in various limits.

The present model extends models like the one presented in Chapter 3 in two directions: partial observability and semi Markov dynamics. Different subsets of these features can be combined to produce three other models, shown in the matrix in Table 4.1.

The top-left corner in the table, the **partially observable semi-Markov model**, corresponds to the complete model presented in this chapter. I have already shown that its learning rule reduces to the learning

rule for the model below it, the **fully observable semi-Markov model,** in the limit of full observability (i.e. the limit in which the observation distribution at each state is sharply peaked at a unique observation that identifies a transition into that state). Most of the tasks modeled in this chapter are fully observable in this sense (the exceptions are those involving reward omission). In these situations, the only differences between the simpler and more complex models are due to the assumption of asymptotic uncertainty in the internal observation model (e.g. Figure 4.7). As long as such uncertainty is minimal, the fully observable model behaves more or less identically to its more complex sibling. For this reason, and because a number of important properties of both models can be presented more clearly in the simplified limit, I use the fully observable version of the model for a number of the simulations that follow.

How do the semi-Markov models on the left of the table relate to the Markov ones on the right? The Markov and semi-Markov dynamics are formally equivalent: an equivalent Markov model can be constructed from any semi-Markov model by subdividing each semi-Markov state into a series of Markov time-marker substates. (See Figure 4.8. For a continuous time semi-Markov process, this equivalence only holds up to the limit of arbitrarily fine discretization error, but here I have assumed a discrete-time semi-Markov process anyway.)

Though the models are equivalently descriptive in terms of dynamics, they differ as to how values are defined, and these distinctions play out as differences in the value learning algorithms and in turn as significant differences in the modeled dopamine behavior. The Markov valuation scheme involves assigning values to all of the many extra time marker states in Figure 4.8, and its learning rules thus engage in learning at every timestep about one of these immediate values. The semi-Markov version defines only a single value for each temporally extended state; thus it learns only sporadically, at state transitions (though the partially observable version learns at every timestep, weighted by the probability of a state transition). As I will demonstrate later in this chapter, this difference gives rise to differences in the way the cost of delays are accounted for, which has implications for the tonic prediction error effects studied in Chapter 3.

With all this in mind, we can define a limit in which the semi-Markov valuation schemes from this chapter reduce to Markov valuation schemes. Starting with the partially observable semi-Markov model, imagine that we impose a maximum dwell duration on each state. If the interval between events can exceed the maximum duration, further states must be introduced to account for the remainder of the interval, and each of these states will learn an intermediate value. If the maximum dwell time is long, the situation will not be much different from the standard semi-Markov model, though long dwell times will be interrupted by an intermediate value learning step. As the maximum stay duration becomes shorter, the number of marker states increases and the valuation scheme approaches a Markov one. Finally, when the maximum dwell time equals the time step size Δt, the partially observable semi-Markov model reduces to the **partially observable Markov model**. In this case, since there is no longer any need to reason about state dwell times, a simpler learning rule (TD in the belief state Markov process) can be used.

If we apply this same construction to the fully observable semi-Markov model, we recover what I have called a **fully observable Markov model** (Table 4.1, bottom right), similar to the model described in Chapter 3 and to previous TD models. Since observations occur sporadically, this model is not actually fully observable in the sense of having each of its states correspond to a unique observation. Rather, each observation gives rise to a unique series of time marker states like a tapped delay line.

Thus I have shown how the various models in Table 4.1 follow as limiting cases of the partially-observable semi-Markov model proposed in this chapter. I will now turn to simulations of the behavior on various tasks of the fully and partially observable semi-Markov models on the left side of the table; at the end of the chapter I will discuss how this behavior differs from the fully observable Markov model of Chapter 3 and its partially observable counterpart.

4.4 Dopamine responses in the model

Here I present simulations demonstrating the predictions of the semi-Markov TD model about the responses of dopamine neurons in various situations. I will discuss basic tasks involving signaled and unsignaled rewards (e.g. Figure 2.2), as well as variations on them that studied variability in event timing (Figures 4.1 and 4.2). Finally, I will discuss tonic error effects in the model, and how they relate to the predictions made in Chapter 3 using a Markov TD model. For each set of tasks, I first present the simulation results, and then discuss

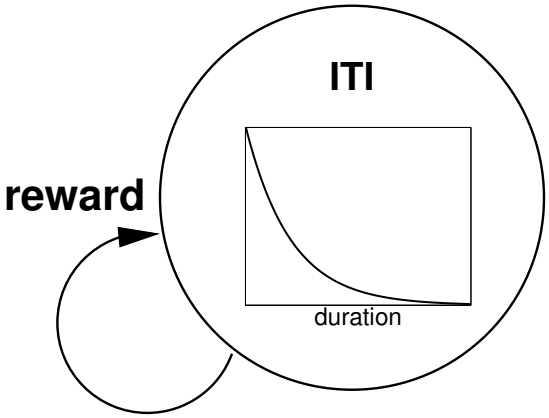

Figure 4.9: Semi-Markov model of free reward experiment, containing a single state.

how the results compare to experimental data and to previous models.

Where appropriate, I will for clarity present results based on the simpler fully observable limit (Equation 4.1) of the partially observable semi-Markov model (Equation 4.3).

4.4.1 Results: Free reward delivery

The simplest experimental finding on dopamine neurons is that they burst when animals receive random, unsignaled reward (Schultz, 1998). The semi-Markov model's explanation for this effect is rather surprisingly different than the usual TD account.

This "free reward" experiment can be modeled as a semi-Markov process with a single state (Figure 4.9). Assuming Poisson delivery of rewards with magnitude r, mean rate λ, and mean inter-reward interval $\theta = 1/\lambda$, the dwell times D are exponentially distributed. Let's examine the TD error, using Equation 4.1. The state's value $\widehat{\mathbf{V}}$ is arbitrary (since it only appears subtracted from itself in the error signal) and $\rho = r/\theta$. The TD error on receiving a reward of magnitude r after a delay d should be:

$$\begin{aligned} \delta &= r - \rho d + \widehat{\mathbf{V}} - \widehat{\mathbf{V}} \\ &= r(1 - d/\theta) \end{aligned}$$

which is positive if $d < \theta$ and negative if $d > \theta$, as illustrated in Figure 4.10, left. That is, the TD error is relative to the expected delay θ: rewards occurring earlier than usual have higher value than expected, and conversely for later-than-average rewards.

Figure 4.10, right top, confirms that the TD error averaged over multiple trials is zero. However, due to the partial rectification of negative error, excitation dominates in the simulated dopamine response (Figure 4.10, right bottom), and net phasic excitation is predicted.

Figure 4.11 compares the situation when reward timing still varies, but less drastically than the Poisson variation described above. Here, I assumed the inter-reward interval was uniform between 4.5 and 5.5 seconds. Though the sign and magnitude of error on each trial still vary according to the duration of the preceding delay, the magnitude of the negative error does not exceed the level of rectification, so the mean rectified error (like the mean TD error) is zero, and no dopamine response is predicted. In general, the extent to which rectification biases the average dopaminergic response to be excitatory depends on how often, and by how much, negative TD error exceeds the rectification threshhold. This in turn depends on the amount of jitter in the term $-\rho d$, with larger average rewards ρ and more sizeable jitter in the interreward intervals d promoting a net excitatory response.

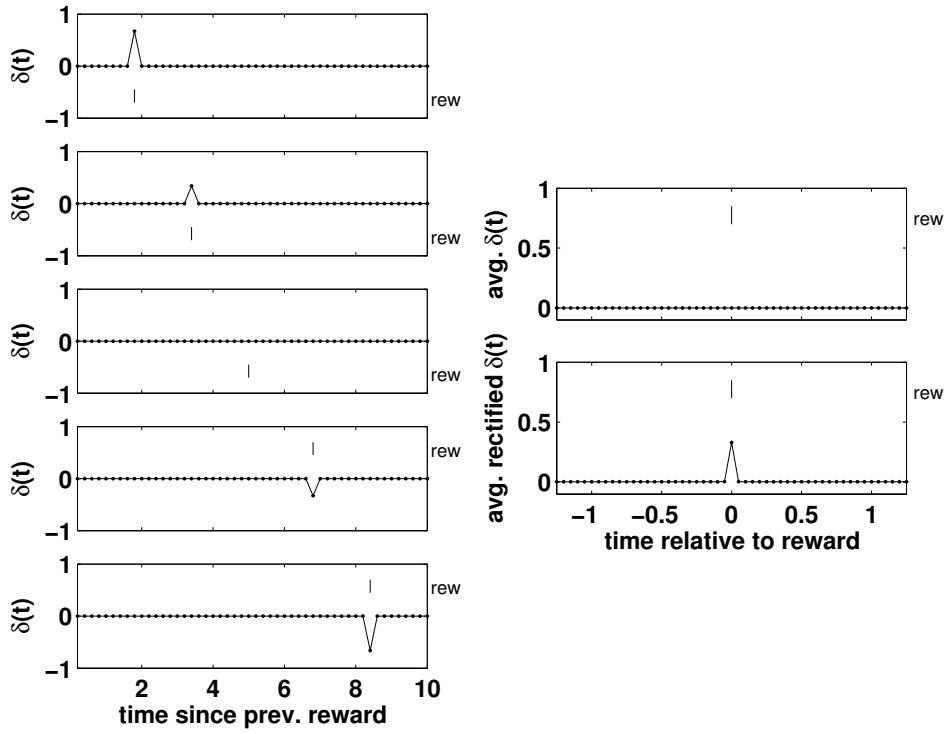

Figure 4.10: TD error to reward delivery in free reward experiment when reward timing varies on a Poisson schedule, using the semi-Markov TD model. Left: Error ranges from strongly positive through zero to strongly negative (top to bottom), depending on the time since the previous reward. Right: Error averaged over trials. Right top: Mean TD error over trials is zero. Right bottom: Mean response over trials with negative errors partially rectified (simulated dopamine signal) is positive. Mean inter-reward interval: 5 sec; reward magnitude: 1; rectification threshhold: 0.1.

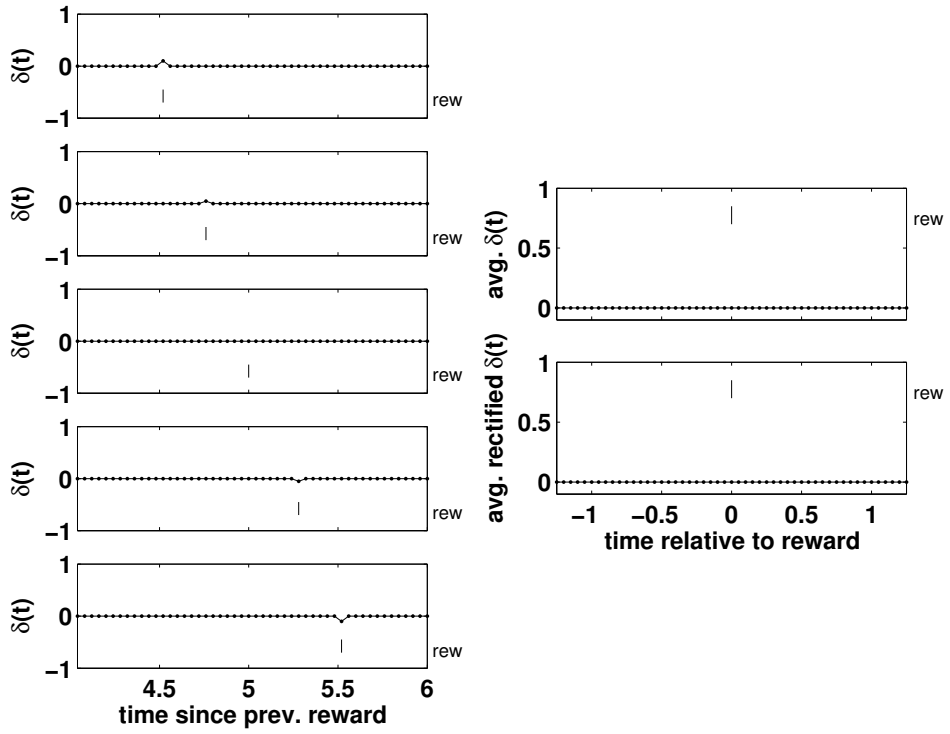

Figure 4.11: TD error to reward delivery in free reward experiment when reward timing varies only slightly, using the semi-Markov TD model. Left: Sign and magnitude of error depend on the time since the previous reward. Right top: Mean TD error over trials is zero. Right bottom: Mean response over trials with negative errors partly rectified (simulated dopamine signal) is also zero, because the negative error never exceeds the rectification level. Inter-reward interval: uniform random between 4.5 and 5.5 seconds; reward magnitude: 1; rectification threshhold: 0.1.

4.4.2 Discussion: Free reward delivery

The simulations in Figure 4.10 reproduce a fundamental property of dopamine neurons, phasic excitation to unpredicted rewards (Figure 2.2, top), but the details of the explanation differ from those offered by previous models, in a way that gives rise to new, testable predictions. Earlier models assumed the response was due to strong, uniform positive prediction error on every trial. In contrast, the new model predicts a pattern of trial-by-trial variability in the dopamine response, with net excitation emerging only in the average over trials. The new model (Figure 4.11) also explains why dopamine neurons can fail to respond to rewards whose timing is somewhat unpredictable (Figure 4.1, "no task," may be an example of this). As discussed in Section 4.2.3, a Markov TD model can only account for this phenomenon when the variability is extremely small.

This is the first example we will encounter of a key difference between the way the Markov and semi-Markov learning schemes account for the costs of delay. The general difference is easiest to see by comparing the Markov and semi-Markov versions of the average reward TD learning rule, Equation 2.12 on page 11 versus Equation 4.1 on page 87. In Markov TD, the cost $-\rho d$ of a delay between, say, a stimulus and reward is apportioned piecewise by subtracting ρ once at each of the d Markov state transitions occurring between the events. In semi-Markov TD the entire amount ρd is subtracted all at once, in a single update occurring at the single state transition, producing negative error for rewards following long delays.

The new model accords with the experimental results in the average, but predicts a testable pattern of trial-to-trial variability in dopamine responses on experiments such as the one discussed here, with the magnitude and direction of dopaminergic responding controlled by the preceding inter-reward interval. There has been no report of dopaminergic responses in this task broken down by the preceding interval, the most direct test of the prediction. There has not even been a published analysis of trial-to-trial variability in dopamine responses on this (or any other) task, which could at least verify whether dopaminergic responses to free reward are often below baseline, as the theory predicts. This (less monumental) prediction can easily be verified by counting — or indeed eyeballing — the spikes shown in Figure 2.2. There is one further piece of suggestive behavioral evidence on this point. The latency to animals' responses in many operant tasks is correlated with the previous interfood interval, with faster responding after shorter intervals ("linear waiting," see Staddon and Cerutti, 2003, for a review). Given the effects of dopaminergic manipulations on impulsivity and response withholding (see Section 2.2.3), it seems possible that this effect reflects enhanced dopaminergic activity after shorter interfood intervals.

There are some practical caveats to testing the prediction that longer-than-average inter-reward delays should provoke dopaminergic inhibition in response to the reward that finally arrives. First is the subtle nature of dopaminergic inhibition, which is difficult to detect due to low baseline firing rates. Even in the canonical reward omission task, inhibition is not visible in 30% of dopamine neurons (Hollerman and Schultz, 1998). It would thus be important to verify that a neuron tested for inhibition on the free reward task is able to show it on the reward omission task. It is also important that the animal be well-trained in the free reward task; the theory predicts that positive error should dominate until the prediction parameters (ρ here and $\widehat{\mathbf{V}}$ in other similar cases) reach their asymptotic values. Finally, because of interval measurement noise, trial-to-trial variability in the animal's subjective intervals will still be an important factor even when trials are sorted by programmed inter-reward delay. Such variability would attenuate (though not wholly eliminate) the predicted effects when averaged over trials. Thus it would be useful to examine the distributions of single-trial spike counts.

4.4.3 Results: Signaled reward, overtraining, and timing noise

When a reward is reliably signaled by a stimulus that precedes it, dopaminergic responses famously transfer from the reward to the stimulus (Figure 2.2). The corresponding semi-Markov model is the two-state model of Figure 4.6. Many of the same issues just discussed in the context of free reward experiments are also seen in this more complicated task.

Here I will assume that the intertrial intervals are exponentially distributed and that the CS-US delay is fixed, and defer discussion of reward omission until the next section. Thus the partially observable semi-Markov model can once again be usefully approximated by its fully observable limit (Equation 4.1).

In this case, the model predicts transfer of the reward response to the signal, as shown in Figure 4.12.

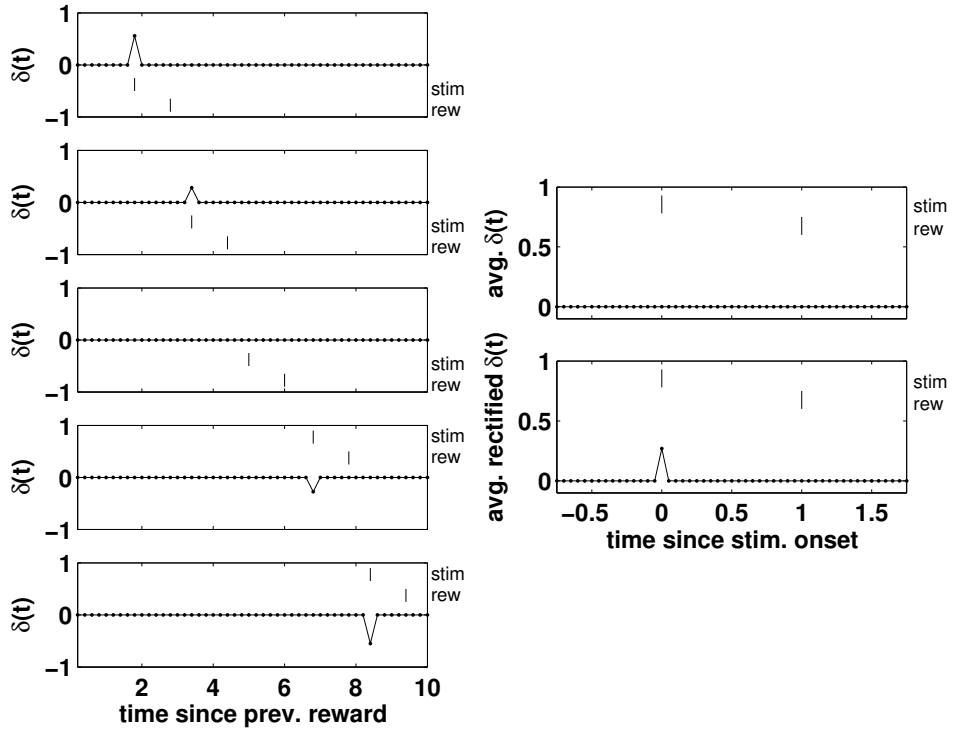

Figure 4.12: TD error to stimulus and reward delivery in signaled reward experiment, using semi-Markov TD model. Left: Error to signal ranges from strongly positive through zero to strongly negative (top to bottom), depending on the time since the previous reward. There is no error to the reward delivery. Right: Error averaged over trials. Right top: Mean TD error over trials is zero. Right bottom: Mean response to stimulus over trials with negative errors partially rectified (simulated dopamine signal) is positive. Mean ITI: 5 sec; stimulus-reward interval: 1 sec; reward magnitude: 1; rectification threshhold: -0.1.

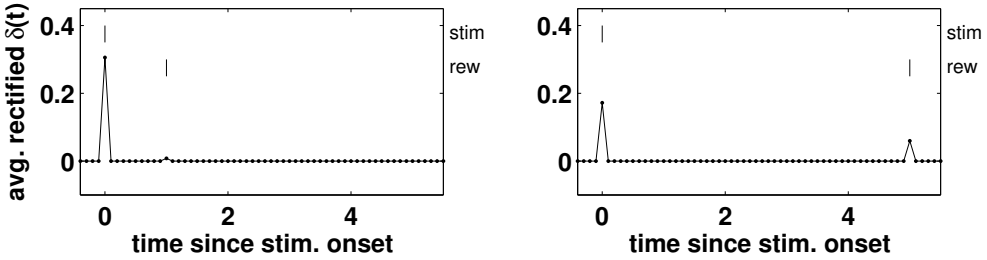

Figure 4.13: Effect of timing noise on modeled dopaminergic response to signaled reward depends on the interval between stimulus and reward (ISI). Left: For ISI=1 sec., error variation due to timing noise is unaffected by rectification and response to reward is minimal. Right: For ISI=5 sec., error variation from timing noise exceeds rectification level, and response to reward begins to emerge. Mean ITI: 5 sec.; reward magnitude: 1; rectification threshhold: -0.1.; coefficient of variation of timing noise: 0.5.

Analogously to free reward, the CS response under this model is excitatory or inhibitory depending on the length of the delay that preceded it; again, the average CS response will be excitatory due to rectification of negative-error trials.

Let us now consider the additional effects of scalar time measurement noise on the modeled dopaminergic response, as described in Section 4.3.7. (So far, I have assumed noiseless timing.) Figure 4.13 demonstrates that this noise has negligible effect when the delay between stimulus and reward is small (only very slight new excitation is seen to the reward), but for a longer delay, the modeled dopaminergic response to the reward re-emerges and cannot be trained away.

4.4.4 Discussion: Signaled reward, overtraining, and timing noise

The results in this task replay many of the same issues discussed in the context of the free reward task.

The difference in reward response between the left and right sides of Figure 4.13 is analogous to the difference between Figures 4.10 and 4.11. Here, timing noise introduces jitter in the measured interval between stimulus and reward, which results in trial-to-trial variability in the sign and magnitude of the response to the reward. For a short stimulus-reward interval, the negative error rarely exceeds the rectification level, and so in the rectified average over trials, near zero response to the reward is seen. But since the timing noise is scalar, the variability is more substantial when the stimulus-reward delay is increased. Negative error now exceeds the rectification level, and a net positive response to the reward is seen. This modeled behavior is consistent with the unpublished finding, discussed in Section 4.1, that the dopaminergic response to reward cannot be trained away when the stimulus-reward interval is long (C. Fiorillo, personal communication, 2002).

Another version of the same effect as shown in Figure 4.11 explains the result (Figure 4.1, "Overtrained" condition) that the response to even the reward predicting stimulus can disappear in overtrained animals, despite what was evidently moderate variability in the intertrial interval. The model again explains this phenomenon as occurring because the moderate temporal jitter in the interval preceding the stimulus fails to produce prediction error that exceeds the rectification threshhold, so positive and negative errors balance in the rectified average. (I do not provide a figure, as the situation is exactly analogous to Figure 4.11.)

The results shown in Figure 4.1 (top) from Schultz et al. (1993) are also explained by the same effects demonstrated in the simulations here. When the trigger stimulus timing does not vary, it evokes no prediction error. Increasing the length and variability of the delay preceding it causes a response exactly analogous to the stimulus depicted in 4.12: per-trial error to the trigger stimulus varies with the delay preceding it, and the average response skews positive due to rectification.

Thus so far we have demonstrated a number of experiments in which the semi-Markov model correctly captures the effect of varying the timing of a stimulus, including some that defeat previous Markov models. I now proceed to consider two more experiments that studied the effect of timing variability more systematically.

4.4.5 Results: State inference and variation in event timing

Here I discuss several more experiments that studied the effect of varying interstimulus timing in a number of ways. In some of these cases, state inference may play an important role; thus in this section I will simulate the system's behavior with the complete model of Equation 4.3 either in addition to or instead of its fully observable limit, Equation 4.1.

Figure 4.14 illustrates a semi-Markov model of the experimental contingencies in the task of Fiorillo and Schultz (2001), in which a cue was followed 1-3 seconds later (chosen uniformly) by reward. Four different simulations of this task, exploring different aspects of the TD model, are shown in Figure 4.15. For each simulation, traces are shown for a range of stimulus-reward delays. Though the original task was delay conditioning — i.e. the cue persisted throughout the stimulus-reward interval — I have treated it here as a trace conditioning experiment with a punctate cue. This has no effect on the reported results.

The top left simulations in Figure 4.15 ("tapped delay line model") illustrate that, as discussed in Section 4.2.3, the tapped delay line TD model of Montague et al. (1996; Schultz et al., 1997) predicts that the delay to reward should have no effect on the response to reward. The rest of the simulations display different versions of the semi-Markov model. In all of these cases, a contrasting and by-now familiar pattern of results

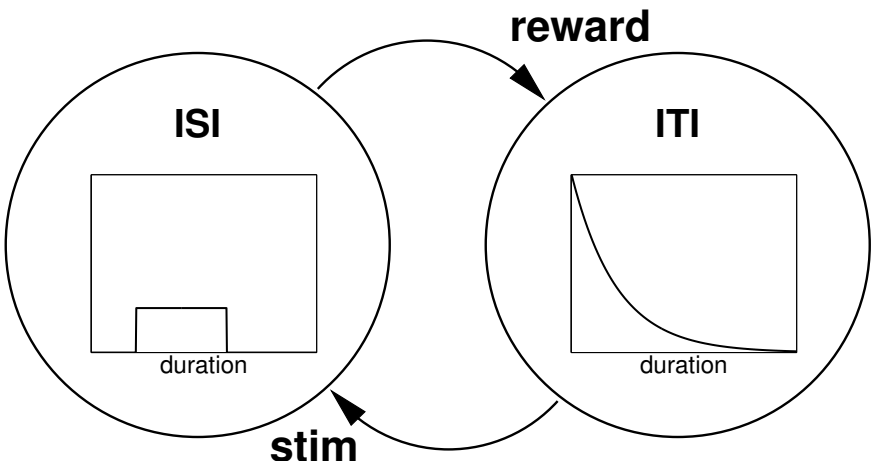

Figure 4.14: Semi-Markov model of the experiment of Fiorillo and Schultz (2001).

is visible: The level of response to reward is graded by the delay preceding the reward, with inhibitory responding predicted for longer-than-average delays.

The simulations on the right side of Figure 4.15 compare the fully observable limit of the model with the complete model including partial observability and state inference. Since this task does not involve any partial observability, it is not surprising that the fully observable limit performs well, but I include the partially observable version to allow comparison with results on some of the other tasks discussed below. For the partially observable model, the inference model was based on the semi-Markov model shown in Figure 4.14, modified as discussed in Section 4.3.4 to include asymptotic uncertainty represented by a 1% chance of each possible anomalous event such as missed reward or self-transition. A shaded bar underneath each trace in the partially observable results reports the state distribution inferred by the internal model (with white for ITI, black for ISI and shades of grey for uncertain intermediate beliefs); these inferred states track the true states accurately.

Finally, the bottom left simulations in Figure 4.15, labeled "Semi-Markov with noise," display the effect of time measurement noise on the predicted results. Here I have used the fully observable limit of the model, whose appropriateness we have just verified. Because timing noise is scalar, it differentially affects the results at longer delays. Specifically, the noise attenuates somewhat the predicted inhibition at longer delays. This is another reflection of a familiar effect: longer true intervals give rise to a range of subjective intervals, and the average over trials is biased in the positive direction due to rectification of the dopamine response.

Next, I consider the effect of reward omission and of rewards expected at a particular time but delivered early or late in occasional probe trials. According to the model, state inference plays an important role in these experiments: As the interval grows following a stimulus with no reward, the state inference process gradually decides that an unsignaled state transition has taken place. The appropriate inference model is the semi-Markov model of a trace conditioning experiment (Figure 4.6), but modified to incorporate asymptotic uncertainty as discussed in Section 4.3.4 and shown in Figure 4.7.

Figure 4.16 shows the effect of reward omission in the new model, compared to the tapped delay line model of Montague et al. (1996; Schultz et al., 1997). The tapped delay line model predicts a single, sharp burst of negative TD error, occurring in the exact tapped delay line bucket in which reward was expected. By comparison, in the semi-Markov model, as time passes without reward, inferred probability mass leaks into the ITI state (shown as progression from black to white in the shaded bar under the trace), accompanied by negative TD error. Thus the predicted dopaminergic inhibition is slightly delayed and smeared out compared to the expected time of reward delivery.

Hollerman and Schultz (1998) generalized the reward omission experiment to include probe trials in which the reward was delivered a half-second early or late. Figure 4.17 compares the tapped delay line model (Montague et al., 1996; Schultz et al., 1997) and my partially observable semi-Markov model on this

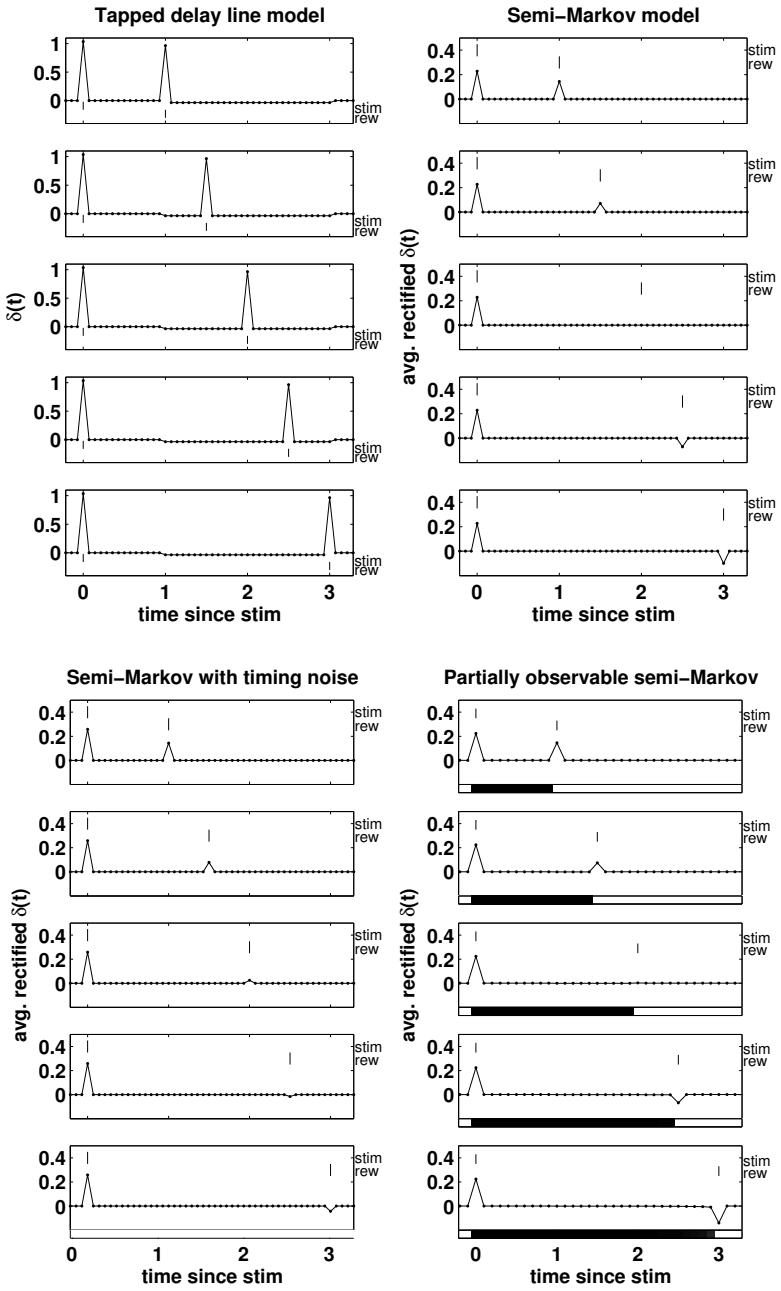

Figure 4.15: Models of dopaminergic response to signal and reward, when the ISI is randomized uniformly as in the experiment of Fiorillo and Schultz (2001). Top left: In the tapped delay line model of Montague et al. (1996), response to reward does not depend on the CS-US interval. Top right: In the semi-Markov model, the reward response is positive for short CS-US intervals and declines to negative for long CS-US intervals. Bottom left: Scalar timing noise selectively affects the average TD error at longer delays, decreasing the inhibition predicted for long delays. Bottom right: The partially observable model behaves the same as the fully observable model on this task. The shaded bar underneath each error trace tracks the system's probability distribution over the hidden state; ITI: white, ISI: black. Mean ITI: 5 sec.; ISI: uniform between 1 and 3 secs.; reward magnitude: 1; rectification threshhold: -0.1; coefficient of variation of timing noise: 0.5; probability of each anomalous event inference model: 1%.

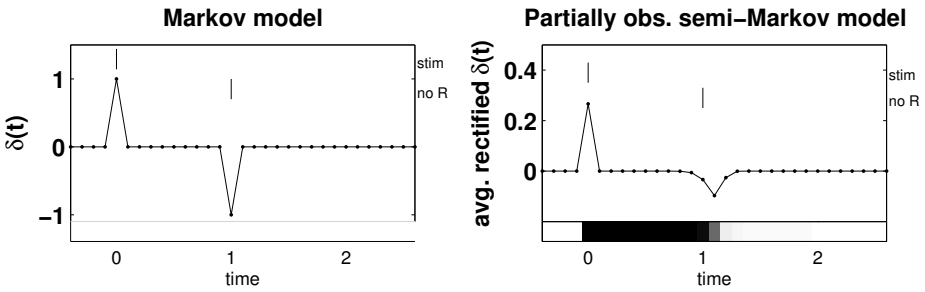

Figure 4.16: Simulated dopamine response to omission of expected reward in the Markov TD model of Montague et al. (1996; Schultz et al., 1997) (left) and the partially observable semi-Markov model presented here (right). In the Markov case, sharp negative error occurs at precisely the time reward was expected. In the semi-Markov case, as time passes without reward, the system infers that an unsignaled state transition must have taken place, and this gradual process is accompanied by negative TD error, which is thus somewhat delayed and smeared-out with respect to the expected reward timing. The shaded bar underneath the error trace tracks the system's probability distribution over the hidden state; ITI: white, ISI: black. Mean ITI: 5 sec.; reward magnitude: 1; rectification threshhold: -0.1.; coefficient of variation of timing noise: 0.08; inference model's probability of unsignaled transition: 2%.

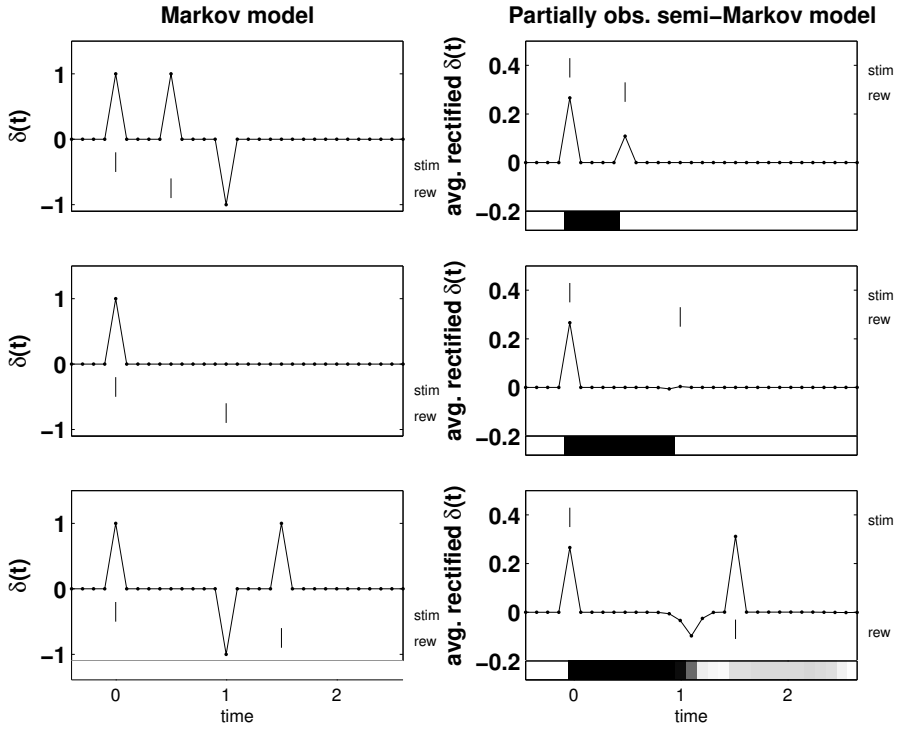

Figure 4.17: Dopamine responses on the task of Hollerman and Schultz (1998), simulated according to the Markov tapped-delay line TD model of Montague et al. (1996; Schultz et al., 1997) and the semi-Markov model presented here. Top: Early reward is followed by spurious inhibition in the Markov model, but because the reward triggers a transition in the semi-Markov model's inferred state (shaded bar, below), it does not show the inhibition. The models perform equivalently for on-time (middle) and late (bottom) rewards. ITI: white, ISI: black. Mean ITI: 5 sec.; reward magnitude: 1; rectification threshhold: -0.1.; coefficient of variation of timing noise: 0.08; inference model's probability of unsignaled transition: 2%.

task. The predicted results contrast in the case of early reward. After an early reward, the tapped delay line model predicts inhibition at the time reward had originally been expected, as discussed in Section 4.2.3. Hollerman and Schultz (1998) reported no such inhibition. In accord with these results, the semi-Markov model instead predicts no inhibition at the time the reward was originally expected. As shown by the colored bar underneath the trace, this is because the model assumes that the early reward has signaled an early transition into the ITI state, where no further reward is expected. While the inference model judges such an event unlikely, it is a better explanation for the observed data than any other path through the state space.

4.4.6 Discussion: State inference and variation in event timing

The simulations in the previous section show a number of cases in which the present model is an improvement on its predecessors, but they also reveal a couple of limitations of the model. Here I discuss these issues in light of the available data.

The Fiorillo and Schultz (2001) results, reproduced in Figure 4.2 (bottom) are particularly important to the present model because they are the only data for which dopaminergic responses have been broken down according to the length of the interval preceding them. As we have seen in a number of circumstances, the model predicts a characteristic pattern of responding in this case (Figure 4.15). The Fiorillo and Schultz (2001) results bear out the model's general prediction that the dopamine response should be graded according to the preceding delay, but they are at best equivocal on the more specific prediction that the response should become inhibitory at longer delays.

As discussed in Section 4.1.1, the single neuron for which data from this task are available (Figure 4.2, bottom) appears to show a pattern of excitation followed by inhibition in response to the reward. Thus whether the response at longer delays appears inhibitory depends to some extent on which phase of the response one considers. Fiorillo and Schultz (2001) computed group averages over 20 neurons using spike counts taken during the early phase of the response, and found (in accord with what can be seen in the single neuron) that the excitatory response wanes toward, but not below, baseline. However, for the neuron shown in Figure 4.2, it seems as though the subsequent inhibition persists somewhat even as the excitation wanes at long delays, so that the net response at long delays might indeed be weakly inhibitory, in accord with the model. That said, the data are quite sparse and difficult to judge by eye, the inhibitory phase is weak, particularly at long delays, and the model suggests no reason why the response should have a burst-pause temporal structure. Another factor may be the impact of time measurement error on the signal (Figure 4.15, bottom left), which attenuates the predicted inhibition at longer delays and may help explain why it is difficult to discern in the data.

In all, the model provides only a somewhat strained explanation for the Fiorillo and Schultz (2001) results. This is one experiment whose results are somewhat more naturally explained in the limiting version of the model in which states are subdivided and values defined in a Markov manner (see Section 4.3.8). In this limit, TD error to a reward would depend on the probability of reward at a particular delay following the stimulus, given that no reward had yet been received. The error would wane toward, but not below, zero at longer delays.

As already mentioned, the data provide more resounding support for the semi-Markov model's account (Figure 4.17) of the results of Hollerman and Schultz (1998). It may be unclear why the theory predicts different results in this case versus the Fiorillo and Schultz (2001) experiment, both of which involve early and late errors. In particular, given the result on the Hollerman and Schultz (1998) experiment, one might wonder why the model does not similarly infer an unsignaled state transition and produce dopaminergic inhibition *prior* to the arrival of a particularly late reward in the simulation of the Fiorillo and Schultz (2001). The answer is that the inference models are presumed to be different in the two cases, reflecting the animals' different experience with reward timing. In the Hollerman and Schultz (1998) experiment, reward was always delivered one second after the leverpress; the inference model reflects this duration as a normal distribution centered around one second. The stimulus-reward delay in the Fiorillo and Schultz (2001) experiment was uniform in the range from one to three seconds. Because the inference model used in the simulations reflects that distribution, it assigns a higher probability to long delays, and does not infer an unsignaled transition when they occur.

Another detail of the results presented here that might be confusing is the progression of belief states after

late reward in the simulation of the Hollerman and Schultz (1998) task (Figure 4.17, bottom right). Before the reward finally arrives, the shaded bar has become almost entirely white, corresponding to an inference that the task is in the ITI state. However, after the reward, it becomes more grey, indicating greater uncertainty between the ISI and ITI states. This may seem odd, since it may seem that reward delivery should only convince the system more forcefully that the task is now in the intertrial interval. The odd behavior is seen because there are multiple anomalous paths through the state space that predict the same anomalous observations. The data could be explained by an unusually long stay in the ISI state, followed by reward on transition to the ITI state. Alternatively, a stay of normal duration in the ISI state could have been followed by an unrewarded transition into the ITI state, where a short stay was then terminated by reward on a transition back into the ISI state. (Recall that such a misplaced reward has a low but nonzero probability.) It is this alternative explanation for the events that leads to increased state uncertainty after the reward, which resolves itself one second later when the ISI should time out but reward is not received.

The data also support the model's prediction (Figure 4.16) that, despite being often referred to as "well-timed," dopaminergic inhibition to missed reward is both delayed and smeared out compared to excitatory responses. (Exactly how delayed and how smeared out depends on the balance between various sorts of uncertainty in the inference model; e.g. how likely is an interstimulus interval of 1.1 seconds compared to an unrewarded transition.) Such an effect is visible to the naked eye in Figure 2.2 on page 16, and aggregate data confirm that dopaminergic pauses indeed have a higher latency than bursts (99 ms on average versus 50–110 ms) and a longer duration (401 ms average versus <200 ms; Hollerman and Schultz, 1998; Schultz, 1998).

For Figure 4.16 it was necessary to assume that the inference model's uncertainty about the inter-stimulus interval was significantly less extreme than suggested by behavioral timing experiments (the coefficient of variation was 0.08 in the simulation versus 0.3–0.5 according to Gibbon, 1977). Uncertainty at the larger levels produced negative error that was both much shallower and unreasonably prolonged compared to the dopamine recordings shown in, e.g., Figure 2.2 on page 16. Perhaps monkeys have more accurate timing processes than pigeons, or perhaps because of the short duration, the monkeys are operating in a pre-scalar accurate timing regime (see Gibbon, 1977). This finding also relates to a somewhat infelicitous issue about the discretization of the error signal in the present model. Event-triggered error is instantaneous and its magnitude is thus invariant to the granularity of the time discretization of the inference process and error signal, but the magnitude of negative error due to inferred state transitions depends on how many discretization "buckets" it is shared between. As the timestep between error points Δt becomes smaller (this is also equivalent to increasing the delay at the same discretization level, since the interval distributions are assumed to be scalar), the inferred state transition when a reward is omitted occurs over more discrete points. Hence, the same amount of negative error is divided between them, reducing the magnitude of error at each point. Thus the model plausibly predicts that as the stimulus-reward delay increases, negative error for a missed reward becomes progressively broader and shallower. However, it is unclear at a given delay what discretization level should be used or how the magnitude of negative error at each discrete point relates to the level of dopaminergic spiking. These issues would be better resolved with a continuous error signal and an account, at the level of spiking, of how the instantaneous error level gives rise to dopaminergic activity.

In this section we have discussed a number of issues with the model that are suggested by the simulations in the previous section. The general finding is that the new model provides a more faithful account of a number of dopaminergic experiments compared to its predecessors — in particular, we have now accounted for all the data in Figures 4.1 and 4.2. We have also identified weaknesses in a couple of areas.

4.4.7 Results: Tonic responding

In this section and the following one, I discuss issues of tonic responding in the semi-Markov model. In this section, I present simulation results about a new class of tonic responses predicted by the present model. In the next section I discuss how these results relate to experimental results, and I also revisit the predictions from Chapter 3 about tonic dopamine activity in light of the new model.

The partially observable semi-Markov model predicts a category of tonic activity not considered in previous models: activity related to the ongoing possibility of an unsignaled state transition, that is, tonic activity related to *uncertainty* in the state. Consider a trace conditioning experiment with different levels of partial

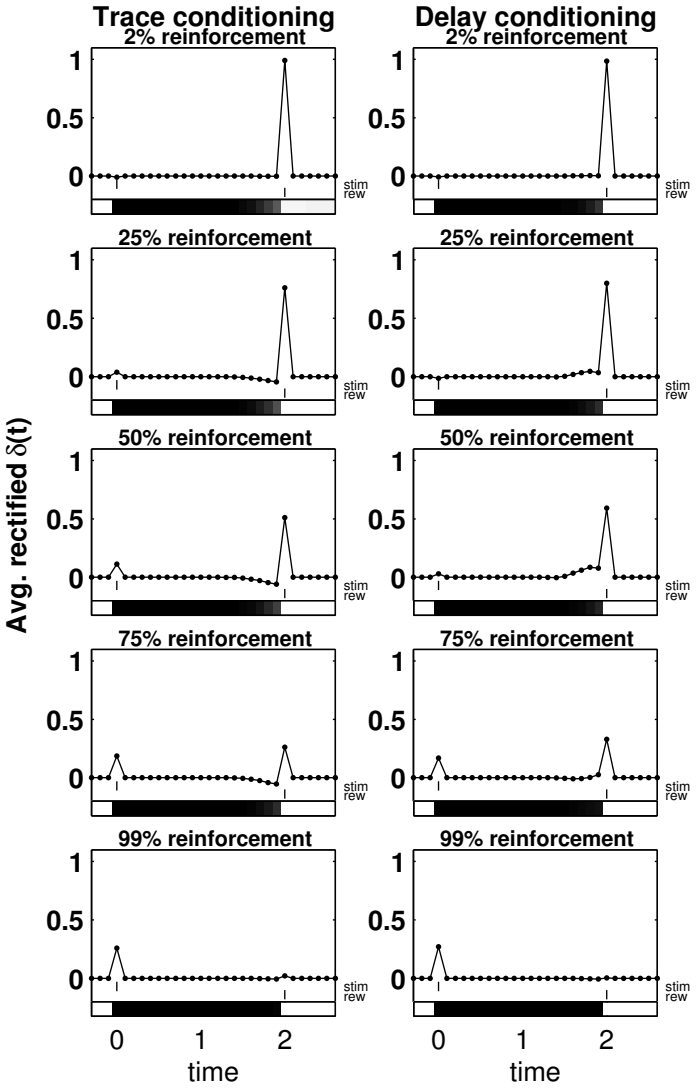

Figure 4.18: Semi-Markov model of average responding to rewarded trials in a partial reinforcement experiment, comparing trace and delay conditioning versions of the task. Left: In the trace conditioning version of the experiment, negative tonic error is seen preceding reward, strongest for 50% reinforcement. Right: In the delay conditioning version of the experiment (in which the persistent stimulus is treated as two punctate stimuli as described in main text), positive tonic error is seen preceding reward, similar to that reported by Fiorillo et al. (2003). Traces are resampled and averaged to simulate the effect of variability in time measurement. ITI: white, both ISI states: black. Mean ITI: 10 sec.; reward magnitude: 1; rectification threshhold: -0.1.; coefficient of variation of timing noise: 0.1.

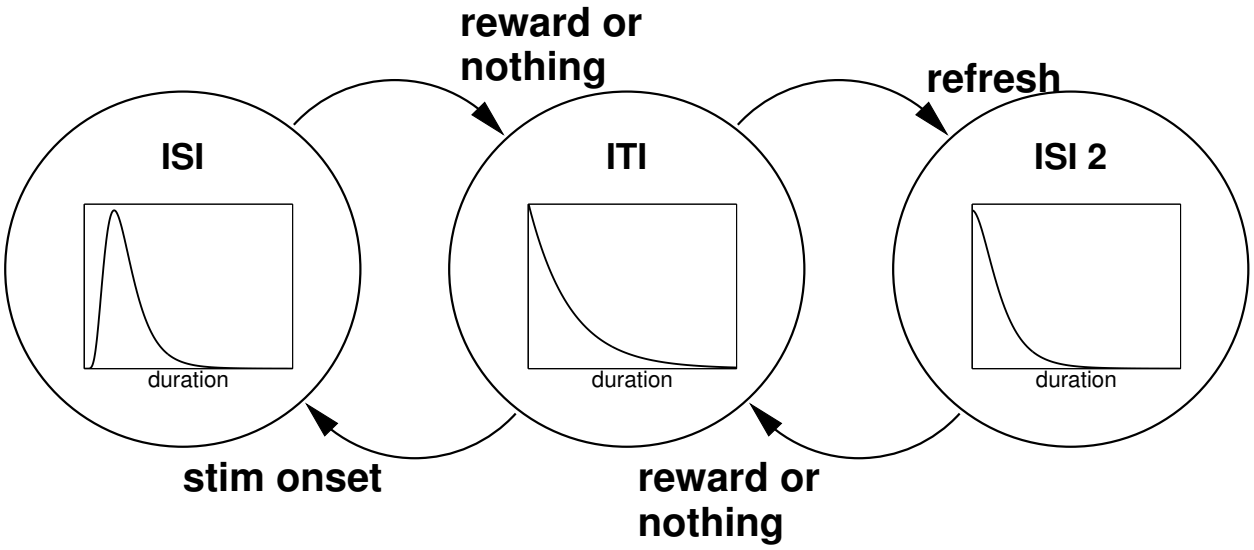

Figure 4.19: Semi-Markov inference model for a delay conditioning partial reinforcement task in which the persistent stimulus is treated as two events: "onset" and "refresh."

reinforcement. A stimulus is followed, two seconds later, by either reward or nothing, with systematically manipulated probabilities. Because of asymptotic temporal uncertainty, the animal does not know exactly when the trace delay has elapsed — and this uncertainty is more dramatic in cases where the animal is less certain that the interval termination will be signaled by reinforcement. Thus, as the fraction of reinforced trials declines, the animal is increasingly uncertain as to whether an unsignaled state transition has taken place. The inference of such an unsignaled transition produces negative TD error, more so as the fraction of reinforced trials *increases*. The product of these two opposing effects, shown in Figure 4.18 (left), is small, ramping tonic negative TD error during the trace delay, strongest for reward probabilities around 50%. (Differences in the level of inhibition for different levels of partial reinforcement, already subtle, are further obscured by the rectification of negative error.) In this figure, I have included the effect of measurement noise in timing according to the method described in Section 4.5.1.[5]

As will be discussed more fully in the next section, this prediction seems related an experiment conducted by Fiorillo et al. (2003). However, that task was delay conditioning rather than trace conditioning — that is, the cue stayed illuminated during the entire two-second period until the reward was either delivered, or not. The reported result had the same form, but the opposite direction, as the prediction reported above: ramping tonic *excitation* of dopamine neurons during the trace delay, strongest for 50% reward probability.

The semi-Markov model concerns point rather than extended events; it needs to be extended to accommodate delay conditioning experiments like that of Fiorillo et al. (2003). The simplest such accommodation would be to assume that a temporally extended stimulus is produced by a semi-Markov state that becomes active on stimulus onset and stays active so long as the stimulus persists. In the present experiment, this would eliminate any uncertainty as to whether there had been an unrewarded state transition, and (because the TD error signal is weighted by the chance of transition, see Equation 4.3), entirely eliminate any tonic

[5]There is an important distinction between *temporal uncertainty* and *interval measurement noise*. The figures in Section 4.4.5 that illustrated the behavior of the partially observable model all included *temporal uncertainty* due to the fact that the inference model's dwell time distributions were not perfectly sharply peaked; the inference model of Figure 4.16 for instance assigned some probability to interstimulus intervals of 1.1 and 0.9 as well as the programmed 1.0 seconds. But the plots themselves were made only from traces in which the interstimulus interval was actually 1.0 seconds, i.e. temporal noise was not added. As discussed in Section 4.3.7, I assume that asymptotic interval uncertainty in the inference model reflects the form of *actual variability in the animal's timing measurements*, so that the animal does indeed sometimes measure the objectively constant 1.0 second interval as lasting 0.9 or 1.1 seconds. For simplicity, I have omitted this effect from the figures of the partially observable model thus far (though it was included in simulations using its fully observable limit, in Figures 4.13 and 4.15, bottom right).

TD error during the period when the stimulus is active. The Fiorillo et al. (2003) data (if they are to be understood in terms of state uncertainty), argue against this approach. On the right side of Figure 4.18, I have modeled the delay conditioning version of the experiment under a different, and somewhat more ad hoc, model of extended stimuli in the semi-Markov framework. In particular, I assume that animals time the the reward interval based only on the stimulus onset, and the stimulus' persistence serves to occasionally *cancel* mounting uncertainty that there may have been an unrewarded state transition. (The intuition behind this approach is that an animal might ignore the persistent stimulus but check back to ensure that it is still active when it decides the interval should have timed out.) Specifically, I model the delay stimulus as a punctate onset event followed by a second punctate "refresh" event that occurs if the inferred probability of an unsignaled transition exceeds some threshhold but the persistent stimulus is still active. In the inference model I assumed, which is illustrated in Figure 4.19, the refresh event occurs on a transition into a third state ("ISI 2") that predicts imminent reinforcement. Such an event is assumed to occur only after a transition into the ITI state. Note that this inference model does not exactly correspond to the actual contingencies of the delivery of the events, which are determined by the inference process itself (since the refresh event is triggered by the inference of a null transition into the ITI state). In individual trials where the delay interval is measured as longer-than-average, this sort of arrangement tends to produce dopaminergic inhibition (due to the presumption of an unrewarded transition), followed by excitation (due to the signal that the stimulus is still persistent); averaged over trials with the negative error partly rectified, the overall effect is tonic excitation, roughly in line with the experiment.

In the following section I discuss further issues surrounding these simulations, and also revisit the separate tonic effects predicted in Chapter 3 in light of the new model.

4.4.8 Discussion: Tonic responding

The results of Fiorillo et al. (2003) are notable in that they represent the only report of tonic responses from recordings of dopamine neurons spiking (as opposed to the neurochemical recordings of dopamine concentrations discussed in the previous chapter). Some results from the experiment are reproduced in Figure 4.20. Due to the similarities between the tonic responding shown on the left subfigure and the simulations of a trace conditioning version of the task shown on the left side of Figure 4.18 (and in spite of the opposing directions of effect), it is tempting to interpret the experimental data in terms of state uncertainty. However, as described in the previous section, it requires a certain degree of imagination to find a model in which state uncertainty has any effect in a delay conditioning experiment like Fiorillo et al.'s.

Under firmer assumptions, the present model predicts graded tonic *inhibition* in a trace conditioning version of the task (Figure 4.18, left). The empirical status of this prediction is unsettled: Two trace conditioning versions of the experiment, one piloted by Fiorillo and another in a separate lab, found no tonic *excitation* of the sort shown in Figure 4.20, left, though it is unclear whether either analysis searched for the possibility, predicted here, of subtler tonic inhibition (Fiorillo et al., 2003, footnote 14; P. Dayan, personal communication, 2002).

The simulations of the delay conditioning version of the task shown in Figure 4.18, right, achieve "tonic" responding in the average by trading off phasic responding between the "onset" and "refresh" stimuli, their relative timing being somewhat smeared out. (This is why responding to the CS onset is depressed in some of the plots; better balancing of these two responses is a parameter-tuning issue.) The data can also be explained, and arguably more naturally, in a model with a Markov valuation scheme, such as the present model when the states are subdivided as described in Section 4.3.8. The explanation is due to Dayan (personal communication, 2002). The idea is that the tonic activity results from phasic positive and negative error from rewarded and nonrewarded trials propagating backward through successive Markov states during the intertrial interval, producing a net positive sustained response due to rectified averaging. This explains one feature of the empirical data that the present model could capture only accidentally (i.e., with careful tuning) if at all: the smooth continuity between the tonic and phasic portions of the response when the response is averaged over all trials as in Figure 4.20, left. The simulations in Figure 4.18 instead show results only for rewarded trials. This does not change the tonic portion of the response, but it was done to allow comparison with the phasic responses to reward, which were reported separately by Fiorillo et al., 2003 and are reproduced in Figure 4.20, right. (The modeled and empirical phasic responses are comparable.)

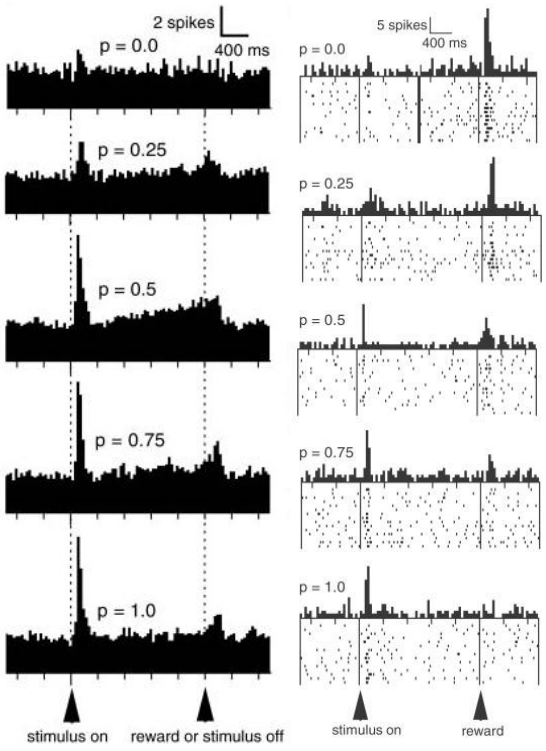

Figure 4.20: Dopamine recordings from a delay conditioning task with partial reinforcement, from Fiorillo et al. (2003). Left: Tonic responding over a population of dopamine neurons, averaged over both rewarded and unrewarded trials. Right: Responses from a single neuron averaged over rewarded trials only to highlight the phasic response to reward.

I now revisit the predictions of Chapter 3 with regard to tonic activity in the dopamine system. The tonic responses in the new model that have been discussed in this section are related to state inference and state uncertainty, but the ones discussed in Chapter 3 have no relation to these and are best discussed in the fully observable limit. Recall that the Markov and semi-Markov valuation schemes represent different ends of a spectrum of models based on the current one (Section 4.3.8), and that their differences play out as a difference in how the Markov and semi-Markov value learning algorithms of Equations 2.12 on page 11 and 4.1 on page 87 apportion the cost of delays. In both cases, this cost is $-\rho$ per timestep. In the Markov algorithm, this factor is subtracted from the error signal at every timestep, which I associated in Chapter 3 with a tonic depression of the signal. But in the semi-Markov version, the $-\rho$ factors are batched up and delivered phasically whenever there is a state transition; TD error is strictly zero whenever there is not a state transition.

Thus the "tonic" effects of the previous chapter are not, under the semi-Markov valuation scheme, tonic at all. The error signal contains the same positive and negative factors — in particular, it still contains a slowly changing component that suppresses the dopamine response and is proportional to the average reward per timestep — but the new model predicts that the same factor affects the dopamine signal with a different, and phasic, temporal pattern. The error will be subtracted from phasic dopaminergic activity that occurs on state transitions that are occasionally triggered by events in the environment. This difference should have no effect on predictions about slow neurochemical measurements such as microdialysis, since these methods are too slow to pick up any particular temporal pattern of responding. So the most important result from that chapter, the explanation for microdialysis measurements of enhanced dopaminergic activity in aversive situations, remains the same. It is also unclear how the difference between the predicted response patterns would play out at the level of the fast-scan voltammetry measurements depicted in Figure 3.6 on page 50, since it is unknown what effect tonic responding would have on baseline dopamine concentrations. However, since the semi-Markov model predicts the suppressive effects of ρ should be visible only in the height of phasic responses and not tonically between them, it might help explain why no effect of the reward rate was visible in the baseline dopamine concentrations in the experiments shown in Figure 3.6 on page 50 (left and center). On the whole, it seems that the extant data do not clearly favor either model, but the contrasting predictions predicted here and in Chapter 3, together with the analysis in this chapter about how these two patterns of responding follow from the same model as different extreme limits, provide a theoretical framework to guide future experimentation.

4.5 Behavior: Timescale invariance in the model

In this section I study how the modeling in this chapter relates to the behavioral data on timing and timescale invariance that was reviewed in Section 4.1.2. First I discuss some issues related to the account (from Section 4.3.7) of scalar variability in the inference system's time measurements. Next I turn to data on timescale invariant acquisition in classical conditioning, and study the speed of value learning in the semi-Markov TD models and also in the hypothesized model learning components.

4.5.1 Scalar variability in timed responses

As discussed in Section 4.1.2, the trial-to-trial variability in timed animal behaviors has a specific, scalar form: the standard deviation of a timed behavior scales linearly with its mean. Here I review how the present model copes with these data, and then discuss how this purely phenomenological account of the data relates to a notable normative attempt to explain this noise in terms of uncertainty.

As described in Section 4.3.7, multiplicative noise can be added to the time measurement processes in the present model by introducing variability in the timestep Δt that clocks the state inference. Since state inference works by counting timesteps between events, such variability will cause the progression of belief states following some event to change faster or slower (as a function of real time) from trial to trial. This variability will be reflected in the system's moment-to-moment value estimates (which are a function of the belief state), and in turn in any timed behaviors that are triggered by particular belief states or value predictions. For instance, if a stimulus is followed an average of ten timesteps later by reward, then the tenth belief state following the stimulus (or its value) might trigger some conditioned response. The distribution

in real time of this CR, relative to the stimulus, will just be ten times the distribution of timesteps Δt. Thus, in all cases, the variability in the timings of such responses will be scalar, because the response times will be distributed as multiples of the distribution of Δt.[6] So the present model accords with experimentally observed scalar variability in timed behaviors (Gibbon, 1977).

This is a phenomenological (as opposed to normative) account of scalar timing, since I have not explained why the timer should behave this way. I now relate the approach to a more normatively grounded approach to the problem, which also has correlates in the present model. Kakade and Dayan (2000, 2002a) explain scalar variability effects as reflecting *asymptotic uncertainty* in animals' estimates of underlying parameters of the world. In the case of a timing study like the peak procedure, the parameter animals are trying to estimate is a temporal interval. The notion — derived from an online statistical estimation procedure known as Kalman filtering — is that even if trial-to-trial measurements of the interval are noiseless and identical, and even if an animal averages a very large number of them to estimate the true interval, the estimate will retain some fixed, asymptotic level of uncertainty if the animal believes that the interval being estimated can *change* from trial to trial. In particular, if the animal assumes a model of change under which the intervals wander in a random walk according to multiplicative (rather than the standard additive) noise, then the asymptotic variability in its interval estimates will be scalar. If, moreover, an animal's behavior on a particular trial reflects a *sample* from its uncertain distribution of intervals, then scalar trial-to-trial variability will result. For instance, Scalar Expectancy Theory (Gibbon, 1977) envisions that at the beginning of each trial in some task involving timing, an animal takes a fresh sample from its (scalar) distribution of expected intervals to reinforcement. The animal treats this sample as its expected time of reinforcement on the current trial, and its response profile varies proportionately.

The present theory directly incorporates this idea of asymptotic uncertainty in beliefs about intervals, in that state dwell-time estimates in the inference models include scalar asymptotic uncertainty; as in the Kakade and Dayan (2000, 2002a) model, this could be justified by a judicious choice of noise model. But such uncertainty by itself doesn't give rise to any trial-to-trial behavioral variability (and thus cannot be connected with the response variability studied by Gibbon, 1977) without some method for sampling over the interval estimates. Directly sampling from the interval distributions doesn't make sense in the present theory, since the timed behaviors are supposed to result from the belief states and value estimates, but the state inference process that produces them necessarily makes use of the full interval distributions. However, the noise added to Δt can be viewed as playing a similar role to sampling over the interval distribution in other theories. The perhaps more natural but less normatively grounded alternative viewpoint is to take this timing noise as a fundamental, unexplained limitation in the timer, and assume that asymptotic uncertainty in interval estimates *models the form* of this measurement noise (rather than modeling change in the world).

4.5.2 Timescale invariance in acquisition

I now examine how the model can be related to timescale invariant acquisition of the sort studied by Gallistel and Gibbon (2000). This is a preliminary and incomplete account because acquisition speed in the present theory is controlled not only by TD learning of values, but also by hypothesized interacting processes for learning an inference model (see e.g. Figure 4.5 and the expectations in Equation 4.3, which are computed from modeled quantities). I have not here provided an explicit account of model learning whose scaling properties could be analyzed. Because of this, I analyze semi-Markov value learning in the fully observable limit (to exclude model learning effects), and show that it is timescale invariant. I sketch some of the issues involved in developing a fleshed-out theory of the model fitting process so that value learning in the partially observable case would have analogous timescale invariance properties, at least approximately. Finally, I note that existing models of acquisition in classical conditioning bear some resemblance to the *model selection* stage in the present framework, which may help explain why they rely on rather different assumptions about the sorts of information being learned.

[6] What is the distribution of Δt? In Section 4.3.7, Δt was determined by dividing some basic timestep by a scale factor drawn from a lognormal or truncated normal distribution. (This method was chosen to ease reasoning about the distribution of timestep counts corresponding to some particular interval of real time, the dual of the question studied here of the distribution of real time intervals corresponding to some particular timestep count.) If we take the scale factors to be lognormally distributed, then Δt is conveniently also lognormally distributed, with the same coefficient of variation.

Consider the effect of halving the speed of events in a conditioning experiment, that is, doubling the inter-stimulus and intertrial intervals. Viewed as a semi-Markov process (Figure 4.6), the structure of events would not change at all — only the state dwell time distributions would rescale. Moreover, the asymptotic values of the two states (using the average-reward return) would not change. The Bellman equation relating the states' values is $V_{ISI} = r - \rho \cdot d_{ISI} + V_{ITI}$, where V_{ISI} and V_{ITI} are the values of the two states, d_{ISI} is the duration of the ISI state, and the reward rate $\rho = r/(d_{ISI} + d_{ITI})$. Slowing down the timescale of events by a factor of two would halve ρ but double d_{ISI}, and the values would remain unchanged.

Similarly, the fully observable algorithm for learning these values, Equation 4.1, follows exactly the same trial-by-trial steps when time is rescaled. That is, consider two conditioning experiments that consist of the same events (stimuli and rewards alternating), but in one case the delays between them are doubled. Since learning is triggered by events, the system in the second case would take exactly the same learning steps, only half as often, and the learned values in either case would be the same at each step. The proof is easy: Substituting the doubled delays $2d_{k'}$ for $d_{k'}$ in Equation 4.2 halves each trial's reward rate estimate ρ_k. But the value update δ_k remains the same because ρ_k is multiplied by the doubled delay, $2d_k$, in Equation 4.1.

Thus the fully observable limit of the present TD algorithm shows timescale invariant acquisition. In a partially observable situation, similar timescale invariance properties can only be approximately true of the full model, Equation 4.3, and only if the processes for learning the inference model are themselves timescale invariant. In the partially observable case, learning can be triggered by events as above, but it can also be triggered by the inference that an unsignaled state transition has taken place. These inferences take place at a timescale determined by the internal model, specifically by the function $\mathbf{D}$ specifying the distribution over dwell time durations in a state. For instance, when a reward is skipped, learning occurs smeared out over several timesteps as the model infers a state transition. For timescale invariance to hold, the batched amount of learning over the course of such an inferred state transition should be invariant to rescaling of the dwell time distributions. Due to the way these distributions and the steps of state inference are temporally discretized, such invariance does not exactly hold true in general, though deviations from it become minimal as the discretization timestep Δt is reduced. Further analysis, and perhaps further algorithmic development, is warranted on this point.

Value learning in the partially observable algorithm can also be timescale invariant only if the processes for learning the inference model it uses (i.e. the functions $\mathbf{D}$, $\mathbf{O}$, and $\mathbf{T}$) are themselves timescale invariant. That is, if the durations between the discrete events are rescaled, the functions should be the same after each event except with $\mathbf{D}$ rescaled. I have not laid out such a learning process here, though this constraint will clearly be an important goal when developing one. The scaling properties of EM algorithms for fitting hidden semi-Markov models have not, to my knowledge, been examined from this perspective. In general, one might hope that if the model learning update steps are event-triggered or gated by a quantity analogous to β that such an invariance property could be arranged, though we may here again encounter problems with learning triggered by an inferred state transition and also with selecting appropriate timescale-neutral priors for $\mathbf{D}$.

A final point is that timescale invariant learning as it has been discussed so far is only a necessary, but not a sufficient, condition for capturing the full pattern of acquisition data studied by Gallistel and Gibbon (2000). In particular, they show that acquisition times are not just timescale invariant but also proportional to the ratio T/I, where T is the inter-stimulus interval and I is the intertrial interval (Figure 2.7 on page 40). The specific dependency of acquisition times on these quantities could best be studied when the full model with all of its interacting stages is complete. (By itself, the value learning stage in the fully observable limit does not satisfy this dependence; it is easy to show that acquisition times in this simplified situation are instead proportional to $1 + T/I$, which is roughly constant given the ranges of T and I normally used experimentally.)

As reviewed in Section 2.3.5, a family of successful models of the acquisition data (Gallistel and Gibbon, 2000; Kakade and Dayan, 2000, 2002a) explains this result by assuming that acquisition is based on a particular statistical test. Specifically, acquisition is assumed to occur when the animal can confidently assert that the reward rate in the presence of the stimulus is higher than the reward rate in its absence. A mystery about those accounts is why the statistical decision is based on comparing simple co-occurrence rates between the stimulus or background and the reward. This seems a rather crude metric because it does not take into account the fact that the reward typically occurs at a specific, deterministic time relative to

the stimulus onset — a fact that animals clearly learn and that models of other aspects of animal learning (such as TD models) incorporate.

I suggest that this apparent conflict between these successful models of different aspects of learning might be explained by noting that the statistical test in the acquisition models sounds like a heuristic for answering a *model selection* question — are the rewards produced by a process with two states or one? (Indeed, it can be formally recast as one.) In the present framework, some way is needed to choose the number of states in the models that are then subject to parameter fitting and value learning, but proper Bayesian model selection in an online setting is probably intractable, given the full generality of the family of hidden semi-Markov models considered here. A reasonable shortcut might be determining the number of states using a simpler family of models involving only fully observable states and untimed, Poisson reward delivery; in this setting, model selection can be performed as in the acquisition models using tests on simple stimulus-reward co-occurrence rates. Thus, we could imagine that an acquisition model like that of Kakade and Dayan (2000, 2002a) serves the role of model selection in the present framework (Figure 4.5) and only when it decides that there are two states can subsequent (timescale invariant) model fitting and value learning stages learn the values that give rise to a conditioned response. This might explain both the T/I dependence in the context of the present model, and the seeming conflict between models of different phenomena that seem to rely on fundamentally different types of learning. On this view, simple co-occurrence rates are useful, but only in an initial model selection phase that occurs prior to more sophisticated sorts of learning.

To recap, in this section I demonstrated that acquisition in the fully observable version of the semi-Markov model is timescale invariant, thanks to the event-driven learning rule and its setting in the timescale-agnostic semi-Markov formalism. Of previous TD models, only that of Suri and Schultz (1998, 1999) could achieve timescale invariant learning, and that only given some modifications I introduced in Section 4.2.7. This result suggests that the semi-Markov formalism provides a useful setting for further study of timescale invariant learning phenomena, hopefully even when partial observability (which adds some complications, which I outlined above) is added to the mix. Finally, I suggested that the detailed dependence on task intervals of animal acquisition times in classical conditioning might be explained in the present framework by a model selection process, and indeed that existing acquisition models could serve that purpose more or less unmodified. Such a hybrid model might help resolve the puzzling conflict between different sorts of learning important to different models of classical conditioning.

4.6 Discussion and open issues

In this chapter, I have presented a dopaminergic model that extends previous TD accounts in two major directions. Previous models used representational schemes liked the tapped delay line that amount to a straight transcription of raw sensory events. The present model replaces these with a more active representational component based on the theory of partially observable processes, whose job is to model the world's underlying processes and to use sensory observations to infer the hidden state of those processes. In addition, the Markov value learning algorithm has been replaced with one based on semi-Markov processes. Because this formalism explicitly incorporates variation in the timing of events, it is useful for reasoning about the expected behavior of dopamine neurons when the timing of events varies, either due to programmed experimental variation or to measurement noise on the part of the animal.

The bulk of this chapter has concerned using this model to account for results about the activity of dopamine neurons in such situations and to make predictions about experiments and analyses that have not yet been done. Notably, the semi-Markov model suggests a testable pattern of covariance between the dopaminergic response to some reward-related event and the delay predicting it; this gives rise to a significantly different explanation from the standard one as to the origins of most familiar excitatory dopaminergic responses. The present model also suggests a broader functional framework for envisioning the interaction of different sorts of learning in the brain, and it also suggests a family of related TD approaches (including some novel and some previously studied) that emerge from the new model in various limits. Finally, and unlike its predecessor, the semi-Markov model has a timescale-invariant learning algorithm, at least in its fully observable limit, so the chapter has also discussed some examples of how the new setting provides an appropriate framework for studying various facets of timescale invariance in animal behavior.

4.6.1 Gaps in the present work

Several issues remain open, suggesting a number of avenues for future work. Here I discuss three theoretical gaps — the problems of model learning, action selection, and additivity — and one empirical one, the neural substrates of the model.

The most glaring oversight in the existing work is that I have not formally treated the problem of model learning. However, simply by recasting the representation learning problem as a model learning problem, the present theory represents an advance over existing TD theories of dopamine, since model fitting and model selection are well defined and well studied problems and are routinely addressed using standard Bayesian machinery such as the Baum-Welch algorithm (Baum et al., 1970) for fitting hidden Markov models. (Of course, these basic methods are not online, but incremental, online variations exist and have been used for behavioral modeling; Courville and Touretzky, 2001.) Analogous representational learning problems in previous TD models (having to do, for example, with learning when tapped delay lines should reset, see the discussion in Section 4.2.5) are not defined in such a way as to allow them to be handled with a similar degree of rigor; in fact, previous models have not viewed them as *learning* problems at all. The new view also suggests a connection with modern theories of cortical sensory function, which are also often based on Bayesian model fitting (e.g. Lewicki, 2002; Lewicki and Olshausen, 1999). Dopamine itself seems to also play a role in this sort of cortical sensory learning (Bao et al., 2001), which would be an interesting angle to explore in future modeling.

As suggested by the discussion of acquisition times, a more fleshed-out account of model learning could also be useful in future for modeling a number of classical conditioning phenomena. The present representational scheme for TD learning was inspired by and builds on Courville and Touretzky's (2001) hidden Markov model theory of how animals learn about and represent the temporal structure of events in classical conditioning; that paper demonstrates online hidden Markov model learning to learn models similar to those used by the present dopaminergic model, and accounts for a number of classical conditioning effects that have not been modeled elsewhere. Meanwhile, representational learning in classical conditioning is a very important problem that has so far received only fairly informal and poorly motivated theoretical treatment. Much of this work concerns "configural learning" (see Section 2.3.1 and the review of Pearce, 1994), for modeling experiments such as negative patterning ("XOR") in which animals must learn that a configuration of stimuli has a different reward association than its elements. Hidden states in the present model play the same role as stimulus configurations in those earlier theories, in that they can represent correlations between groups of stimuli, and Bayesian model selection and model fitting provide a well-founded account for understanding how such representations should develop in response to learning. Advancing these ideas is a major focus of current work in which I am involved (Courville et al., 2003).

Also missing from the present model is any discussion of the problem of action selection, which is of course what all of this value function learning is supposed to support. This is an extremely important gap and a major target for future development of the model. The main computational complication with regard to action selection in a partially observable context such as the present model is that action choice can (and optimal action choice often must) depend, in a continuous manner, on the belief state. That is, if I believe I am in state A with 90% probability and state B with 10% probability, I may choose to take action 1; if I am equally likely to be in either state, I may take action 2, and if I am more likely to be in state B, I may take yet a third action. This is a powerful (but correspondingly computationally costly) feature in that it allows the degree of uncertainty to play a role in decision-making. For instance, algorithms based on POMDPs can learn automatically that when state uncertainty is high, it can be advantageous in terms of long-term payoff to take information-gathering actions. But since different actions, potentially occurring anywhere in the continuous belief space, lead to different future values, the value function itself must be defined over the continuous belief space rather than the discrete space of hidden states. Without decisions, expected future value behaves in an orderly manner and can be exactly captured, as I have in this chapter, as linear in the hidden states. Properly accounting for decision-making with uncertainty will require the use of a more sophisticated function approximation method to approximate the continuous structure of the value function. There has been some impressive work combining physiology, behavior and modeling to study how animals make decisions under a very limited sort of state uncertainty (Gold and Shadlen, 2002); the present account, when augmented to include decision-making, potentially offers a more general framework that could encompass these ideas.

Also, I have not spoken in any detail about the neural substrates of the model. Statistical modeling and Bayesian inference are considerably more abstract than tapped delay lines, which, for all their faults, admit of a quite straightforward neuronal implementation involving waves of activity propagating through a series of neurons (Montague et al., 1996). But as already mentioned, the hidden semi-Markov models discussed here are really only a temporal generalization of the sorts of latent variable generative models that have recently become quite popular as theories of processing in early sensory cortices (Lewicki, 2002; Lewicki and Olshausen, 1999); in this context there has been some effort toward developing neurophysiologically realistic implementations for these algorithms. When comparing the neural substrates of the present model against those envisioned for previous TD models, the important high-level difference is that the new model underlines the importance of the state representation s, and of the rather sophisticated learning and computation that underly it. Broadly, the state inference and world modeling components of the theory could possibly be identified with either the striatum or its afferent cortices; one interesting upshot of the present model is that sustained anticipatory firing of neurons in striatum and prefrontal cortex could reflect the representation of semi-Markov states (which are active in the model for extended intervals between events), rather than value predictions as has often previously been assumed (Suri and Schultz, 2001). This account improves on the value-based account in that it would explain why sustained anticipatory firing is seen preceding events other than rewards. (For a value-based take on the same phenomenon, see Montague and Baldwin, 2003). I return to this point from an experimental viewpoint in the next chapter.

As previously mentioned, also missing from the present model is any additivity between states and any higher-order semi-Markov dynamics. A generative model with a factored state combining multiple simultaneously active elements is necessary to give a convincing account of some conditioning phenomena such as conditioned inhibition; for this reason I have not modeled the dopaminergic results on conditioned inhibition here. To see why higher-order dynamics would be useful, consider events o and o', separated in time by some constant interval, with a third event o'' occurring between them at some random time. This can't be expressed in the semi-Markov formalism — the best that can be done is for o to be followed by o'' after some random interval, and o'' by o' after a second random interval, losing the tight temporal relationship between o and o'. Behaviorally, this would mean that animals would be unable to predict the timing of o' accurately on the basis of its temporal relationship with o, since they would instead be forced to predict its timing based on its muddier relationship with o''. This seems a highly counterintuitive prediction and unlikely to fully hold up empirically; in fact, in somewhat related experiments (see e.g. the "gap experiment" of Roberts, 1981, and the review by Staddon and Cerutti, 2003), stimuli are graded in the extent to which they can serve as time markers or "reset" an ongoing timing process, with many factors seemingly controlling these effects. The present theory could be extended to include these characteristics, though the combination of a factored state with partial observability makes tractable inference difficult due to "explaining away" effects (see e.g. Ghahramani and Jordan, 1997).

4.6.2 The dopamine response magnitude

One potential problem with the present account involves the magnitude it predicts for dopaminergic responses. As can be seen by examining the axes of almost any figure in this chapter, the TD errors modeled in this chapter are lower in absolute terms than those in previous TD models. Average rectified TD error to unexpected rewards or reward-predicting stimuli in the present simulations ranged between about 0.1 for early reward in the simulation of the Hollerman and Schultz (1998) experiment to about 0.4 in the free reward experiment, and approach 1 only for reward delivered after a habitually unreinforced stimulus in a partial reinforcement task. TD errors in the original Markov model in all of these cases would be near 1, the reward magnitude. The reason for this discrepancy is that, in the new model, there is zero asymptotic *TD error* in the average over trials to almost all of the events usually associated with dopamine activity — that is, these events would produce negative or positive errors from trial to trial that exactly balance out in the average. The model predicts excitation in the average *firing rate* measured to these events, because the rectification of negative error in the dopamine signal biases the mean response toward excitation. (See, e.g., Figure 4.10 for one example of this.) However, the magnitude of this response is small, since the negative errors are only partially rectified and also since the trial-to-trial errors being averaged are themselves small, since they result from small trial-to-trial differences in delay costs for events whose occurrence can be predicted but

whose exact timing cannot be.

Obviously, the absolute response scale between models is arbitrary; the question is whether *among* the responses simulated using the semi-Markov model across different experiments, some are vanishingly small relative to others. Viewed from this perspective, the response scale does not seem quite as troubling, since the canonical sorts of dopamine responses are all, on the present model, more or less attenuated by the same factors. Most importantly, the response to free reward, which is normally taken as the benchmark "totally unpredicted" reward against which other dopamine response magnitudes are compared, is itself, on the present account, only due to rectified averaging of zero-mean responses that measure the cost of an unpredictable delay to an *otherwise predicted* reward. The response to free reward is stronger than other responses in the present simulations because its timing is more variable (the magnitude of the rectified average is driven by the range of possible event timings as a fraction of the average trial length). This means that the relatively large response to free reward in the present simulations is due to the use of a true Poisson schedule of reward delivery. Intertrial interval scheduling in actual experiments is not so variable, usually involving a variable tail added to a fixed delay. For instance, in the Hollerman and Schultz (1998) experiment, intertrial intervals were chosen randomly to be between 4 and 6 seconds, out of a total trial cycle lasting around ten seconds. Such constrained variability, which seems to be typical, would tend to bring the average response to free reward down toward the level of the other responses simulated in this chapter.

Thus the only outliers with respect to the relative magnitude of dopaminergic bursts simulated here are the response to rewards occurring after a stimulus that predicts 0% partial reinforcement and to rewards on free reward trials selected for very short ISIs. In these cases, rather uniquely, the model predicts strong, positive asymptotic TD error that is not the result of a rectified average. The analysis involving free reward trials has never been performed. Data on partial reinforcement have been published (Fiorillo et al., 2003), though they are sparse. For the single neuron depicted there, the reward response for 0% partial reinforcement does indeed look unusually robust, though it does not seem to exceed the CS response to the extent predicted by Figure 4.18. (For the population of studied neurons, the Fiorillo et al. paper reports only normalized CS and reward response magnitudes, so we cannot make any magnitude judgments about the population.) One explanation for the seeming discrepancy is that in a version of the present model that incorporated additivity between multiple semi-Markov processes, rewards following stimuli known to predict no reward are likely to be partially credited to a background reward delivery process, which would attenuate the modeled response. Also, there is a suggestion from very recent experiments comparing the responses to different magnitudes of juice (Tobler et al. 2002, reviewed in Section 2.2.1) that there may be some sort of gain control of the dopaminergic response. Depending on its details, which are of course totally unclear, such a mechanism could compensate for variation in response scale under a semi-Markov model.

4.6.3 Implications of the internal model

Suri (2001) and Dayan (2002) have produced TD theories of the dopamine system that, like the present one, also include an internal world model. Both focus on the role of the world model in facilitating action planning in a fully observable context; the present work is novel among dopaminergic models for instead exploring a *representational* role for world modeling. These functions are by no means mutually exclusive; indeed, the "successor representation" algorithm used by Dayan (2002) for planning in a dopaminergic model had been originally conceived as a method for representation learning (Dayan, 1993).

Suri (2001) also uses iteration in the internal model to recover future value estimates, limiting the reliance on TD-style sample backups. This points to an interesting question about the motivation for the present theory (as well as the other model-based theories). Given that we assume a full world model, why should the brain not solve it directly for the values (and policies) using something like value iteration? Why, instead does the brain seem to use sample backups — i.e., TD — to recover the values, when those samples don't provide any information that isn't, in principle, already encoded in the model? One possibility is that if the world is nonstationary, and the model thus always changing, continually relearning values with value iteration is infeasible while incrementally updating values with TD can track reasonably up-to-date estimates. (A cleverer approach to this problem from the artificial reinforcement learning literature is prioritized sweeping; Moore and Atkeson, 1993.)

Models with the basic structure of the one used here — TD in the belief space — are not uncommon in

reinforcement learning (dating back to Chrisman, 1992); one motivation for work of this sort seems to be that if the model is incorrect, or certain parts of it not well established, then TD over real samples of actual trajectories in explored parts of the state space seems less prone to corruption than estimates derived by solving the model directly. This is in line with the general philosophy of the present model, in that the idea is that the model constructs a representation in which TD can operate. The actual details of the learning rule used here do not really support such a separation, however: since there is so much conditioning on the model in the TD update rule itself, it's unclear that it would perform usefully given a bad model. The Markov limit of the present model is better in this respect, since in this model it is possible to directly run TD over the inferred belief states, with no further use of the model in the update rule.

4.6.4 Implications for experiment and comparison of Markov and semi-Markov models

The modeling in this chapter and the previous one contrasted the Markov and semi-Markov valuation schemes for TD learning, which predict different dopaminergic behavior, and which occur as different limits of the present model. This classification provides a theoretical framework for planning future experiments and for analyzing existing ones.

One key difference between the two approaches' predictions, which could be addressed experimentally, is to study whether the costs of delays are assessed with a tonic, reward-rate-sensitive depression of the dopamine system (as described in Chapter 3), or instead, whether these costs are batched up and delivered only during occasional phasic dopamine responses, as the present semi-Markov model envisions. In addition, the semi-Markov valuation scheme suggests that it will be important to reanalyze data from dopamine experiments in terms of the trial-by-trial distributions of spiking, in order to search for a predicted correlation between trial-by-trial response strengths and the lengths of delays in experiments where inter-event timing can vary. The semi-Markov model predicts a characteristic pattern for this correlation across a number of experiments, with inhibition of dopamine neurons below baseline levels predicted in response to events occurring later than average. Interpretational difficulties with the rather sparse results of Fiorillo and Schultz (2001) make it difficult to decide this matter definitively at present; similarly, on the issue of tonic responding, the voltammetry data on striatal dopamine concentrations studied in the previous chapter are not sufficiently clear to resolve the contrasting predictions of the Markov and semi-Markov accounts.

Though I have not presented simulations of the version of the present model with Markov rather than semi-Markov valuations (i.e. the top-right model in Table 4.1), it addresses several of the same results as the semi-Markov version. As I have noted, it provides an arguably more natural account for the results of Fiorillo and Schultz (2001) about rewards delivered on a uniform schedule and also for the results of Fiorillo et al. (2003) on tonic responses in a partial reinforcement experiment. However, the semi-Markov version of the model has advantages in other situations; notably, it predicts that dopaminergic neurons can cease responding to reward-related events even when their timing is somewhat unpredictable. This phenomenon has never been explored systematically in experiments (though Figure 4.1, bottom, may be one example), but assuming that every event's timing is somewhat unpredictable due to animals' interval estimation noise, even the standard dopamine experiments in which programmed intervals are constant demonstrate an (uncontrolled) version of this variability tolerance, with rather substantial variability. This seems to be a particular problem for the Markov model, since I can find no way to modify the theory to tolerate the necessary level of jitter. Also, the semi-Markov formalism provides a more natural framework for studying timescale invariance properties in animal learning and behavior, though as I have noted, more work is needed on this point with regard to partial observability. It may be useful in the future to study intermediate models along the graded reduction from the semi-Markov to the Markov valuation scheme (see Section 4.3.8), to see if they can balance the contrasting advantages of each account.

Finally, and importantly, the semi-Markov formalism has also been useful in the present work because it significantly eases analyzing and describing situations when event timing can vary; having achieved a clearer understanding of these issues in the semi-Markov domain, it will now be easier to bring these insights back into Markov models should future data support this.

4.7 Appendix: Mathematical formulae and derivations

Here I present and sketch derivations for the formulae for inference in the partially observable semi-Markov model of Section 4.3. Inference rules for similar hidden semi-Markov models have been described by Levinson (1986); Guedon and Cocozza-Thivent (1990). I also sketch the proof of the correctness of the TD algorithm for that model.

The chief quantity necessary for the TD learning rule of Equation 4.3 is $\beta_{s,t} = P(s_t = s, \phi_t = 1 | o_1 \dots o_{t+1})$, the probability that the process left state s at time t. To perform the one timestep of smoothing in this equation, we use Bayes' theorem on the subsequent observation:

$$\beta_{s,t} = \frac{P(o_{t+1}|s_t = s, \phi_t = 1) \cdot P(s_t = s, \phi_t = 1 | o_1 \dots o_t)}{P(o_{t+1}|o_1 \dots o_t)} \tag{4.5}$$

In this equation, and several below, we have made use of the Markov property, i.e., the conditional independence of s_{t+1} and o_{t+1} from the previous observations $o_1 \dots o_t$ and states $s_1 \dots s_{t-1}$ given the predecessor state s_t. In semi-Markov processes (unlike Markov processes) this property holds only at a state transition, i.e. when $\phi_t = 1$.

The first term of the numerator of Equation 4.5 can be computed by integrating over s_{t+1} in the model: $P(o_{t+1}|s_t = s, \phi_t = 1) = \sum_{s' \in \mathcal{S}} \mathbf{T}_{s,s'} \mathbf{O}_{s',o_{t+1}}$.

Let's call the second term of the numerator of Equation 4.5 $\alpha_{s,t}$. Computing it requires integrating over the possible durations of the stay in state s:

$$\begin{aligned}
\alpha_{s,t} &= P(s_t = s, \phi_t = 1 | o_1 \dots o_t) && (4.6)\\
&= \sum_{d=1}^{d_{max}} P(s_t = s, \phi_t = 1, d_t = d | o_1 ... o_t) && (4.7)\\
&= \sum_{d=1}^{d_{max}} \frac{P(o_{t-d+1} \dots o_t | s_t = s, \phi_t = 1, d_t = d, o_1 \dots o_{t-d}) P(s_t = s, \phi_t = 1, d_t = d | \mathbf{o}_1 \dots \mathbf{o}_{t-d})}{P(o_{t-d+1} \dots o_t | o_1 \dots o_{t-d})} \\
&= \sum_{d=1}^{d_{max}} \frac{\mathbf{O}_{s,o_{t-d+1}} \mathbf{D}_{s,d} P(s_{t-d+1} = s, \phi_{t-d} = 1 | o_1 ... o_{t-d})}{P(o_{t-d+1} \dots o_t | o_1 ... o_{t-d})} && (4.8)
\end{aligned}$$

where d_{max} is the maximum time the system could have dwelt in s. If the last non-empty observation in $o_1 ... o_t$ occurred at time t', then there must have been a state transition at time $t' - 1$ and $d_{max} = t - t' + 1$. The derivation makes use of the fact that the observation o_t is empty with probability one except on a state transition. Thus under the hypothesis that the system dwelt in state s from time $t - d + 1$ through time t, the probability of the sequence of null observations during that period equals just the probability of the first, $\mathbf{O}_{s,o_{t-d+1}}$.

Integrating over predecessor states, the quantity $P(s_{t-d+1} = s, \phi_{t-d} = 1 | o_1 ... o_{t-d})$, the probability that the process *entered* state s at time $t - d + 1$, equals:

$$\sum_{s' \in \mathcal{S}} \mathbf{T}_{s',s} \cdot P(s_{t-d} = s', \phi_{t-d} = 1 | o_1 \dots o_{t-\tau}) = \sum_{s' \in \mathcal{S}} \mathbf{T}_{s',s} \cdot \alpha_{s',t-d}$$

Thus α can be computed recursively, and prior values of α back to the time of the last non-empty observation can be cached, allowing dynamic programming analogous to the Baum-Welch procedure for hidden Markov models (Baum et al., 1970).

Finally, the normalizing factors in the denominators of Equations 4.8 and 4.5 can be computed by similar recursions, after integrating over the state occupied at $t - d$ (Equation 4.8) or t (Equation 4.5) and the value of ϕ at those times. Though we do not make use of this quantity in the learning rules, the belief state over state *occupancy*, $\mathbf{B}_{s,t} = P(s_t = s | o_1 \dots o_t)$, can also be computed by a recursion on α exactly analogous to Equation 4.6, substituting $P(d_t \geq d | s_t = s)$ for $\mathbf{D}_{s,d}$.

The two expectations in the TD learning rule of Equation 4.3 are easily derived. They are:

$$E[\widehat{\mathbf{V}}_{s_{t+1}}] = \sum_{s' \in \mathcal{S}} \widehat{\mathbf{V}}_{s'} P(s_{t+1} = s' | s_t = s, \phi_t = 1, o_{t+1})$$

$$= \sum_{s' \in \mathcal{S}} \widehat{\mathbf{V}}_{s'} \frac{\mathbf{T}_{s,s'} \mathbf{O}_{s',o_{t+1}}}{\sum_{s''} \mathbf{T}_{s,s''} \mathbf{O}_{s'',o_{t+1}}}$$

and

$$\begin{aligned} E[d_t] &= \sum_{d=1}^{d_{max}} d \cdot P(d_t = d | s_t = s, \phi_t = 1, o_1 ... o_{t+1}) \\ &= \sum_{d=1}^{d_{max}} d \cdot P(d_t = d | s_t = s, \phi_t = 1, o_1 \dots o_t) \\ &= \frac{\sum_{d=1}^{d_{max}} d \cdot P(s_t = s, d_t = d, \phi_t = 1 | o_1 \dots o_t)}{\alpha_{s,t}} \end{aligned}$$

where $P(s_t = s, d_t = d, \phi_t = 1 | o_1 \dots o_t)$ is computed as on the right hand side of Equation 4.6.

The proof of correctness of the TD algorithm of Equation 4.3 is as follows. We assume the inference model correctly matches the process generating the samples. With each TD update, $\widehat{\mathbf{V}}_s$ is nudged toward some target value with some step size $\beta_{s,t}$. Analogous to the more standard stochastic update situation with constant step sizes, the fixed point is the average of the targets, weighted by their probabilities and also by their step sizes. Fixing some arbitrary t, the update targets and associated step sizes β are functions of the observations $o_1 \dots o_{t+1}$, which are samples generated with probability $P(o_1 \dots o_{t+1})$ by a semi-Markov process whose parameters match those of the inference model. The fixed point is:

$$\widehat{\mathbf{V}}_s = \frac{\sum_{o_1 \dots o_{t+1}} P(o_1 \dots o_{t+1}) \cdot \beta_{s,t} \cdot (r_{t+1} - \rho_t \cdot E[d_t] + E[\widehat{\mathbf{V}}_{s_{t+1}}])}{\sum_{o_1 \dots o_{t+1}} P(o_1 \dots o_{t+1}) \cdot \beta_{s,t}}$$

The expansions of $\beta_{s,t}$, $E[d_t]$, and $E[\widehat{\mathbf{V}}_{s_{t+1}}]$ are all conditioned on $P(o_1 \dots o_{t+1})$, where this probability from the inference model is assumed to match the empirical probability appearing in the numerator and denominator of this expression. In the partially observable framework, the reward r_{t+1} is a function of the observation o_{t+1}, which we could write $r(o_{t+1})$. Thus we can marginalize out the observations in both sums, reducing the fixed point equation to

$$\widehat{\mathbf{V}}_s = (\sum_{s' \in \mathcal{S}} [\sum_{o \in \mathcal{O}} \mathbf{O}_{s',o} \cdot r(o) - \sum_d \rho_t \cdot d \cdot \mathbf{D}_{d,s} + \widehat{\mathbf{V}}_{s'} \mathbf{T}_{s,s'}])$$

which (assuming $\rho_t = \rho$) is just Bellman's equation for the value function, and is also (of course) the same fixed point as value iteration.

Chapter 5

The organization of reinforcement learning in the striatum

5.1 Introduction

In this chapter, I depart from this thesis' primary methodology of computational modeling, and report complementary efforts to address the same issues empirically, by recording the activity of neurons in the striatum of freely moving rats. The goal of this work is to improve our understanding of the organization of decision-making, both in the abstract sense of what computational steps are involved, and in the concrete, anatomical sense of how these processes are laid out in the brain. Apart from obligatory methodological details, the presentation here will focus on the high-level functional and theoretical implications of the data that are relevant to the overall project of this thesis. These experiments were performed in the laboratory of William Skaggs and with significant technical assistance from Judith Joyce Balcita-Pedicino. Here I present preliminary results from the subset of the recordings that have been analyzed so far; data analysis is ongoing.

TD models of the dopamine system (ranging from Houk et al., 1995, through Dayan, 2002) have long assumed that the dopaminergic TD system is part of a broader actor/critic decision-making architecture (Sutton, 1984), involving an "actor" module that learns to make decisions under the guidance of a "critic" module that specializes in predicting the future values of action outcomes. While this is by no means the only architecture for policy learning using TD methods — in particular, as far as I can see, extant data on the behavior of dopamine neurons are equally compatible with a Q-learning approach (Watkins, 1989) that collapses actor and critic functions into a single learning process — the presumed separation of these functions fits well with more generic notions about the functional organization of the brain, specifically the distinction between dorsal and ventral subregions of the basal ganglia.

The interactions between cortex and the basal ganglia are famously organized as a set of distinct loops (Alexander et al., 1986), in which specific areas of cortex project to associated regions of the striatum (the input structure for the basal ganglia), which in turn return projections, through several further basal ganglia substructures, back to their cortical afferents. One of the most striking divisions is between the dorsal part of the striatum, which predominantly receives inputs from sensorimotor cortex, and the ventral striatum (also called the *nucleus accumbens*), which is more connected with limbic structures such as the hippocampus, the basolateral amygdala, and the orbitofrontal cortex. Thus while the basal ganglia as a whole are thought to have a broadly motoric function, their ventral circuits and particularly the ventral striatum are thought to have a unique function in processing the motivational and emotional information with which their limbic inputs are implicated ("the limbic-motor gateway"; Mogenson et al., 1980). This general separation of striatal function is supported by targeted lesion and drug infusion studies (Everitt and Robbins, 1992; Robbins and Everitt, 1996), but it has not previously been reflected in recordings of the behavior of striatal neurons. This is one goal of the present work, in which both dorsal and ventral striatal neurons were recorded in animals performing a task that was designed to tease apart differences in firing related to their putative functions. Specifically, the neurons' behavior was recorded across reversals in task contingencies intended to distinguish firing related to actions from firing related to their predicted outcomes.

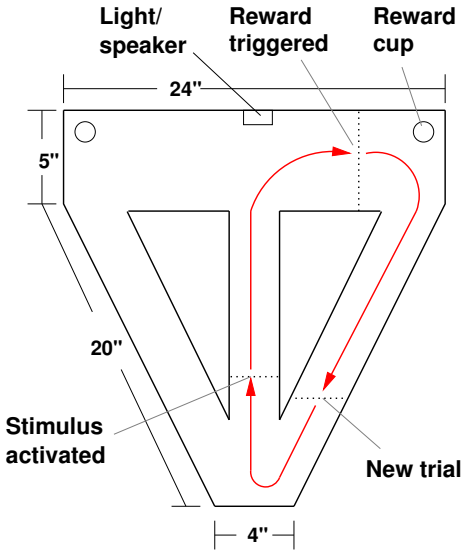

Figure 5.1: The experimental apparatus, showing the positions at which trial events are triggered.

Striatal function and organization are relevant to TD models of the dopamine system because the striatum is strongly interconnected with the dopamine neurons, serving both as one of their major outputs and major inputs. The dopamine neurons are also organized along dorsal/ventral lines, with two clusters of neurons (the substantia nigra pars compacta, SNc, and the ventral tegmental area, VTA) associated respectively, though not exclusively, with the dorsal and ventral striatum. One early puzzle in interpreting dopaminergic recordings was an inability to find distinctions in the firing properties of these two groups of neurons that would reflect the functional distinctions between their associated areas of striatum. Instead, neurons recorded in both dopaminergic areas behave similarly (Schultz, 1998). In actor/critic models, this non-result is explained by the fact that both the actor and the critic share the same error signal; thus Montague et al. (1996) proposed that the SNc and VTA carry the same error signal in the service of distinct functions at distinct targets, in one case for learning action selection functions (at dorsal striatal targets) and in the other for reward-prediction functions (perhaps in the ventral striatum or elsewhere). The present experiment is grounded in, and aims to test, this idea of a dorsal/ventral breakdown between actor and critic functions in the striatum.

The rest of this chapter is organized as follows. Section 5.2 details the experimental methods. Sections 5.3 and 5.4 detail the behavioral and neurophysiological results, and Section 5.5 discusses their implications.

5.2 Experimental methodology

5.2.1 Behavioral task

Six male Sprague-Dawley rats (hereafter referred to by number: 142, 146, 147, 160, 165 and 182) were food-deprived to 85% of their free-feeding weights and trained to perform cued decisions for food pellets or chocolate milk in a continuous T-maze (Figure 5.1) in daily sessions lasting between 20 minutes and an hour. During one or two initial sessions, each animal was acclimatized to the maze and given food pellets there. In several subsequent sessions, the animal was trained to traverse the maze, ascending the central corridor, turning to one of the corners, and returning to the start via one of the long diagonal arms. This was accomplished by allowing the animal to follow trails of food pellets along the appropriate path, and also by blocking inappropriate directions of travel with a cardboard barrier. Use of the barrier, and the availability of food at locations other than the feeding stations at the two maze corners, were gradually eliminated.

When an animal could reliably traverse the route, receiving reward only at the corner stations for a full

Before:	light→left→milk	tone→right→food
After **cue-action** reversal:	**tone**→left→food	**light**→right→milk
After **cue-outcome** reversal:	light→left→**food**	tone→right→**milk**

Table 5.1: Summary of two sorts of contingency reversals used in the experiment.

	Reward types	Reversal types	Recording sites	notes
142	food only	cue-action	both	tended to alternate
146	both	cue-reward	both	ventral data not yet analyzed
147	food only	none	both	
160	both	cue-action	dorsal	died early
165	both	cue-action	ventral	
182	both	cue-action	both	data not yet analyzed

Table 5.2: Summary of the treatments to which each subject was exposed.

maze circuit, cues were introduced. Starting at this point, the animal's position was tracked using a camera and either a reflective collar (pre-implantation) or LEDs on his headstage (post-implantation), and cue and reward delivery were controlled by a clock sequence running on a Datawave Discovery system (Datawave Technologies, Longmont CO). Also starting at this point, for five animals (146, 160, 165, and 182), chocolate milk reward was substituted for food pellets at one of the feeding stations, delivered in small drops through a tube. (Animals 142 and 147 were rewarded only with food pellets throughout the experiment.) Rats were acclimatized to the unfamiliar chocolate milk using a bottle placed in their home cage overnight. Note also that the subjects were not water-deprived, though they were not provided water during the experimental sessions; the chocolate milk was presumably rewarding insofar as it was nutritive. Particularly in early training sessions, however, it was evident on observation that the animals preferred the food to the chocolate milk.

When the animal left the maze start and crossed a threshhold while ascending the central corridor (Figure 5.1), a light or tone cue was activated. A light signaled a left turn; the cue was extinguished and reward was delivered at the left feeding station if the animal crossed a threshhold in the left branch of the maze. Tone signaled a right turn and reward was delivered in the right corner for correct performance. If the animal made the wrong turn — that is, if it crossed the threshhold in the incorrect arm prior to crossing the threshhold in the correct arm — then the cue was extinguished and no reward was delivered. In either case, the animal could start a new trial by returning to the start of the maze and again ascending the middle corridor. The stimulus, tone or light was chosen purely at random for each trial with equal probabilities.

Once they could perform the cued task reliably (more than 90% of trials correct), all animals except for number 147 were exposed to repeated *reversals* in the task contingencies. Reversals were of two types (Table 5.1). All but one of the animals were exposed to reversals in the *cue-action* contingencies. Where the light had previously signaled a left turn, after the reversal it signaled a right turn, and the opposite for the tone. Cue-reward contingencies were maintained across the reversal, so if a light had been rewarded with chocolate milk on the left prior to the reversal, it would be rewarded with chocolate milk on the right afterward. Animal 146 instead received reversals in the *cue-reward* contingencies, with the cue-action contingencies held constant. Thus, before a reversal, a light signaled a left turn and chocolate milk reward, and afterward it still signaled a left turn but for food reward. Note that both sorts of reversals dissociate reward type from turn direction. An initial attempt to alternate both sorts of reversals in animal 182 was abandoned after it seemed to confuse him; thereafter he was subject only to reversals in the cue-action contingencies. Table 5.2 summarizes the different treatments to which each subject was exposed.

After animals regained accurate behavior, the contingencies were reversed again (back to the original contingencies); this serial reversal and relearning process was repeated until the animals were able to relearn quickly enough that stable and accurate behavior under both sets of contingencies could be observed in a single recording session. For animals 160, 165, and 182 (after abandoning the additional cue-outcome

reversals) this took only 5-10 additional training sessions; animal 142 required three months of additional daily training, after which relearning was still an order of magnitude slower than the others due to a strong post-reversal tendency to alternate between sides without regard to the cues. This strong difference between subjects is probably due to the use of two different rewards with the more successful subjects; it is a well-known behavioral phenomenon (called the "differential outcome effect") that learning of a conditioned discrimination is faster if the two responses are differently rewarded. Animal 146's behavior was largely stable over the cue-reward reversals, so prolonged retraining was unnecessary.

5.2.2 Surgery

Following successful behavioral training, each rat was anesthetized with injections of xylazine and ketamine (and further periodic boosters of each during surgery) and placed in a stereotaxic device. A custom-made "hyperdrive" containing 12 individually drivable tetrodes and two single-channel reference electrodes was positioned above the right medial striatum, approximately 1.5mm anterior to bregma and 1.5mm right of the midline. Figure 5.2 shows the intended electrode placement. The tetrodes were made from 25-micron nichrome wire insulated with polyimide. The microdrive was cemented to skull screws using dental acrylic. Immediately after surgery, the electrodes were lowered about 5mm into the dorsal striatum, and the wound was covered with antibiotic/anesthetic ointment. Over the next several days, the wound was cleaned and ointment was reapplied until the animal had healed. After about 7 days, the animal was reintroduced to the recording chamber for one session of behavioral retraining, followed by recordings. For rat 165, the electrodes were further advanced to a depth of about 6.5mm, consistent with the nucleus accumbens, prior to recording.

Apart from the drive placement and a thicker gauge of tetrode wire (chosen to promote straight passage for deeper recordings), the methods and microdrive design follow those reported in more detail elsewhere for hippocampal recordings (Skaggs et al., 1996).

5.2.3 Recordings

For recording, the top of the hyperdrive was connected to a headstage containing preamplifiers and position tracking LEDs, and tethered with flexible cables and a pulley/counterweight system to a Cheetah parallel recording system (Neuralynx, Tucson AZ) consisting of eight 8-channel amplifiers, programmable filters and A/D converters. Data for each channel were filtered to 600-6000 Hz, sampled at 32KHz, and stored on disk for later offline analysis using a Pentium 4 computer running Windows NT and Neuralynx Cheetah acquisition software.

Each daily recording session lasted roughly 30 minutes and contained around 200 trials. (Those involving animal 142 were as much as three times longer due to poor behavior.) For those animals experiencing reversals, the recordings began with the animal behaving under the contingencies he had most recently experienced in the previous day's recordings; once a sufficient number of correct trials had been recorded, the contingencies were reversed and recording continued until more stable, correct behavior had been recorded. Neural activity was also recorded each day during a baseline period of quiet resting in a nest.

After each day's recordings, all tetrodes were advanced 80-160μm deeper into the striatum. In this way, over the course of recordings, striatal neurons were sampled from a variety of depths across the dorsal and ventral portions of the structure. No attempts were made to "tune" electrode placement based on proximity to neurons, reducing the chance of a bias in neuronal sampling.

5.2.4 Histology

After some 30 recording sessions, animals were humanely sacrificed and their brains and hyperdrives removed for analysis. (Animal 160 lost his headstage after only six recording sessions and was sacrificed prematurely.) Brains were sliced, stained with a Nissl stain to reveal cell bodies, and mounted on slides. These slices were examined under a microscope to verify proper electrode placement. This analysis revealed a few tetrodes that had strayed medially across the ventricle that separates the striatum from the septum (see Figure 5.3 for an example). These were excluded from further analysis. Apart from these, the tetrodes were well-placed and straight, but it was not possible to ascertain from the weak tracks in the slices the ultimate depth that

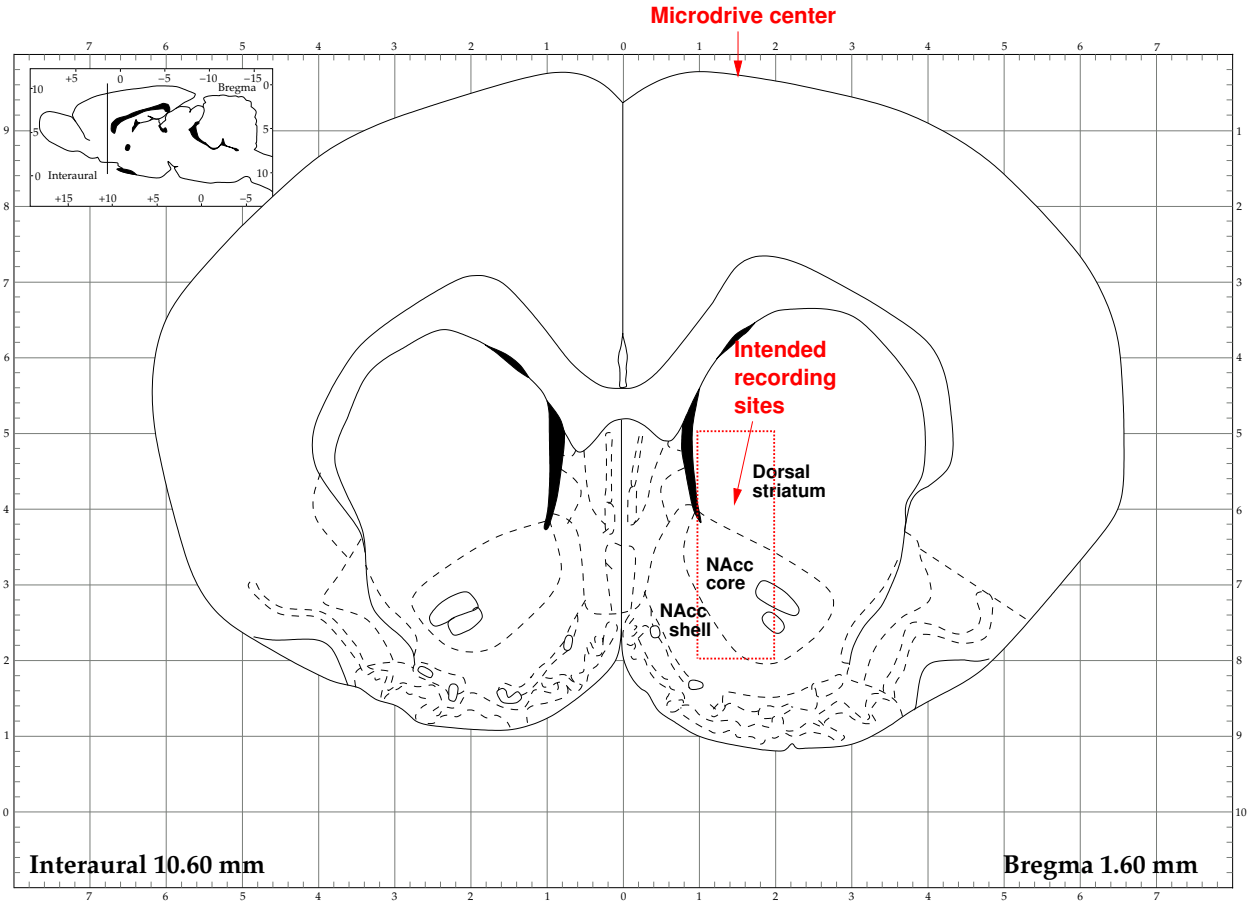

Figure 5.2: Diagram of rat brain showing intended microdrive placement and recording sites. Adapted from Paxinos and Watson (1996). NAcc: Nucleus accumbens (ventral striatum).

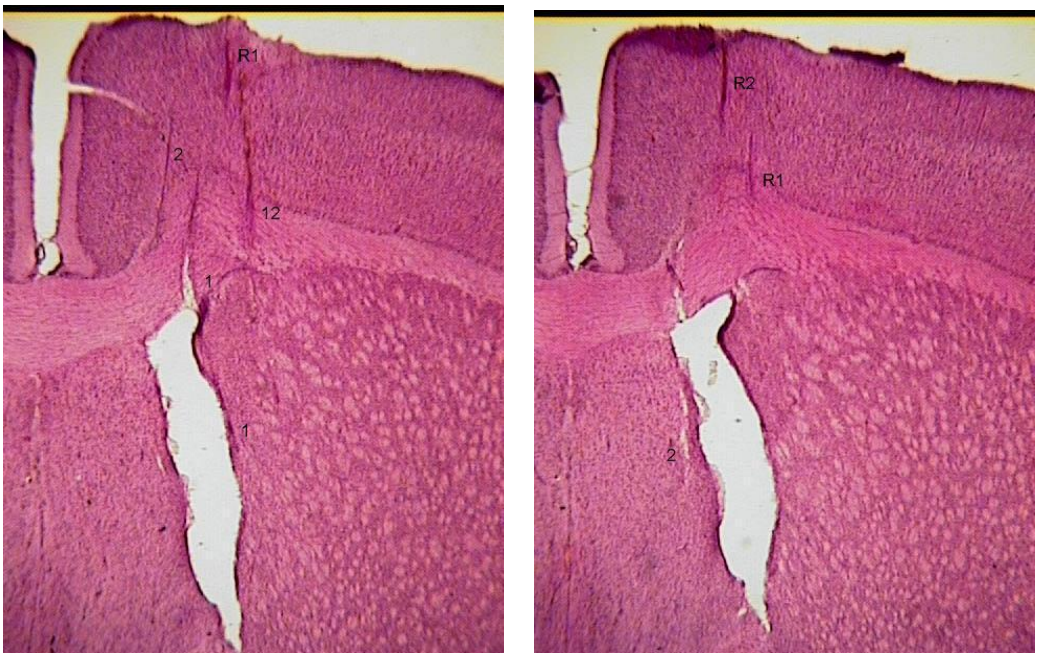

Figure 5.3: Two examples of histological slices from animal 142, showing parts of identified electrode tracks. Numbers show identified electrodes ("R1" and "R2" are reference electrodes.) Electrode 1 passes through the ventricle (the light gap surrounded by dark brain matter) but emerges in the striatum; in the second slice, electrode 2 is seen to enter the septum.

they had reached. This was instead estimated by measuring the actual length of electrodes protruding from the removed hyperdrive; these measurements revealed that a number of tetrodes had become stuck at depths above their intended destination. From this information together with the daily turn records, the depth of each day's recording on each tetrode was estimated.

5.2.5 Data analysis

Recorded spike events were sorted into multiple "clusters," each putatively corresponding to the group of spikes originating from a single neuron, based on multiple waveform features across the four channels of each tetrode. This was done either manually using XCLUST software (M. A. Wilson) or by using MCLUST software (A. D. Redish) to manually clean up a rough initial cut performed automatically using KLUSTAKWIK (K. D. Harris). Only clusters judged reasonably complete and well-isolated were subject to further analysis.

One challenge in visualizing and analyzing the behavior of the neurons is that the animals performed the task self-paced. Thus the intervals between events varied from trial to trial, making it difficult to ascertain to what events neuronal firing was related using standard methods like a peri-event time histogram. To address this problem, I devised a method for creating histograms that approximate peri-event time histograms aligned on multiple events, to produce a timeline capturing each neuron's profile of firing throughout an entire circuit of the maze. The events on which trials were aligned correspond to the animals crossing various spatial threshholds, which are drawn as dotted lines in Figure 5.4. Note that while we can accurately gauge the time of stimulus and reward delivery (which are triggered by the tracked position crossing spatial threshholds), there is no way to measure when the reward is actually consumed. We estimate this by the animal's arrival at the food cup. Similarly, the point marked "turn" does not literally correspond to any actual measurement of turn initiation, but to the animal arriving at a spatial location near where he must turn left or right.

To create a histogram of firing, for each trial, the interval between each pair of adjacent events (e.g. between the stimulus onset and the turn) was divided into a set number of temporal bins of equal length,

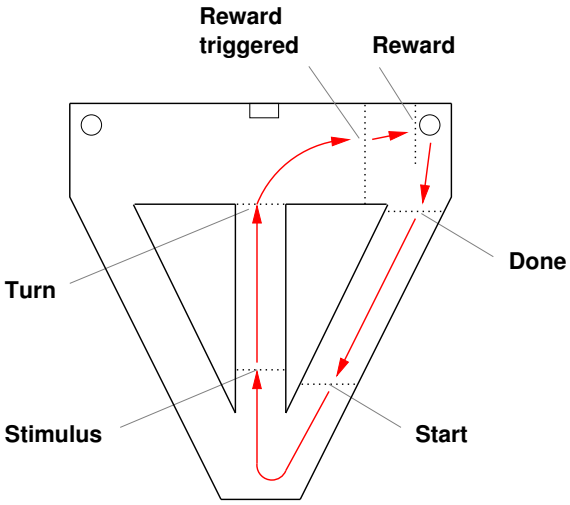

Figure 5.4: Illustration of the events on which spike traces are aligned in constructing a timeline of neuronal firing over the course of a trial.

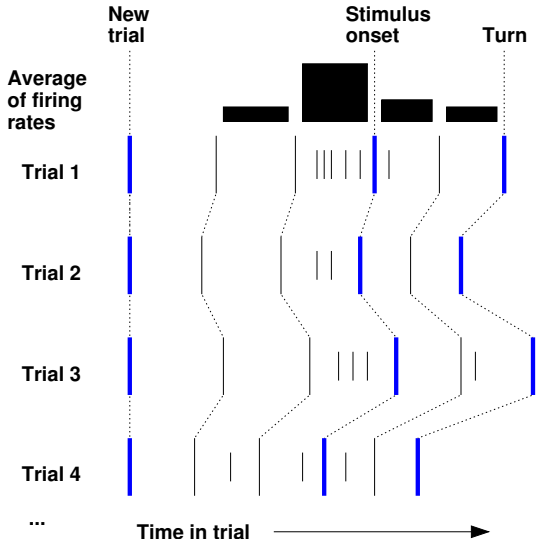

Figure 5.5: Schematic illustrating how the intervals between events (thick lines) are divided into equal-sized temporal bins on each trial, and rates of spike firing (short lines) are averaged between these bins to produce a histogram.

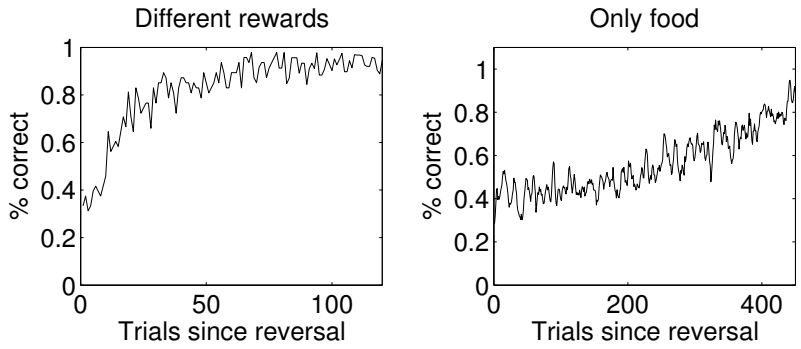

Figure 5.6: Learning curves following stimulus-action reversals. Left: Percentage of correct behavior by number of trials following a reversal in 48 sessions with animals 160, 165, and 182, who were rewarded with chocolate milk on one side of the maze and food on the other. Right: Presumably due to the differential outcomes effect, animal 142, who received only food reward, learned a great deal slower (note different x axis scaling).

the number chosen so that the bin length would correspond to roughly 100 ms in a typical trial.[1] A firing rate was computed for each temporal bin in each trial as the spike count divided by the bin duration in that trial, and these rates were averaged between trials to produce an average firing rate for that phase of the trial. The process is illustrated in Figure 5.5. Spike rasters are drawn underneath the histogram; due to analogous alignment difficulties, these are not raw rasters but linearly stretched between events so that the spikes fall underneath the histogram bar to which they contribute. In some cases, these results were compared to more traditional analysis methods such as peri-stimulus event histograms for a single event or spatial firing rate maps, in order to verify that conclusions were not due to some artifact of this analysis. No such anomalies were detected.

Because the trials are self-paced, there are analogous difficulties in determining whether neuronal firing differs significantly between different subsets of trials, e.g. those involving a right versus a left turn. To do this, the firing rate between a pair of adjacent events (spike count divided by interevent interval) is treated as a random variable whose value is sampled on each trial. Differences in firing were then detected using a t-test on these firing rates. In the case of right versus left turns, a neuron was considered to distinguish the trajectories if its firing rates differed significantly between any pair of adjacent events, e.g. stimulus and turn or turn and the reward trigger.

5.3 Results: behavior

Most of the animals were able quickly to resume correct behavior after a reversal in stimulus-action contingencies. Figure 5.6, left, displays the percentage of correct choices as a function of the number of trials since the last reversal, averaged over all recording sessions with animals 160, 165 and 182, a total of 48 sessions. Though after a reversal, behavior briefly dropped below chance, it recovered to above-chance levels within about 20 trials and to an asymptotic value of better than 90% correct after another 40 trials. As shown in the right side of the figure, animal 142 learned a great deal more slowly (note the difference in x axis scaling). The difference between the behavior of this animal, who received only food reward, and the others, who received both food and chocolate milk, is presumably due to the differential outcomes effect. This trace is averaged over 21 recording sessions and temporally smoothed using a Gaussian kernel with 5-trial support.

[1] Specifically, the number of bins was the modal interevent timing for each pair of events in some 48,000 trials across 133 recording sessions with six animals, divided by 100 ms and rounded. To facilitate comparisons, the same numbers of bins were used for visualizing data recorded from all animals and all sessions; thus the actual length of time corresponding to a bin varies due to between-trial, between-session and between-animal variability. However, within a given session, interevent timings were fairly stable, and a very few outlier trials in which an interevent interval far exceeded the average for that session were excluded from analysis.

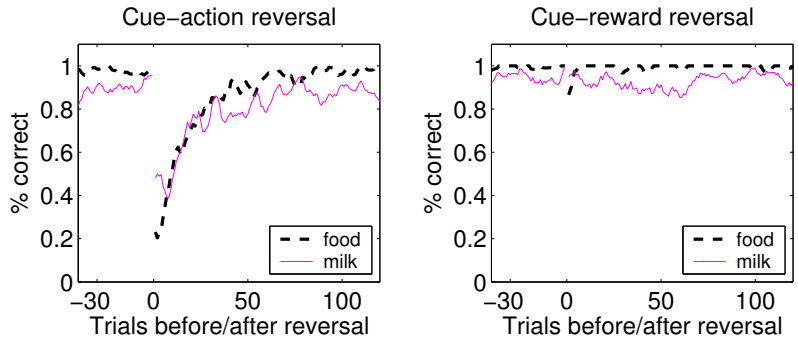

Figure 5.7: Percentage of correct choices preceding and following reversal, broken down by trials in which the reward for correct behavior would have been food versus chocolate milk. Left: Higher error rates for chocolate milk over food in 48 cue-action reversals with animals 160, 165, and 182. The reward sides are switched at the reversal, which briefly reverses the apparent preference, but the behavior quickly recovers. Right: The same preference is visible in 28 cue-reward reversals with animal 146.

The overall averages in this plot are depressed by the results of a few sessions in which the animal failed to learn the reversal; on most days, however, it was possible to record a substantial block of post-reversal trials in which the animal behaved at least 80% correctly.

It is also interesting to note that the animals consistently appeared to prefer the food reward to the chocolate milk. Anecdotally, this preference was often evident in their behavior; for instance, when the cue signaling that a trial was to be rewarded with chocolate milk was activated, animals could sometimes be seen to rear back, momentarily shying away from the stimulus. (The observation of such a response to a stimulus signaling presumably worse-than-average reward fits well with the ideas in Chapter 3 concerning average-adjusted TD error and the putative involvement of serotonin both in signaling negative prediction error and in energizing withdrawal responses.) More quantitatively, the error rate was higher for trials that were to be rewarded with milk (12%, over all 26,000 trials recorded from animals 146, 160, 165 and 182) than for trials that were to be rewarded with food (5%). This difference was highly significant (p not measurably different from zero given the floating point tolerances of Matlab), using a two-proportion z-test for Bernoulli trials.

Figure 5.7 gives an indication of the time course by which these preferences react to reversals in the stimulus contingencies. The left side of the figure separates the learning curve from Figure 5.6, left, into separate curves for trials in which the animal was asked to turn in the direction that would have resulted in food versus chocolate milk. Behavior before the reversal is also shown. (Due to the reduced number of sessions contributing to each data point, it was necessary to temporally smooth the data with the Gaussian kernel described above.) Asymptotically, the animals make most of their errors on the chocolate milk side; this preference flip-flops only briefly just after the reversal (when the food and chocolate milk sides are reversed but the previous preference is apparently still in effect), but quickly returns as the task is relearned. Though error bars are not shown, the preference for food re-emerges as significant about 40 trials after the reversal (two-proportion z-test on blocks of 5 trials following reversal, $p < .05$).

The right plot in Figure 5.6 illustrates the same preference in the behavior of animal 146, over 28 recording trials with cue-reward reversals. The behavioral distinction between trials in which the stimulus predicts food versus chocolate milk, and the fact that the preference recovers so quickly following a reversal in the stimulus contingencies, provide evidence that animal 146 is aware of the cue-reward contingencies, despite frequent reversals in their relationships. [2] In this case, the preference for food re-emerged as significant

[2]This evidence is not decisive since the illustrated behavior could have been produced without any representation of the stimulus-reward contingencies. For instance, imagine that the behavior was driven by two competing stimulus-response motor habits, which were reinforced by the subsequent rewards. If the food reward were more effective than the chocolate milk at "stamping in" the choice that preceded it, then this pattern of errors might be expected. This seems unlikely since it does not explain the anecdotal observations, mentioned above, of the animals rearing back at the onset of the stimulus that predicted

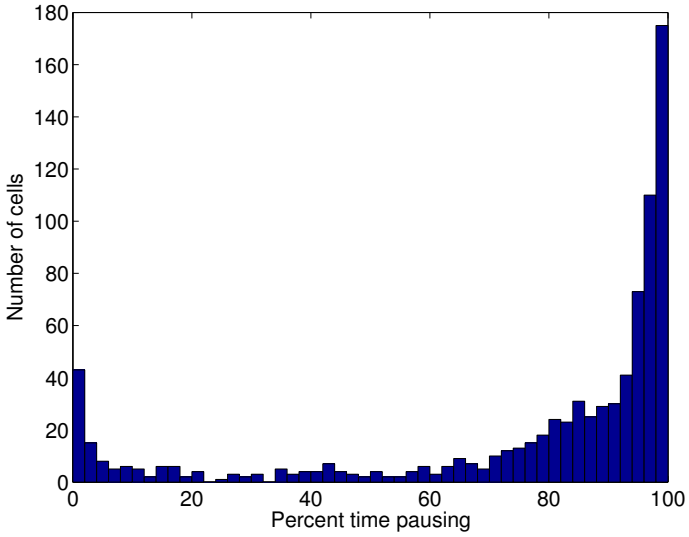

Figure 5.8: Histogram over 820 recordings of the percentage of time the neuron spent in interspike intervals lasting more than 1.3 seconds, a measure of the burstiness of responding used by Redish et al. (2002) to classify striatal neurons. Most neurons group into the extremes, with tonic neurons (probably corresponding to interneurons) on the left and phasic neurons (probably corresponding to medium spiny projection neurons) on the right.

in the second block of 5 trials following the reversal (two-proportion z-test on blocks of 5 trials following reversal, $p < .05$). This is an important point for the interpretation of subsequent neuronal data, since we are interested in determining whether the neurons are sensitive to the predicted reward type. If the animals themselves were unaware which reward was forthcoming, then there would be no use searching for neural correlates of this prediction.

5.4 Results: neurophysiological recordings

On the basis of their spiking behavior, the 820 clean and complete recorded units fell into a variety of classes, seemingly consistent with striatal neuronal types known from previous studies. About 85% of the recordings showed low firing rates and bursty firing patterns, and were likely from medium spiny neurons. The remaining neurons fired tonically (though their continuous activity was often nonetheless modulated in relation to task events) and probably corresponded to various sorts of interneurons. Figure 5.8 shows that tonic and phasic neurons separate clearly on a histogram of how the recorded neurons distribute according to the percentage of time each neuron spent in interspike intervals lasting more than 1.3 seconds. This is an index of burstiness that was used by Redish et al. (2002) to classify striatal neurons, and my results are similar to theirs. (They used a threshhold of 5 seconds on interspike intervals, which proved too long for the present recordings. The difference is probably due to the fact that their experiment was conducted on a larger maze, with a longer total cycle time. Since bursty neurons tend to be active at particular locations in the circuit, circuit time is related to pause duration.) The tonic neurons were heterogeneous, including slowly firing neurons and a smaller group of exceptionally active neurons, a finding also in accord with previous studies. Figures 5.9 and 5.10 display example histograms of several tonic and phasic neurons' firing patterns over the course of a maze circuit, for the purpose of demonstrating broadly the sorts of responses seen. Hereafter I group all neuronal types together for analysis.

Several points are obvious in Figures 5.9 and 5.10. First, neuronal firing rates change systematically in relation to task phase. The tonic neurons show various sorts of modulations, including inhibitory periods,

chocolate milk. Such Pavlovian behavior would seem to require a true representation of the stimulus-outcome contingencies.

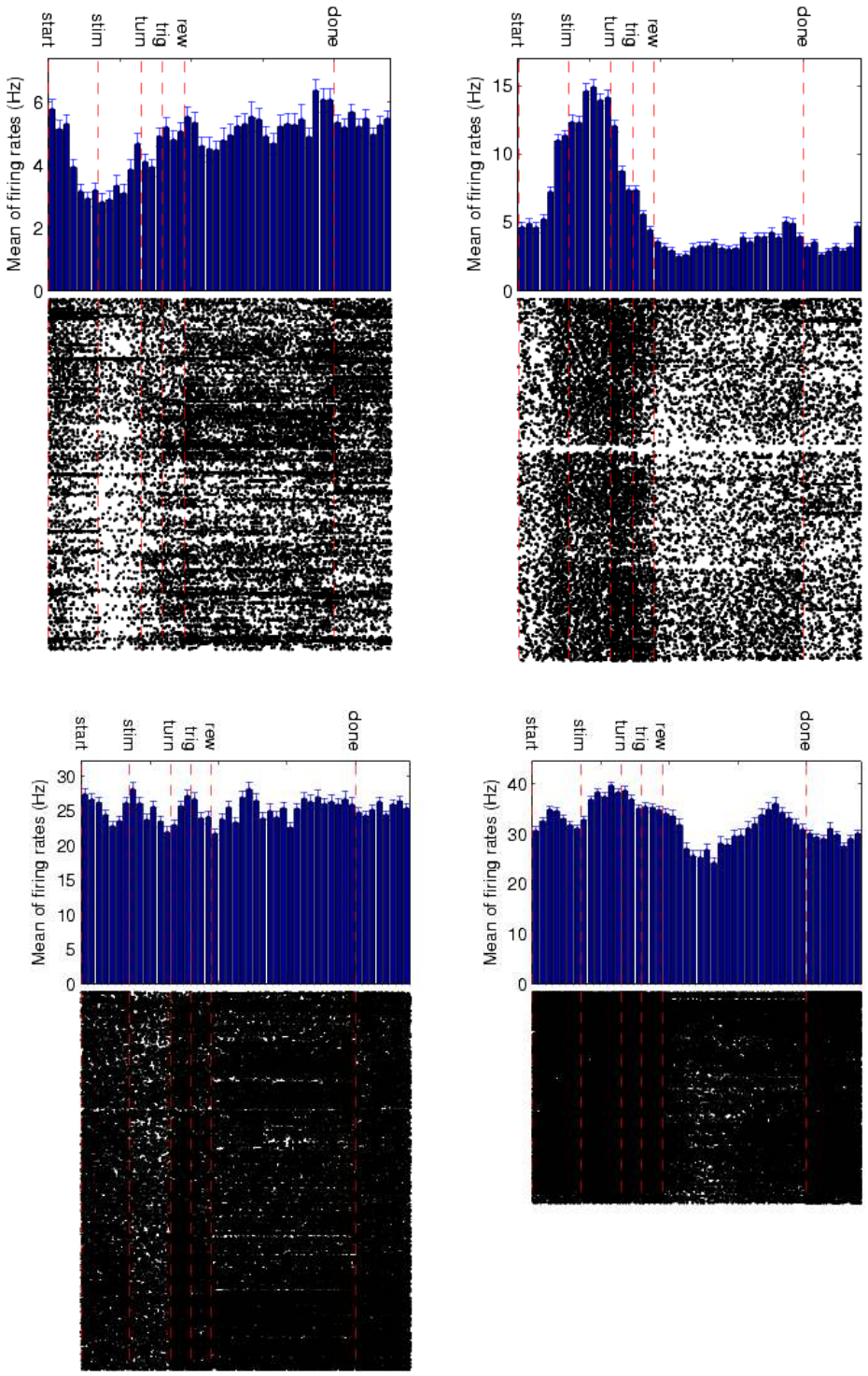

Figure 5.9: Firing traces of several tonic neurons. These were heterogeneous in firing rate (note differences in y-axis scaling) and usually showed complex modulations or oscillations in their firing over the course of a trial. The unit IDs, from left to right and top to bottom, are 142-83-8-3, 142-78-2-1, 142-90-6-4, and 160-25-5-1.

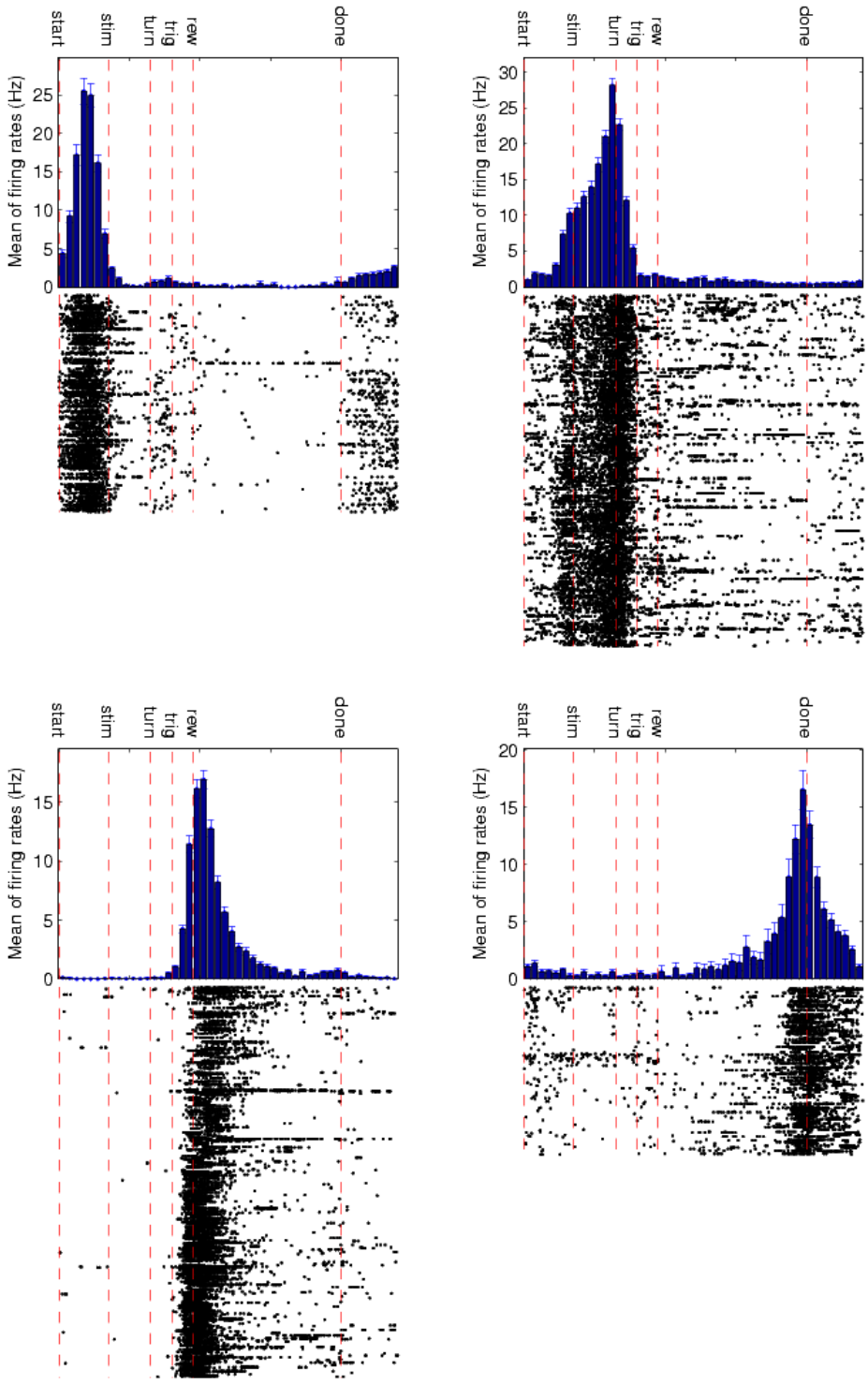

Figure 5.10: Firing traces of several phasic neurons, showing tuning to different phases of the task. These are representative of the majority of responsive neurons and probably correspond to medium spiny neurons. The unit IDs, from left to right and top to bottom, are 160-25-11-3, 142-79-8-1, 142-84-8-1, and 160-22-10-3

while the phasic neurons show tuned excitation. In light of the schema of Schultz et al. (see Figure 2.3 on page 24) it might be tempting to conclude that these neurons fall into several classes, each class tuned (with either anticipatory or responsive firing) to one of the controlled events in the task, such as the stimulus onset or reward delivery. In fact, as discussed below, responses seem to tile the whole task timeline more evenly than this viewpoint would suggest. Second, there is considerable trial-to-trial variability in the neuronal responses. The present figures combine all trials and also all reversal conditions; later I will present results that break down responding in various subsets of trials in order to capture some of this variability. There is a further sort of variability, which is that even in recordings that seem otherwise to be well-isolated, complete and stable, the overall firing rate seems sometimes to drift around while responding nevertheless follows roughly the same temporal envelope. My attempts to correlate these changes in overall firing rates to task events or animal behavior have not so far been successful, though they are ongoing.

Figure 5.11 displays similar histograms of firing for all recorded neurons that met a test of minimal responsiveness on the task. One row of each colored figure corresponds to a single recorded neuron; the color at a particular point was assigned based on the extent to which that neuron's firing is tuned to that point in the task. That is, coloration was assigned based on firing histograms like those shown in Figures 5.9 and 5.10, but the summed height of histogram bars was normalized and colors assigned based on these normalized values. Warmer colors, corresponding to higher normalized firing rates, are thus displayed for those bins to which a neuron's firing was most sharply tuned, regardless of its overall firing rate. Neurons with very low firing rates (those not exceeding 3 Hz in some histogram bucket) were excluded from this figure for fear that unresponsive neurons would misleadingly be displayed as sharply tuned to some point in the task; e.g. a neuron that fired one spike at random would show up, on this analysis, as maximally tuned to the point in the trial in which the spike occurred. In these figures, neurons are sorted by the location of their peak firing.

The topmost plot in Figure 5.11 displays all 334 responsive neurons out of 820 reasonably clean and complete recordings so far analyzed.[3] Most neurons responded preferentially at some point in the task, and their preferred response times tiled the task timeline rather evenly. There appear to be neurons that coded for any particular point in the task, as can be seen in the diagonal banding that runs the length of the figure. However, it would be a mistake to conclude from this figure that, as a population, the neuronal code was equally sensitive to all points in the task and did not favor specific events such as rewards; in part because of difficulties involving trial alignment and the actual temporal lengths of different histogram bins, it would be difficult to pass quantitative judgment on the degree of population responsiveness at different points in the task. Nonetheless, in light of the classification scheme of Schultz et al. (Figure 2.3), which envisions strictly clustered firing, the interesting result here is the relative evenness by which firing fields are distributed throughout the task timeline. Caveats about the histogram process also presumably account for the break in the figure's diagonal band after the animal arrives at the food cup and before it returns to the bottom of the maze to start the next trial. This is the time period in which the animals' behavior and also the interevent timings were most variable — for instance, on some trials they walked around the top of the maze, sniffed over the edge of the maze, or took a quiet break. The lack of well-tuned neuronal firing in the figure during this period presumably relates to poor trial alignment due to behavioral variability, as opposed to the neuronal code actually not representing whatever was going on in any individual trial at that point.

The middle and bottom plots of Figure 5.11 display subsets of the full neuronal library in an attempt to explore differences between firing in the dorsal and ventral parts of the striatum. The middle plot, consisting of 206 neurons recorded at depths consistent with the dorsal striatum (< 6.5 mm deep), appears similar to the full population plot. The bottommost figure contains 128 neurons recorded at depths consistent with the ventral striatum in animals 142, 147 and 165. Due to a smaller number of neurons (and generally poorer recordings), these data are preliminary. Also, as a population, the ventral neurons were less sharply tuned than the dorsal ones; for this reason, it was necessary to change the scale of coloration to improve legibility of this figure. (Specifically, the normalized firing rates in the ventral plot were multiplied by 1.5 and then subject to the same color mapping as the other plots.) Given all of these caveats, the firing fields of the

[3] In a few cases in which a tetrode was stuck near a neuron and its position remained stable, the same neuron may have been recorded in several sessions. As no attempt has been made to remove such duplicates here, there are likely a few cases of repetition in the plots in Figure 5.11, and the counts of recorded and responsive neurons are likely slightly inflated accordingly. Later, when results are presented as to how many neurons pass various statistical tests about their behavior, recordings that appeared to be duplicates were eliminated by hand.

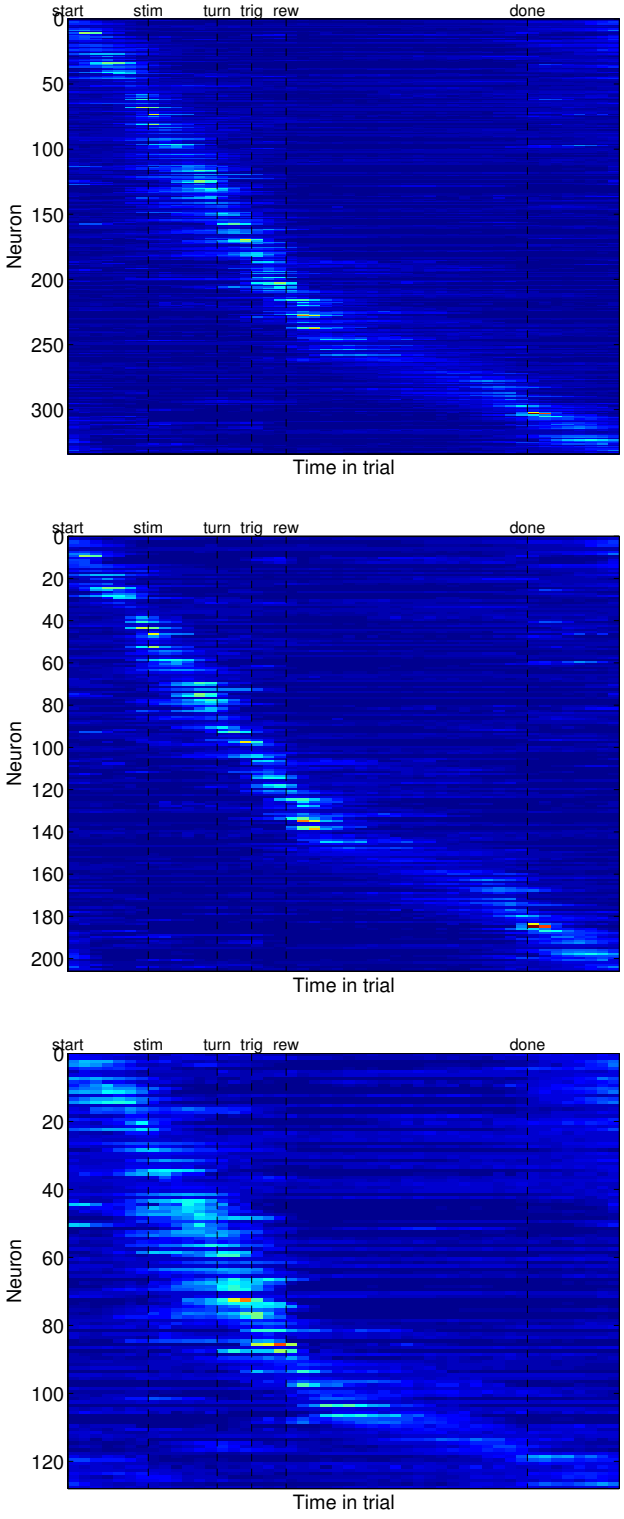

Figure 5.11: Timelines of neuronal firing (shown as colored traces, one row per neuron) over the population of recorded neurons. Top: All recorded neurons that met a test of minimal responsiveness. Middle: Neurons recorded in the dorsal striatum. Bottom: Neurons recorded in the ventral striatum.

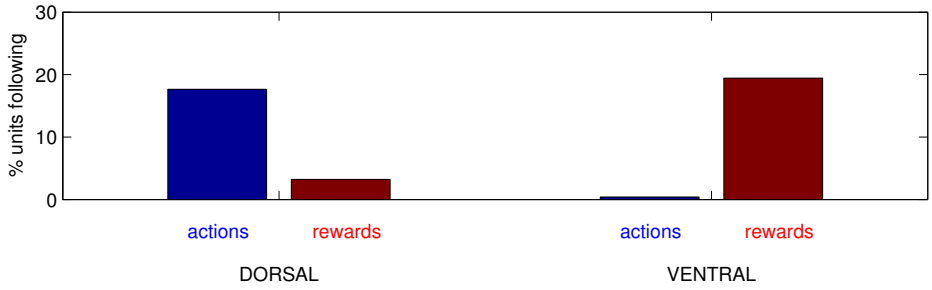

Figure 5.12: Summary of the percentage of recorded neurons that significantly distinguished either turn direction or reward type, broken down by depth of recording.

neurons shown here seem to have a more pronounced tendency than the dorsal population to cluster during the period between the start of the trial and the reward consumption. In particular, very few neurons are seen that fired preferentially during the period when the animal was returning to the bottom of the maze to start a new trial (these would have appeared at the lower right hand corner of the plot). Though the differences on this analysis are slight, they are consistent with a more pronounced distinction seen on a more sensitive analysis of how the neuronal populations respond to reversals, as discussed next.

In order to determine whether dorsal and ventral neurons were sensitive to actions or instead to their outcomes, the preferences of neurons before and after contingency reversals were compared. Contingency reversals were studied in animals 146, 160, and 165, from whom 434 units have so far been analyzed. (Reversals were also studied in rat 142, but those data are not useful for this comparison since both alternatives were rewarded with food.) Neurons were screened for those that, during the periods before and after a reversal considered separately, fired at significantly different rates during trials involving left and right turns (or, equivalently, since the directions and outcomes are confounded in a single reversal condition considered alone, during trials rewarded by food versus chocolate milk). Significantly different firing was detected using a t-test ($p<.005$) on trial-by-trial firing rates between the two blocks of trials. Firing rates were separately tested using spike counts taken during the intervals between the stimulus and the turn, the turn and the food delivery, and the food delivery and the rat's arrival at the food cup. Neurons displaying significant firing rate differences during any of these intervals were classified as distinguishing the two blocks of trials. (Neuronal firing was not considered during food consumption, to avoid any chance that the results would be corrupted by chewing- or licking-related electrical artifacts or by neuronal firing related to differences in the two consumptive behaviors rather than the rewards *per se*.)

If a neuron showed significantly different firing between left- and right-turn trials both before the reversal and after, then it was possible to disambiguate whether the preferences were primarily related to the turn directions or to their outcomes, by testing for a reversal in the neuron's side preference. If the neuron fired more during trials with left (or right) turns during both reversal conditions, it was classified as related to the actions; if it instead reversed its side preference between conditions, firing more during chocolate milk (or food) trials in each condition, it was classified as related to the rewards. One neuron was found with ambiguous preferences, depending on during which interval of the trial the firing rates were measured; on examination, this was classified as an action-related neuron, as that phase of the response was more robust. Figure 5.12 summarizes the patterns of sensitivity to turn direction versus reward type in dorsal versus ventral striatum. In all, dorsal neurons tended to follow the turn directions on reversal (33/187) rather than the reward type (6/187), while the opposite pattern was seen in ventral striatum, where 48/247 neurons followed the reward type and only one neuron followed the actions. Though these numbers may seem low as a percentage of all neurons, recall that more than half the neurons were unresponsive and many of the rest were active only during periods of the task other than the period relevant to this analysis, that between stimulus and reward.

Though the contrast between dorsal and ventral units is striking, there are two caveats to this result. First, the ventral data are taken entirely from one animal, 165, while the bulk of the dorsal data are from

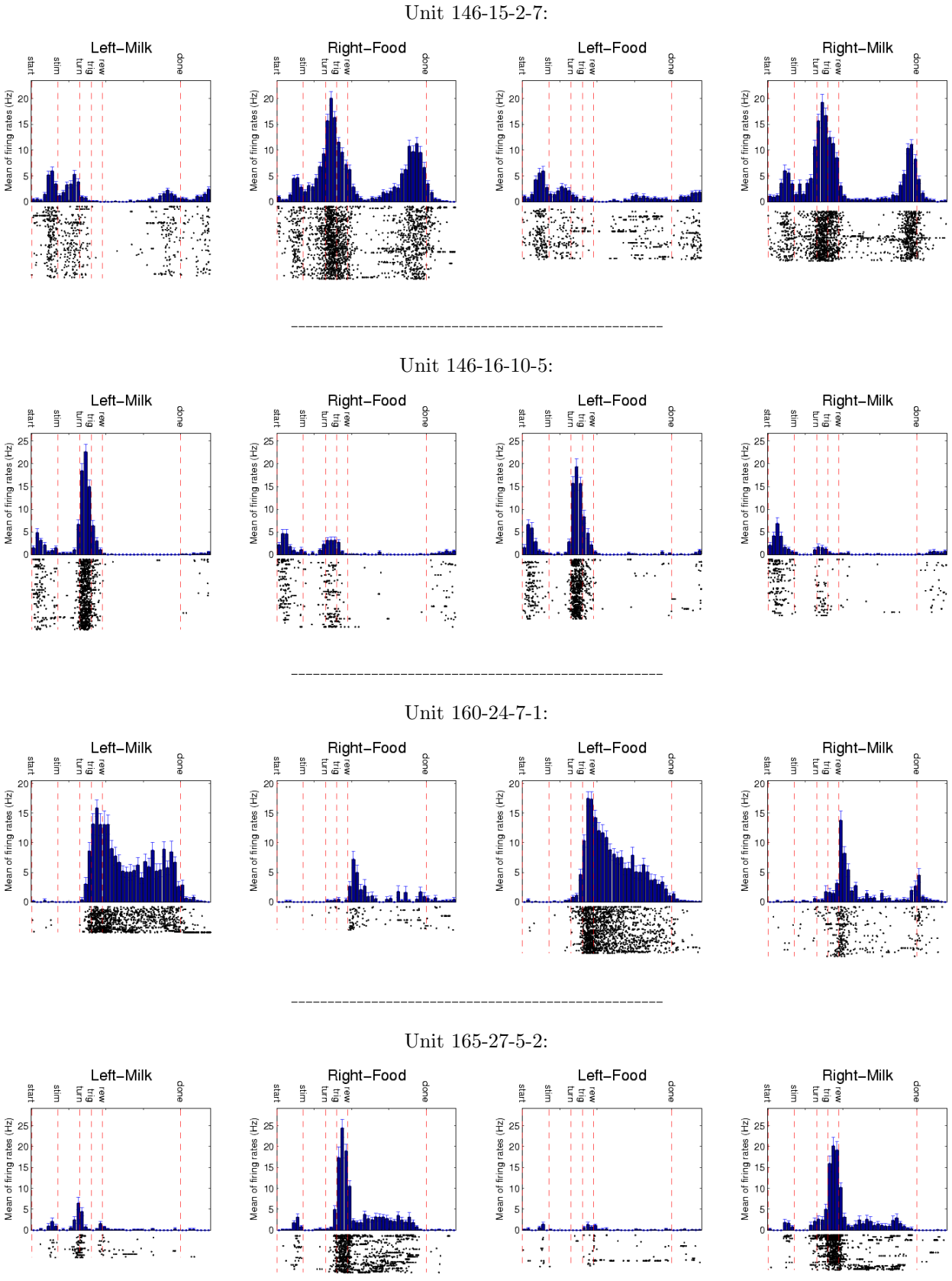

Figure 5.13: Plots of four neurons recorded from the dorsal striatum, showing correct trials subdivided by turn direction and reward type. As is typical of such neurons, the neurons followed the turn direction rather than the reward type on reversal.

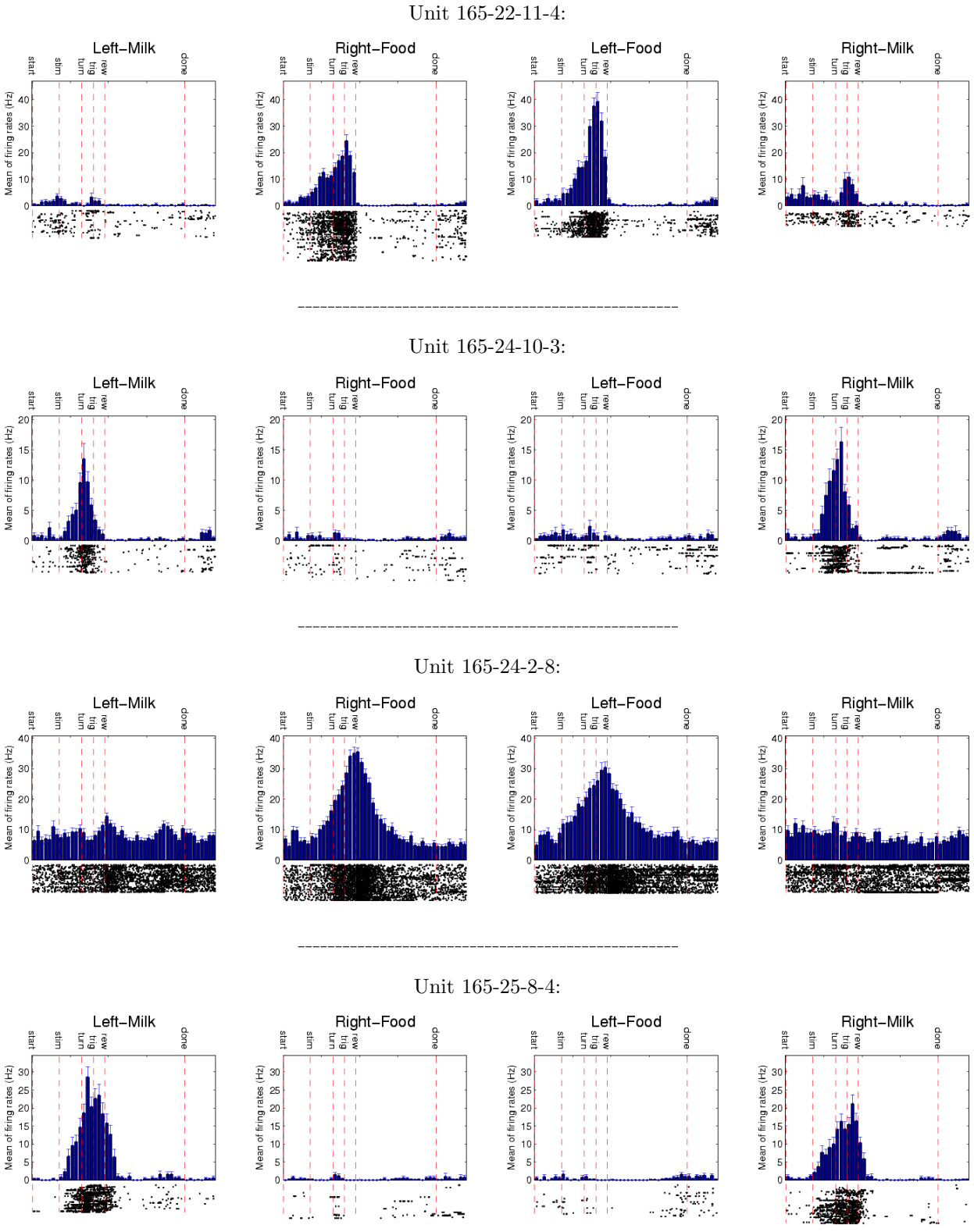

Figure 5.14: Plots of four neurons recorded from the ventral striatum, showing correct trials subdivided by turn direction and reward type. As is typical of such neurons, the neurons followed the reward type rather than the turn direction on reversal.

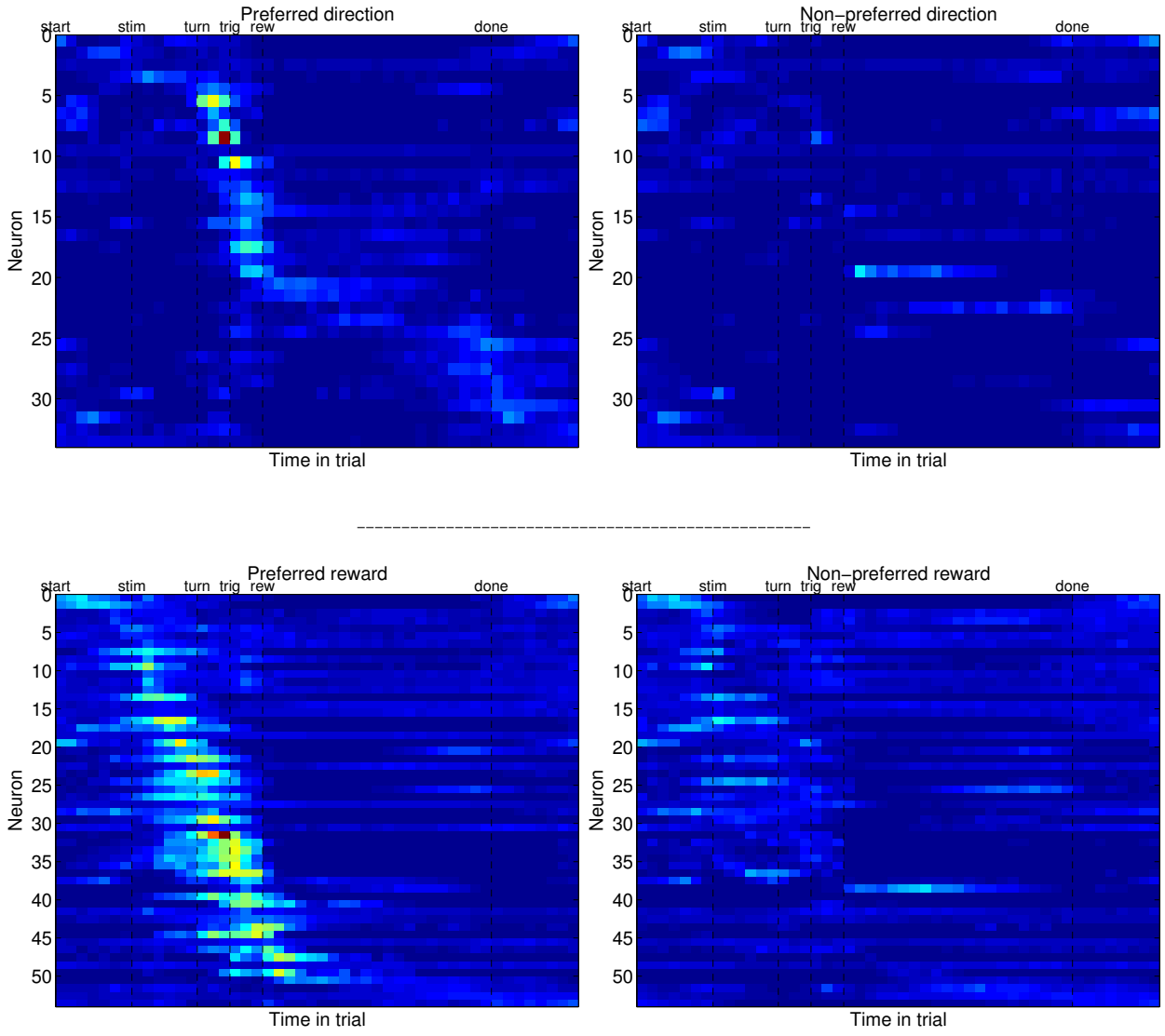

Figure 5.15: Population plots of all neurons that were significantly sensitive to either turn direction (top) or reward type (bottom). Separate traces are shown for trials with the preferred (left) versus non-preferred (right) direction or reward.

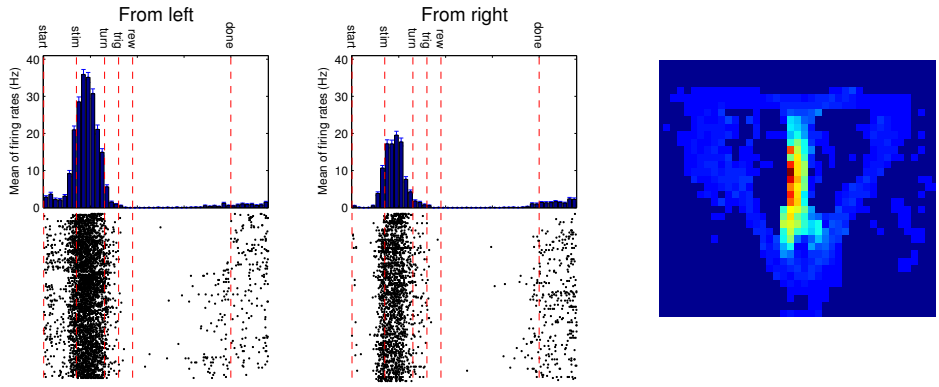

Figure 5.16: Left: Firing rate traces for neuron 146-16-10-4,broken down by trials in which the animal arrived at the T base from the left versus the right diagonal arm, this making a left versus a right turn to begin the next trial. The neuron fired preferentially during and after having made a left turn there. Right: Spatial firing rate plot for the same neuron, showing that the bulk of the firing occurs while the animal's head (where position is measured) has already completed the turn.

animals 146 and 160, so the possibility cannot yet be excluded that the difference is due to some uncontrolled difference between the subjects. One piece of evidence mitigates against this possibility, however: One tetrode from animal 165 was stuck in dorsal striatum and recorded two neurons that behaved like the dorsal neurons in the other animals. One of these neurons is illustrated in the bottom plot of Figure 5.13. Thus there is a small within-animal effect of depth. The other caveat about this finding is that in animal 165, food rewards were always signaled by tone and milk rewards by light. Thus the possibility cannot be eliminated that the ventral neurons are actually related to the stimulus type rather than the reward type. However, given that the response timings do not line up with the periods during which the stimuli were active, this seems unlikely.

Figure 5.13 displays firing traces for a number of the dorsal neurons considered above, showing correct trials broken down by turn direction and reward type. As is typical of dorsal neurons, on reversal, these units' firing preferences followed the turn direction rather than the reward type. (Though I speak of neuronal firing as being related to turns, in many cases it may be tied to other variables related to the animal's behavior, such as locomotion direction or spatial location. The general point is that such firing is related to some aspect of the animal's behavior, and not to expectations about reward.) Note also that these neurons are not related to the stimulus quality (light versus tone), since in animals 160 and 165, from whom the lower two traces were taken, the light always signaled chocolate milk and the tone food. Figure 5.14 illustrates some examples of contrasting neurons from the ventral striatum, which followed the rewards on reversal. Among these neurons is a fast-spiking tonic interneuron (unit 165-24-2-8), demonstrating that tonic neurons behave like their phasic neighbors on reversal.

Figure 5.15 shows population plots of all neurons that significantly distinguished either the turn directions or the rewards, made in the same manner as Figure 5.11. Separate plots are shown for trials with the preferred turn direction or reward versus the non-preferred one. Each firing rate trace was normalized by the total area under the histogram for the preferred direction or reward, so that, as in Figure 5.11, coloration in the preferred trial plots indicates degree of well-tunedness rather than absolute firing rate (with warm colors such as yellow and red marking the time bins to which a neuron is most sharply tuned). Since the plots for the non-preferred trials were normalized by the same factor, their paler coloration shows that firing rates were much lower during nonpreferred trials relative to preferred ones. Recall that these neurons were identified based on their firing between the points marked "stim" and "rew." As can be seen in the figure, some of the neurons shown are primarily tuned to some part of the trial outside this interval, but in these cases, the neurons showed a secondary response or background firing rates during the relevant period that significantly distinguished the turn directions or rewards.

Reward-related responses also tended to be more strongly *anticipatory* than the action-related responses.

As can be seen by comparing the time courses of the traces in Figures 5.13 and 5.14, or the top and bottom plots in Figure 5.15, action-related neurons often fire quite late in the trial (i.e. near reward delivery at the point marked "rew"), while neurons related to reward type often show ramping anticipatory firing that begins as soon as the stimulus is activated. Consistent with this observation, 30 of the 54 neurons related to reward type already showed significantly different firing rates for the two expected rewards using spike counts taken during the interval between stimulus onset and the animal reaching the top of the maze, well before the reward was actually delivered. In contrast, the spiking of only 1 of the 34 action-related neurons distinguished the turn direction in the same period. The remaining 33 neurons distinguished the turn direction during the period after the animal reached the top of the maze. Since this is around the time the turn is executed, these neurons might be classified as *reactive* rather than *anticipatory:* their firing is correlated with ongoing or past actions. Because it is difficult to define exactly when an extended behavior such as a turn begins and ends, it is not possible to definitively compare the time course of the behavior to the neuronal firing.

Further demonstrating reactive coding, action-related firing rates during the interval between the stimulus onset and the arrival at the top of the maze were more often correlated with direction of the *previous* turn (36/187 dorsal neurons), from the diagonal return path back up into the base of the T. It is important to again note that it is difficult to know exactly when the "turn" begins and ends: Position is tracked by head location, and for some of the time while the animal's head is ascending the T, his body is still bent around the corner, so the "turn" is thus in some sense still being executed. Figure 5.16 displays the time course of firing for a dorsal neuron that showed such a correlation, broken down by trials in which the animal made a left versus a right turn at the T base. (The neuron's firing did not distinguish left versus right turns at the top of the maze, though these data are not illustrated here.) Note that firing peaks after the stimulus onset, and the neuron does not completely cease firing until the reward triggers — after the *subsequent* turn is executed. The figure also displays a spatial firing rate map for the same neuron, which shows that the bulk of the firing occurs when the animal is ascending the T. (This was constructed by binning the animal's position on a 64x64 grid, and dividing the number of spikes that occurred while the animal was in each location by the amount of time the animal spent there.) While the timing of the turn is difficult to discern exactly, this plot makes clear that the peak response of the neuron lags the turn initiation, and that the response continues after the turn is surely complete.

5.5 Discussion

In this section, I have presented the results from a striatal recording experiment that used reversals in task contingencies to study what sort of information is represented by neuronal firing in different parts of the striatum, in order to explore more general ideas about the functional organization of decision making in the brain. The results demonstrate rather a different view of the population responsiveness than had been suggested by previous experiments in primates, and they demonstrate, preliminarily but for the first time, distinctions in neuronal firing properties between striatal subareas that reflect general notions about the organization of striatum derived from other experimental techniques. Here, in keeping with the basic focus of this thesis, I focus on broad considerations about the implications of these data for the organization of reinforcement learning in the brain.

One of the most interesting and novel results reported here is the finding, illustrated in Figure 5.11, that the preferred firing fields of striatal neurons, at least in the dorsal region, tile the entire task timeline fairly evenly. This is in contrast to to the predominant view from primate recordings (illustrated in Figure 2.3 on page 24, from Schultz et al., 1995), that neuronal responses fall into a few discrete classes, related to the three or four controlled events and actions in the task. This difference may very well be due to real differences in the patterns of results between this experiment and the primate ones, perhaps stemming from differences in the behavioral tasks used. The striatal recording experiments summarized by Schultz et al. (1995) were primate tasks involving very simple, punctate and well-controlled behaviors, such as a cued lever-press for juice reward. The present task is more fluid and complex, involving as it does continuous locomotion, and lends itself poorly to conception as a series of a few discrete events. Thus it is possible that the more continuously varying tuning patterns reported here are due to the more continuous task structure. The alternative is that the present results are due to the novel analysis methods used here, which could imaginably have turned up a similar result if applied to the monkey data. Given data resembling those

shown in Figure 5.11, but analyzing them with peri-event time histograms keyed to four or five discrete task events, it might be hard to detect a continuously varying pattern of tunings. In the present work, the use of a modified firing rate histogram to visualize the neuronal firing profile over a whole trial enabled the discovery this continuous variation in the preferred firing times of the neurons. A related point is that there has been a tendency in some previous striatal studies to assume that a neuron codes for the event that occurs closest to the time its firing peaks (though this is not universally true and some studies have addressed this criticism using devices like contingency reversals or unrewarded probe trials); for instance, neuronal firing around the time that food is delivered is often assumed to be reward-related while firing around the time of an action is assumed to be tied to decision-making. The use here of contingency reversals to more carefully determine the events to which firing is tied revealed a number of cases in which this assumption would have been erroneous, e.g. firing that peaked around time of the turn but was shown on reversal to follow the anticipated outcome rather than the action (Unit 165-24-10-3 in Figure 5.14).

The view of striatal representation derived from the primate experiments had been extended directly into theories mapping various reinforcement learning functions onto the striatal neurons. Specifically, it was proposed (e.g. by Houk et al., 1995) that firing around the time of an action supported competition between medium spiny neurons representing different stimulus-response habits — the actor half of an actor/critic system — and that anticipatory firing before rewards represented the value functions learned in the critic (e.g. Suri and Schultz, 2001). Finally, there is a stream of thinking going back to Montague et al. (1996; see also Dayan et al., 2000) that these decision-making and predictive/evaluative functions break down in the striatum along dorsal and ventral lines.

How do these ideas hold up in light of the present results? The results presented here support the idea that the ventral striatum specifically is involved in learning to predict future rewards, in that ventral neurons are likely to distinguish between different rewards but not different actions (Figure 5.14), and this firing is predictive in nature. The coding of reward predictions separately from action choices also supports the idea that reinforcement learning in the brain is organized along actor-critic lines rather than using a system like Q-learning that combines the functions of reward prediction and decision-making. The present experiment probed reward sensitivity by searching for neurons that responded differently to two different sorts of rewarding outcomes; this is different from the standard view of value functions in TD models, in which rewards of all sorts are collapsed to a single scalar measure of value. This is in keeping with a number of more recent TD theories of the dopamine system (e.g. the theory of Dayan, 2002, and the theory presented in Chapter 4) in which animals are assumed to learn about specific future outcomes as well as their generic rewarding consequences. There might very well also be neurons in the striatum representing a more traditional scalar value function; depending on the relative values of the food and the chocolate milk, the present analysis might not have flagged them. Also of course, a scalar value function can be arrived at by summing over outcome-specific values, so the predictive representation studied here could play a part in a TD system involving a scalar error signal; however, that error signal could not by itself be used to learn the vectorial predictions.

The results about the dorsal striatum are a bit more confusing in terms of the model. The fact that neurons recorded there tend to follow actions rather than rewards on reversal (Figure 5.13) would seem to support the theory that the dorsal striatum is involved in decision-making; however, the fact that the representation is not obviously anticipatory and in some cases seems even to lag the decisions (Figure 5.16), casts some doubt on this idea. Perhaps, however, neurons concerned with decision-making and control should not be strongly anticipatory; instead they should be separated from the actions they control with only minimal delay. It is also difficult in the present paradigm to define when the "turn" actually took place. Thus these doubts about response timings should not carry a great deal of weight. The results shown in Figure 5.11, coupled with the modeling in Chapter 4, do suggest an interesting alternative hypothesis, however: dorsal striatal cells might be involved in representing the *state* of the task. Though previous TD models have not paid a great deal of attention to where in the brain the task state was represented, Chapter 4 underlined the extent to which the state representation in a realistic TD system is a complex and abstract entity that requires a great deal of computation to derive. Insofar as the responses shown in Figure 5.11 tile the state space of the task quite effectively, the dorsal striatum would seem to be a reasonable candidate for this function. On this account, decisions would be made further downstream in the basal ganglia, or in motor cortex; striatal neurons would represent past actions with a slight lag since past actions are often a

relevant part of present behavioral state.

The results shown in Figure 5.11 differ strongly from those reported by Jog et al. (1999) in a similar task. They found that, while neurons in naive animals distributed their firing throughout a T-maze traversal, as the animals were further trained, striatal firing tended to cluster at the start and end of the trial, dramatically so for "overtrained" animals that had experienced about 500 trials of training. The result was interpreted in terms of a putative striatal involvement in the "chunking" or automization of behavioral sequences. In the present experiment, all animals were grossly overtrained in the sense of Jog et al. (1999), but no such clustering is seen. The seeming inconsistency may be due to a difference in the behavioral task, or in the recording site. In contrast to the present experiment, Jog et al. (1999) did not use a continuous T-maze — animals started behind a closed gate, were signaled to run the maze by the opening of the gate, and after receiving reward were ushered to retrace their steps to the start box, where the gate was closed. Thus, in this task, the "beginning" and "end" of a discrete trial are much more meaningful concepts than in the continuous, self-paced task described in this chapter, and it is possible that this distinction accounts for the differences in population sensitivity. The difference in recording sites is probably more important. The more lateral recording site of Jog et al. (1999) is in a part of the striatum more strongly associated with sequential behavior in lesion experiments (Cromwell and Berridge, 1996). Moreover, the "chunking" effect is not seen in preliminary results from a replication of the Jog et al. (1999) experiment conducted in the same laboratory but with a more medial recording site similar to the one used in this chapter (Kubota et al., 2002).

The present results may also seem inconsistent with a report from primate recordings by Hassani et al. (2001) of significant sensitivity to different sorts of fruit juice reward in neurons recorded throughout the striatum. This is due to a difference in analysis. I searched for neurons meeting a relatively strict definition of reward sensitivity: those that showed preferential firing on rightward versus leftward trajectories and then reversed that preference when the same trajectory was associated with the opposite reward. Obviously there are many weaker sorts of reward sensitivity; for instance, a neuron could fire on rightward trajectories, but more vigorously if that turn was rewarded with chocolate milk rather than food. Such a neuron has a degree of sensitivity to both the reward and the action, but my reversal test would classify it as an action-related neuron. Most of the neurons reported by Hassani et al. (2001) are modulated by both the action and the outcome in this manner. My own data also reveal many instances in which it seems likely that an action-related response was further modulated by the expected outcome, or vice versa. Because the overall firing rates of the neurons in my experiment are sometimes subject to unexplained variability, such modulation cannot be definitively demonstrated with the present data. A third phase of recording, in which the reversed contingencies were restored to their original state, would help with this. In any case, the strict test used here would seem to capture the primary correlate of the neuron, and the test also seems justified by the fact that it reveals a very clear distinction between the dorsal and ventral neurons. The results of Hassani et al. (2001) should remind us, however, that it would be improper to conclude from the data presented here that information about either rewards or actions is *entirely unrepresented* in any portion of the striatum.

Chapter 6

Concluding remarks

In this thesis I have presented a family of theoretical accounts of the dopamine system that extends previous theories in a number of directions. The work provides an enriched theoretical framework for understanding existing experimental data about the system, and for planning and analyzing future experiments. Most notably, richer theoretical models were introduced that provided correspondingly more accurate explanations for the responses of dopamine neurons. Chapter 3 introduced an average reward TD model that incorporated learning about long-run reward expectancies. The theory model was used to explain results about slow-timescale tonic responses in the system and to suggest a model of dopamine-serotonin interactions. Chapter 4 presented a broader family of TD models, incorporating partial observability and semi-Markov dynamics, and demonstrated that models much like the one studied in Chapter 3 are a special case of this family. The semi-Markov setting is congenial for studying situations where event timing can vary, and the new TD model was used to explain dopamine responses recorded in a number of such situations.

I have also explored several connections with behavioral data suggested by the new models, including in Chapter 3 with models of opponent interactions in conditioning and with tasks designed to probe optimal decision-making. The semi-Markov formalism used in Chapter 4 is particularly hospitable for analyzing several manifestations of timescale invariance in animal behavior. Also, the work in this thesis provides a basis for understanding how the dopamine neurons are part of a broader set of brain systems involved in learning, prediction, and decision-making. Theoretical ideas along these lines are presented in Chapters 3 and 4, and Chapter 5 presents new experimental results designed to probe the organization of these functions in the striatum, one of the dopamine system's chief targets. The experiment broadly supports the idea of "actor/critic" models that reward prediction and action selection are functionally and anatomically separate in the brain.

Finally, the work presented here paves the way for future theoretical and experimental investigations of a number of issues, some of which I now review.

6.1 Future directions

Many questions are raised by the theories presented here, which could fruitfully be addressed in future theoretical and experimental work.

First, the family of theories presented here that might explain published dopamine responses is broad; available data provide few strong constraints to help decide between the possibilities. New experimental work and new analysis of existing data are called for. On the issue of Markov versus semi-Markov valuation schemes, it would be important to record dopamine spiking and analyze baseline responding levels while manipulating reward rates (to provide a more direct test than the voltammetric data offer of the predictions of tonic responding discussed in Chapter 3). In addition, the semi-Markov scheme of Chapter 4 could be rather directly tested by analyzing how trial-by-trial dopamine responses covary with the delays preceding them; the data to allow this should largely already be available, though not published in a form that would allow the analysis. Another set of experiments suggested by Chapter 4 would investigate dopaminergic responding in situations designed to probe the system's handling of partial observability; along these lines

it would be interesting to place animals in situations with ambiguous reward expectancy and see how the dopamine neurons respond when the ambiguity is resolved in one direction or another. Analysis of between-neuron variability in this sort of a task might help resolve the speculation in Chapter 4 that the dopamine signal might be vector rather than scalar.

Questions in chapter 3 about how different sources of error are distributed between opponent error signals could best be resolved by systematically recording dopamine responses to negative and positive prediction triggered both directly (in appetitive and aversive conditioning experiments) and indirectly (in appetitive and aversive conditioned inhibition). Recordings of serotonin neurons during the same tasks, and particularly the aversive ones, are also warranted, given Chapter 3's plausible speculation that serotonin carries the negative prediction error channel.

The present work also suggests new directions for future modeling of dopamine and related brain systems. Notably, there are several further theoretical issues whose clarification might impact our understanding of the system and its behavioral implications. Chief among these are two components suggested by the discussion of partial observability in Chapter 4: Theories of decision-making in a partially observable setting and of learning an internal model of the world's hidden contingencies. The action selection problem is technically challenging and also hampered by a lack of dopamine recordings in a partially observable context (see above), but data recorded in other brain systems during decision making under sensory uncertainty (Gold and Shadlen, 2002), and also the data presented in Chapter 5, might be useful. Meanwhile, the problem of world modeling is more computationally straightforward, and future work in this area would build on strong precedents in both the neurophysiological (Lewicki, 2002; Lewicki and Olshausen, 1999) and behavioral domains (Courville and Touretzky, 2001; Courville et al., 2003). The study of world modeling in a partially observable semi-Markov setting would also help resolve significant questions, discussed in Section 4.5.2, about how such learning could be conducted in a timescale invariant manner, and thus help forge stronger connections with the behavioral literature on acquisition in classical conditioning.

Finally, all of these potential theoretical advances, and the ones already described in this thesis, crosscut another set of more empirical issues that could be addressed in future modeling. This involves developing a more detailed physiological theory of how all of these computations are implemented in neural tissue. Particularly interesting here are the questions of world modeling, state inference and representation, and action selection, which are the most abstract parts of the present framework. The data in Chapter 5 should be helpful in this regard, since it addresses both decision-making and prediction.

6.2 Summary of contributions

Here I briefly enumerate some of the major points made in the present work, highlighting the reasons why they constitute advances for our understanding of computation in the brain.

Overall contributions:

- Advances computational models of the dopamine system, a brain system thought to be important for learning, motivation and motor control and to such phenomena as drug addiction and Parkinson's disease. The general improvements fall along several dimensions:
  - The computational algorithms used to model the system are made more realistic by eliminating a number of simplifying assumptions and addressing a number of complexities ignored by previous models.
  - The work improves the fidelity of the model at explaining the known behaviors of dopamine neurons, offering new explanations for some phenomena, explaining other phenomena that were inconsistent with previous models, and making new predictions about the results of experiments and analyses not yet performed.
  - The work also enhances the connections between neurophysiological theories of dopaminergic function and psychological data about the behavior of animals in a variety of conditioning experiments, moving toward a unified theory of animal behavior and the brain systems that support it.

  - The reach of the theory is extended from an account of the dopamine neurons themselves to new or refined hypotheses about the functions of a number of brain areas with which the dopamine system interacts.

Chapter 3:

- Proposes a model of the dopamine system that relaxes the artificial assumption from most previous models that learning and prediction take place in a series of discrete trials, and investigates the implications of this correction for the expected behavior of dopamine neurons. While previous modelers had suggested adopting an infinite horizon return in place of the episodic one, none had previously noted that this change had any implications for the behavior of the modeled system. The present work unmasks these effects by making use of a version of the TD rule that is equivalent in a limit to the one considered by others, but which clearly segregates the effects of predictions outside the previous trial.

- Demonstrates that the inclusion of long-term predictions in the model introduces slowly changing background inhibition to the dopamine signal, thereby extending the TD account to incorporate slow-timescale ("tonic") dopamine behaviors. Slow timescale dopamine activity had previously been considered an entirely separate phenomenon, outside of the scope of existing computational theory. Moreover, confusion between tonic and phasic effects had sometimes been exploited to criticize computational theories of phasic dopamine activity. Tonic behaviors of dopamine neurons predicted by the new model include:

  - A tonic excitation in response to aversive events. Hints of excitation observed in slow neurochemical recordings (and more confusingly and controversially in unit recordings, which the present theory helps to clarify) had long been taken as a major obstacle to the reward-prediction models of the dopamine system. Prior to this work, this effect was the dopaminergic phenomenon that had remained most seriously anomalous under the TD models. The resolution proposed here clarifies matters, and taken together with work by Kakade and Dayan (2001a, 2002b), helps to dislodge a persistent strain of theory that had held that dopamine should be viewed as an exclusively attentional rather than reward-related signal.
  - Habituation effects in dopamine release. These had been observed in fast neurochemical recordings at dopamine target structures, but never given a satisfactory computational explanation.
  - Novel predictions about dopamine behavior in several experiments yet to be conducted.

- Establishes parallels between the present computational theory and the *opponency* theory of Solomon and Corbit (1974), a classic in psychology. The original psychological theory, while compelling, was essentially phenomenological; there was little attempt to explain the described behaviors in terms of rational computation and no notion of the brain systems subserving the phenomena. By studying Solomon and Corbit (1974)-like phenomena in a TD system, the present work also clarifies the relationship between opponency in their sense, and a different sort of opponency that is also ubiquitous in psychological theories, that between systems associated with appetitive and aversive events.

- Leverages these ideas about opponency, together with data on how the dopamine and serotonin systems interact, to extrapolate the TD theory of dopamine function into a proposed theory of the function of the serotonergic system of the dorsal raphe nucleus (one of several serotonergic systems in the brain). Though this work is quite speculative, it is among the first detailed computational theories of serotonergic function ever put forward, an advance made possible by the novel approach of supplementing consideration of the relatively poor data on serotonin with the relatively well understood data on dopamine. The dorsal raphe system's proposed role as an opponent, aversive channel in the TD model allows the model to incorporate a number of explanations for behavioral phenomena from opponent psychological theories, including for instance ideas about conditioned inhibition and extinction.

- Investigates the implications of data about animal foraging and decision behaviors for TD models of the dopamine system. Notably, results from many decision tasks support the dominant psychological theory that animals discount delayed rewards hyperbolically in their delays, which might be seen as a

challenge to the ubiquity of exponential discounting in reinforcement learning models. However, here I demonstrate that an exponentially discounted TD model that incorporates reasonable assumptions about measurement noise in animals' interval timing processes can quantitatively account for decision data regarding delay discounting as well as animal sensitivity to variability in the delays and amounts of expected rewards.

Chapter 4:

- Proposes a model of the dopamine system that incorporates sound reasoning about uncertainty in the timing of events and in the underlying situation ("state of the world") relevant to prediction, and which, additionally, is agnostic with respect to the timescale of events ("timescale invariant"). The new model:

  - Embodies a novel hypothesis about the interactions between a cortical sensory/representational system whose job is to model and infer the state of the world, and a dopaminergic reward-learning system that predicts integrated future values in this inferred state space. The algorithm proposed for this interaction combines semi-Markov dynamics and partial observability, and model-based and model-free reinforcement learning techniques.
  - Explains dopamine behavior on a variety of experiments with programmed variation in the timing between events. Most such experiments had not been convincingly explained by previous models.
  - Explores the effects on the dopamine response of *subjective* timing variability due to noise in the animal's time measurement processes, which had not previously been considered in dopaminergic models. Together with the partial rectification of negative TD error in the dopaminergic signal, this feature allows the model to explain why dopaminergic responses to reward can transfer to a stimulus that reliably predicts it with a short, but not a long, deterministic latency. Similarly, it can explain the previously unexplained phenomenon of the extinction of all dopaminergic responding in overtrained animals despite small programmed variation in the interval between trials.
  - Predicts a testable pattern by which trial-by-trial variation in the magnitude of the dopaminergic response to an event should correlate with the length of the delay preceding the event. This pattern is present in most common situations in which dopamine neurons are excited, and embodies a novel (but self-consistent) explanation for why a response is observed asymptotically in all these cases. The pattern results from a semi-Markov method of accounting for the costs of delays, which is an alternative to the Markov account detailed in Chapter 3. Though it is not yet possible to definitively rule out either possibility, the distinction developed here should guide future experimentation and analysis.
  - Forges connections with a large behavioral literature on variability in animals' timed behaviors that had not previously been considered from the perspective of models of this sort.
  - Provides a framework for studying timescale invariance properties in animal learning, and suggests a way that existing models of these phenomena could be integrated into the present theory.

Chapter 5:

- Presents the results of a recording study designed to study the organization of decision-making in rat striatum by dissociating neuronal firing related to action choice or execution from firing related to the expected outcomes of actions.

- Presents data about the firing properties of the population, suggesting that as a group striatal neurons' preferred firing times are distributed rather evenly throughout the entire execution of the task, rather than being specialized toward representing a few key events like cues and rewards, as previously had been thought.

- Presents preliminary data showing that neurons in the dorsal striatum seem most often to be representing information about the animal's actions, while neurons in the ventral striatum usually represent the expected outcomes of those actions. While such a distinction is to be expected based on data from other experimental methods, this functional organization had not previously been well demonstrated in a neuronal recording study.
- Interprets these data as supporting theories in which the brain makes decisions using an actor/critic system, in which behavioral responses and their expected outcomes are learned separately.

6.3 Summary of modeling results

In this section, I review the major modeling results in this thesis. In particular, I list the major experiments simulated here, with reference to the original data and to my simulations. In this section, I do not review predictions from simulations about experiments that have not been performed, and I also in all but a few important cases omit mention of experiments whose explanation under the models is discussed but not explicitly simulated in this thesis. Finally, I do not review results from Chapter 5, since these are from experiment rather than modeling.

As this thesis contains several variations on the TD model, I also note which version was used in each simulation. The major variations on the model are referenced by number below:

1. The average reward TD model (Equation 3.3 on page 45);
2. The same model, with the error signal divided in various ways between putative dopaminergic and serotonergic subsignals (Equations 3.5 through 3.8 on page 56);
3. The single-trial exponentially discounted return incorporating temporal uncertainty (Equations 3.16 and 3.17 on page 69), used for simulating behavioral choices;
4. The partially observable semi-Markov TD model (Equation 4.3 on page 89); and
5. The fully observable limit of this model, Equation 4.1.

Here, then, are the experiments modeled in this thesis:

Basic phasic dopamine responses:

- Dopamine neurons respond to unpredicted rewards, and the response transfers to stimuli that reliably predict reward (Schultz, 1998; data reproduced in Figure 2.2 on page 16). These phenomena are modeled using the average reward model (Model 1) in Figure 3.2 on page 46, using the version including serotonin (Model 2) in Figures 3.10 on page 57 and 3.11 on page 58, and using the semi-Markov TD model (Model 5) in Figures 4.10 on page 96 and 4.13 on page 99.
- When a predicted reward is omitted, dopamine neurons exhibit a well-timed pause in their background firing (Schultz, 1998; data reproduced in Figure 2.2 on page 16). This phenomenon is modeled using the average reward model (Model 1) in Figure 3.2 on page 46, using the version including serotonin (Model 2) in Figures 3.11 on page 58 and 3.12 on page 58, and using the partially observable semi-Markov TD model (Model 4) in Figure 4.16 on page 103.
- Only 50-75% of recorded dopamine neurons display any particular class of phasic response (Schultz, 1998). As discussed in Section 4.3.5 on page 91, Model 5 explains this result if we assume a particular form for the way that dopamine neurons code the vector error signal.

Phasic responses in a conditioned inhibition experiment

- Dopamine neurons respond with inhibition (sometimes preceded by excitation due to overgeneralization) to a conditioned inhibitor presented alone, but subsequently show no response at the time a reward would have been omitted (Tobler et al., 2001). These data are reproduced using a variation on the dopamine/serotonin model (Model 3), in Figure 3.13 on page 59.

Phasic responses when event timing can vary

- When the timing of a reward (Fiorillo and Schultz, 2001) or reward-predictive event (Schultz et al., 1993) is strongly unpredictable on the basis of previous events, dopamine neurons respond to it asymptotically (data reproduced in Figures 4.1 on page 75 and 4.2 on page 76). Versions of this effect are simulated using Model 5 in Figures 4.10 on page 96, 4.12 on page 99, and 4.13 on page 99.

- When the reward or stimulus timing varies, but only slightly, dopamine neurons do not respond to the event (W. Schultz, personal communication, 2003; also data from Ljungberg et al., 1992, reproduced in Figure 4.1 on page 75, may be an example of this). Versions of this phenomenon are simulated using Model 5 in Figure 4.11 on page 97 and 4.13 on page 99.

- When the stimulus-reward delay is long but deterministic, the dopamine response to a reward cannot be trained away (C. Fiorillo, personal communication, 2002). Model 5 explains this as resulting from subjective variability in the interevent timing due to time measurement noise (Figure 4.13 on page 99).

- In an experiment by Hollerman and Schultz (1998), dopamine neurons responded to a reward delivered earlier-than-usual in probe trials, but did not subsequently pause at the time reward was originally expected. These results are simulated using Model 4 in Figure 4.17 on page 103.

Magnitude of phasic responding

- In a partial reinforcement experiment, phasic dopamine responses to the stimulus are graded by likelihood of reinforcement, and responses to the reward are graded inversely by the extent to which the reward was expected given the prior stimulus (Fiorillo et al., 2003; data reproduced in Figure 4.20). These data are modeled using the partially observable semi-Markov model (Model 4) in Figure 4.18 on page 106.

- The dopaminergic response to some reward or reward omission is proportional to the the estimated degree of prediction error engendered by that event, though the proportionality is truncated for large, negative prediction errors (Bayer and Glimcher, 2002; data reproduced in Figure 3.9 on page 55). As discussed in Section 3.4.6 on page 60, Model 3 explains this pattern of results given a particular assumption about how the signal is divided between dopaminergic and serotonergic channels.

- In an experiment by Fiorillo and Schultz (2001) dopamine responses to a reward whose timing could vary were largest when the reward occurred earliest, and declined for later rewards. These data are simulated using models 4 and 5 in Figure 4.15 on page 102.

Tonic dopamine responses

- When dopamine concentrations are measured using voltammetry at target areas, the response to repeated stimulation habituates with a slow timescale (Kilpatrick et al., 2000) and the level of habituation varies inversely with the rate of stimulation (Montague et al., 2003; both sets of data are reproduced in Figure 3.6 on page 50). These results are modeled in Figure 3.3 on page 47, using the average reward model (Model 1).

- Dopamine concentrations, measured using slow neurochemical means, rise during aversive stimulation such as footshock (e.g. Abercrombie et al., 1989), and there are also some indications from recordings of dopamine neurons spiking of a slow response in similar situations (Schultz and Romo, 1987; Guarraci and Kapp, 1999; data reproduced in Figure 3.7 on page 52). These results are modeled in Figure 3.4 on page 48, using the average reward model (Model 1).

- Background dopamine neuron firing rates show a sustained elevation during the period between stimulus and reward in a partial reinforcement task, in delay but evidently not trace conditioning (Fiorillo et al., 2003; data reproduced in Figure 4.20). The response is largest when the reinforcement fraction is near 50%. These data are modeled using the partially observable semi-Markov model (Model 4) in Figure 4.18 on page 106.

Behavioral results

- When asked to choose between large and small rewards after different delays, animals seem to choose in accord with a single-trial hyperbolically discounted return (e.g. Mazur, 1987). The data are reproduced, and modeled using Model 3, in Figure 3.16 on page 68.
- When asked to choose between two possible rewards whose timings may vary, animals seem to choose in accord with the "expectation of the ratios" (Equation 2.17 on page 36), i.e. the expected hyperbolically discounted reward magnitude (e.g. Mazur, 1984). The data are reproduced, and modeled using Model 3, in Figure 3.17 on page 69.
- Many behavioral measures of affective response to the onset and offset of motivationally significant events (such as heart rate under electric shock) display a characteristic pattern of response, habituation, rebound, and re-habituation (Solomon and Corbit, 1974). This pattern is reproduced using a low-pass filtered version of Model 1, in Figure 3.8 on page 54.
- Trial-to-trial variability in animals' timed behaviors is *scalar* in the sense that the standard deviation of response time scales linearly with the delay being timed (Gibbon, 1977). As discussed in Section 4.5.1 on page 110, it is easy to incorporate such variability in Models 4 and 5.
- Acquisition of a conditioned response in a delay conditioning experiment is *timescale invariant;* i.e., the number of trials to acquisition is invariant to contractions or dilations in the timescale of events (Gallistel and Gibbon, 2000). As discussed in Section 4.5.2, this basic property is also true of Model 4, and it seems likely that in future work Model 5 could be augmented with a module for learning its inference model that similarly respected timescale invariant acquisition.

Bibliography

E. D. Abercrombie, K. A. DiFrischia, and M. J. Zigmond. Differential effect of stress on in vivo dopamine release in striatum, nucleus accumbens, and medial frontal cortex. *Journal of Neurochemistry*, 52:1655–1658, 1989.

C. D. Adams. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 34B:77–98, 1982.

C. D. Adams and A. Dickinson. Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 33B:109–112, 1981.

G. K. Aghajanian and G. J. Marek. Serotonin and hallucinogens. *Neuropsychopharmacology*, 21:16S–23S, 1999.

G. E. Alexander, M. R. Delong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381, 1986.

T. Aosaki, M. Kimura, and A. M. Graybiel. Temporal and spatial characteristics of tonically active neurons of the primate's striatum. *Journal of Neurophysiology*, 73:1234–1252, 1995.

T. Aosaki, H. Tsubokawa, A. Ishida, K. Watanabe, A. M. Graybiel, and M. Kimura. Reponses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *Journal of Neuroscience*, 14:3969–3984, 1994.

P. Apicella, S. Ravel, P. Sardo, and E. Legallet. Influence of predictive information on responses of tonically active neurons in the monkey striatum. *Journal of Neurophysiology*, 80:3341–3344, 1998.

P. Apicella, E. Scarnati, T. Ljungberg, and W. Schultz. Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *Journal of Neurophysiology*, 68:945–960, 1992.

L. C. Baird. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of the International Conference on Neural Networks, Orlando, FL, June*, 1994.

S. Bao, V. T. Chan, and M. M. Merzenich. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, 412:79–83, 2001.

A. Barto and M. Duff. Monte Carlo matrix inversion and reinforcement learning. In *Advances in Neural Information Processing Systems 6*, pages 687–694, 1994.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:834–46, 1983.

A.G. Barto and R.S. Sutton. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research*, 4:221–235, 1982.

M. Bateson and A. Kacelnik. Preferences for fixed and variable food sources: Variability in amount and delay. *Journal of the Experimental Analysis of Behavior*, 63:313–329, 1995.

M. Bateson and A. Kacelnik. Rate currencies and the foraging starling: The fallacy of the averages revisited. *Behavioral Ecology*, 7:341–352, 1996.

L. E. Baum, T. Petrie, G. Soulds, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

W. M. Baum. Optimization and the matching law as accounts of instrumental behavior. *Journal of the Experimental Analysis of Behavior*, 36:387–403, 1981.

H. M. Bayer and P. W. Glimcher. Subjective estimates of objective rewards: Using economic discounting to link behavior and brain. In *Society for Neuroscience Abstracts*, volume 28: 358.6, 2002.

R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, N.J., 1957.

K. C. Berridge and T. E. Robinson. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28:309–369, 1998.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

C. M. Bradshaw and E. Szabadi. Choice between delayed reinforcers in a discrete-trials schedule. *Quarterly Journal of Experimental Psychology*, 44B:1–16, 1992.

S. J. Bradtke and M. O. Duff. Reinforcement learning methods for continuous-time Markov Decision Problems. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 393–400, Cambridge, MA, 1995. MIT Press.

T. S. Braver, D. M. Barch, and J. D. Cohen. Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, 46:312–328, 1999.

W. J. Brogden. Sensory preconditioning. *Journal of Experimental Psychology*, 25:323–332, 1939.

J. Brown, D. Bullock, and S. Grossberg. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, 19(23): 10502–10511, 1999.

H. T. Chang C. J. Wilson and S. T. Kitai. Firing patterns and synaptic potentials of identified giant aspiny interneurons in the rat neostriatum. *Journal of Neuroscience*, 10:508–519, 1990.

T. Caraco, W. U. Blanckenhorn, G. M. Gregory, J. A. Newman, G. M. Recer, and S. M. Zwicker. Risk-sensitivity: Ambient temperature affects foraging choice. *Animal Behavior*, 39:338–345, 1990.

R. Cardinal, N. D. Daw, T. W. Robbins, and B. J. Everitt. Local analysis of behavior in the adjusting delay task for assessing choice of delayed reinforcement. *Neural Networks*, 15:603–616, 2002.

R. N. Cardinal, T. R. Robbins, and B. J. Everitt. Choosing delayed rewards: Perspectives from learning theory, neurochemistry and neuroanatomy. In *Choice, Behavioral Economics and Addiction*. Elsevier, 2003. (in press).

R. N. Cardinal, T. W. Robbins, and B. J. Everitt. The effects of d-amphetamine, chlordiazepoxide, alpha-flupenthixol, and behavioural manipulations on choice of signaled and unsignaled delayed reinforcement in rats. *Psychopharmacology*, 152:362–375, 2000.

R. M. Carelli and S. A. Deadwyler. Cellular mechanisms underlying reinforcement-related processing in the nucleus accumbens: Electrophysiological studies in behaving animals. *Pharmacology, Biochemistry and Behavior*, 57:495–504, 1997.

K. Chergui, M. F. Suaud-Chagny, and F. Gonon. Nonlinear relationship between impulse flow, dopamine release and dopamine elimination in the rat brain in vivo. *Neuroscience*, 62:641–645, 1994.

L. Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *National Conference on Artificial Intelligence*, pages 183–188, 1992.

R. M. Church. Evaluation of quantitative theories of timing. *Journal of the Experimental Analysis of Behavior*, 71:253–256, 1999.

R. M. Church and M. Z. Deluty. Bisection of temporal intervals. *Journal of Experimental Psychology: Animal Behavior Processes*, 3:216–228, 1977.

R. M. Church and J. Gibbon. Temporal generalization. *Journal of Experimental Psychology: Animal Behavior Processes*, 8:165–186, 1982.

A. C. Courville, N. D. Daw, G. J. Gordon, and D. S. Touretzky. Model uncertainty in classical conditioning. 2003. Submitted to NIPS 2003.

A. C. Courville and D. S. Touretzky. Modeling temporal structure in classical conditioning. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 3–10, Cambridge, MA, 2001. MIT Press.

H. C. Cromwell and K. C. Berridge. Implementation of action sequences by a neostriatal site: A lesion mapping study of grooming syntax. *Journal of Neuroscience*, 16:3444–3458, 1996.

T. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck. Solving semi-Markov decision problems using average reward reinforcement learning. *Management Science*, 1999.

N. D. Daw, A. C. Courville, and D. S. Touretzky. Dopamine and inference about timing. In *Proceedings of the Second International Conference on Development and Learning*, pages 271–276. IEEE Computer Society, 2002a.

N. D. Daw, A. C. Courville, and D. S. Touretzky. Timing and partial observability in the dopamine system. In *Advances in Neural Information Processing Systems 15*, 2003. in press.

N. D. Daw, S. Kakade, and P. Dayan. Opponent interactions between serotonin and dopamine. *Neural Networks*, 15:603–616, 2002b.

N. D. Daw and D. S. Touretzky. Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, 32:679–684, 2000.

N. D. Daw and D. S. Touretzky. Operant behavior suggests attentional gating of dopamine system inputs. *Neurocomputing*, 38:1161–1167, 2001.

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

P. Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.

P. Dayan. Improving generalisation for temporal difference learning: The successor representation. *Neural Computation*, 5:613–624, 1993.

P. Dayan. Computational modelling. *Current Opinion in Neurobiology*, 4:212–217, 1994.

P. Dayan. Motivated reinforcement learning. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 11–18, Cambridge, MA, 2002. MIT Press.

P . Dayan and T. J. Sejnowski. Exploration bonuses and dual control. *Machine Learning*, 25:5–22, 1996.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA, 2001.

P. Dayan and B. W. Balleine. Reward, motivation and reinforcement learning. *Neuron*, 36:285–298, 2002.

P. Dayan, S. Kakade, and P. R. Montague. Learning and selective attention. *Nature Neuroscience*, 3: 1218–1223, 2000.

P. Dayan and T. Long. Statistical models of conditioning. In M. J. Kearns M. I. Jordan and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 117–123. MIT Press, 1998.

J. F. W. Deakin. Roles of brain serotonergic neurons in escape, avoidance and other behaviors. *Journal of Psychopharmacology*, 43:563–577, 1983.

J. F. W. Deakin and F. G. Graeff. 5-ht and mechanisms of defence. *Journal of Psychopharmacology*, 5: 305–316, 1991.

M. R. DeLong, M. D. Crutcher, and A. P. Georgopoulos. Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. *Journal of Neuroscience*, 3:1599–1606, 1983.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

J. C. Denniston, H. I. Savastano, and R. R. Miller. The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer and S. B. Klein, editors, *Handbook of Contemporary Learning Theories*, pages 65–117. Erlbaum, Hillsdale, NJ, 2001.

M. J. Detke. Extinction of sequential conditioned inhibition. *Animal Learning and Behavior*, 19:345–354, 1991.

R. L. DeValois, I. Abramov, and G. H. Jacob. Analysis of response patterns of lgn cells. *Journal of the Optical Society of America*, 56:966–977, 1966.

P. B. Dews. The theory of fixed-interval responding. In W. N. Schoenfeld, editor, *The Theory of Reinforcement Schedules*. Appleton-Century-Crofts, New York, 1970.

A. Dickinson. Actions and habits — the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B — Biological Sciences*, 308:67–78, 1985.

A. Dickinson. Bolles' psychological syllogism. In M. E. Bouton and M. S. Fanselow, editors, *Learning, Motivation and Cognition*, pages 345–367. American Psychological Association, Washington, D.C., 1987.

A. Dickinson. Instrumental conditioning. In N. J. Mackintosh, editor, *Animal Learning and Cognition*, pages 45–79. Academic Press, San Diego, 1994.

A. Dickinson and B. Balleine. The role of learning in motivation. In C. R. Gallistel, editor, *Stevens' Handbook of Experimental Psychology (3rd ed.) Vol. 3: Learning, Motivation and Emotion*. Wiley, New York, 2002.

A. Dickinson and G. R. Dawson. Motivational control of instrumental performance: The role of prior experience with the reinforcer. *Quarterly Journal of Experimental Psychology*, 40B:113–134, 1988.

A. Dickinson and M. F. Dearing. Appetitive-aversive interactions and inhibitory processes. In A. Dickonson and R. A. Boakes, editors, *Mechanisms of Learning and Motivation*, pages 203–231. Lawrence Erlbaum, Hillsdale, N.J., 1979.

K. Doya. Metalearning and neuromodulation. *Neural Networks*, 15:495–506, 2002.

W. K. Estes. Discriminative conditioning: Ii. effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *Journal of Experimental Psychology*, 38:173–177, 1948.

B. J. Everitt and T. W. Robbins. Amygdala-ventral striatal interactions and reward-related processes. In J. P. Aggleton, editor, *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pages 401–429. Wiley-Liss, New York, 1992.

C. D. Fiorillo and W. Schultz. The reward responses of dopamine neurons persist when prediction of reward is probabilistic with respect to time or occurrence. In *Society for Neuroscience Abstracts*, volume 27: 827.5, 2001.

C. D. Fiorillo, P. N. Tobler, and W. Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299:1898–1902, 2003.

P. J. Fletcher. Conditioned place preference induced by microinjection of 8-oh-dpat into the dorsal or median raphe nucleus. *Psychopharmalogy*, 113:31–36, 1993.

P. J. Fletcher. Effects of combined or separate 5,7-dihydroxytryptamine lesions of the dorsal and median raphe nuclei on responding maintained by a drl 20s schedule of food reinforcement. *Brain Research*, 552: 219–245, 1995.

P. J. Fletcher and K. M. Korth. Activation of 5-ht1b receptors in the nucleus accumbens reduces amphetamine-induced enhancement of responding for conditioned reward. *Psychopharmacology*, 142:165–174, 1999.

P. J. Fletcher, K. M. Korth, and J. W. Chambers. Selective destruction of brain serotonin neurons by 5,7-dihydroxytryptamine increases responding for a conditioned reward. *Psychopharmacology*, 147:291–299, 1999.

K. J. Friston, G. Tononi, G. N. Reeke Jr., O. Sporns, and G. M. Edelman. Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, 59:229–243, 1994.

C. R. Gallistel. Can a decay process explain the timing of conditioned responses? *Journal of the Experimental Analysis of Behavior*, 71:264–271, 1999.

C. R. Gallistel and J. Gibbon. Time, rate and conditioning. *Psychological Review*, 107(2):289–344, 2000.

R. Ganesan and J. M. Pearce. Effect of changing the unconditioned stimulus on appetitive blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, 14:280–291, 1988.

K. Gao, D. O. Chen, J. R. Genzen, and P. Mason. Activation of serotonergic neurons in the raphe magnus is not necessary for morphine analgesia. *Journal of Neuroscience*, 18:1860–1868, 1998.

K. Gao, Y. H. Kim, and P. Mason. Serotonergic pontomedullary neurons are not activated by antinociceptive stimulation in the periaqueductal gray. *Journal of Neuroscience*, 17:3285–3292, 1997.

P. A. Garris, J. R. C. Christensen, and R. M. Wightman. Real-time measurement of electrically evoked extracellular dopamine in the striatum of freely moving rats. *Jorunal of Neurochemistry*, 68:152–161, 1997.

P. A. Garris, M. Kilpatrick, M. A. Bunin, D. Michael, Q. D. Walker, and R. M. Wightman. Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation. *Nature*, 398:67–69, 1999.

P. A. Garris and R. M. Wightman. Different kinetics govern dopaminergic transmission in the amygdala, prefrontal cortex, and striatum: An in vivo voltammetric study. *Journal of Neuroscience*, 14:442–450, 1994.

I. Geller and J. Seifter. The effects of mepobromate, barbutrate, d-amphetamine and promazine on experimentally induced conflict in the rat. *Psychopharmacology*, 1:482–492, 1960.

Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–273, 1997.

J. Gibbon. Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, 84:279–325, 1977.

J. Gibbon. On the form and location of the psychometric bisection function for time. *Journal of Mathematical Psychology*, 24:58–87, 1981.

J. Gibbon. Ubiquity of scalar timing with a Poisson clock. *Journal of Mathematical Psychology*, 36:283–293, 1992.

J. Gibbon and R. M. Church. Time left: Linear versus logarithmic subjective time. *Journal of Experimental Psychology: Animal Behavior Processes*, 7:87–108, 1981.

J. Gibbon and R. M. Church. Sources of variance in an information processing theory of timing. In H. L. Roitblat, T. G. Bever, and H. S. Terrace, editors, *Animal Cognition*, pages 465–488. Erlbaum, Hillsdale, NJ, 1984.

J. Gibbon, C. Malapani, C. L. Dale, and C. R. Gallistel. Toward a neurobiology of temporal cognition: Advances and challenges. *Current Opinion in Neurobiology*, 7:170–184, 1997.

M. A. Gluck and C. Myers. Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3:491–516, 1993.

J. I. Gold and M. N. Shadlen. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.

J.H. Goodman and H. Fowler. Blocking and enhancement of fear conditioning by appetitive CSs. *Animal Learning and Behavior*, 11:75–82, 1983.

F. G. Graeff. On serotonin and experimental anxiety. *Psychopharmacology*, 163:467–476, 2003.

F. G. Graeff and N. G. Silveira Filho. Behavioral inhibition induced by electrical stimulation of the median raphe nucleus of the rat. *Physiology and Behavior*, 21:477–484, 1978.

L. Green, N. Fristoe, and J. Myerson. Temporal discounting and preference reversals in choice between delayed outcomes. *Psychnomic Bulletin and Review*, 1:383–389, 1994.

S. Grossberg. Some normal and abnormal behavioral syndromes due to transmitter gating of opponent processes. *Biological Psychiatry*, 19:1075–1118, 1984.

S. Grossberg, editor. *Neural Networks and Natural Intelligence*. MIT Press, Cambridge, MA, 1988.

S. Grossberg. The imbalanced brain: From normal behavior to schizophrenia. *Biological Psychiatry*, 48: 81–98, 2000.

S. Grossberg and N. A. Schmajuk. Neural dynamics of attentionally modulated Pavolivan conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, 15:195–240, 1987.

S. Grossberg and N. A. Schmajuk. Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, 2:79–102, 1989.

F.A. Guarraci and B.S. Kapp. An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential Pavlovian fear conditioning in the awake rabbit. *Behavioral Brain Research*, 99:169–179, 1999.

Y. Guedon and C. Cocozza-Thivent. Explicit state occupancy modeling by hidden semi-Markov models: Application of Derin's scheme. *Computer Speech and Language*, 4:167–192, 1990.

O. K. Hassani, H. C. Cromwell, and W. Schultz. Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *Journal of Neurophysiology*, 85:2477–2489, 2001.

M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

D. J. Heeger, G. M. Boynton, J. B. Demb, E. Seidemenn, and W. T. Newsome. Motion opponency in visual cortex. *Journal of Neuroscience*, 19:7162–7174, 1999.

R. J. Herrnstein. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4:267–272, 1961.

R. J. Herrnstein. Aperiodicity as a factor in choice. *Journal of the Experimental Analysis of Behavior*, 7: 179–182, 1964.

R. J. Herrnstein. On the law of effect. *Journal of the Experimental Analysis of Behavior*, 13:243–266, 1970.

R. J. Herrnstein. Experiments on stable suboptimality in individual behavior. *American Economic Review*, 81:360–364, 1991.

R. J. Herrnstein and W. Vaughan. Melioration and behavioral allocation. In J. E. R. Staddon, editor, *Limits to Action: The Allocation of Individual Behavior*, pages 143–176. Academic Press, New York, 1980.

P. C. Holland. Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. *Animal Learning and Behavior*, 25:373–399, 1997.

J. R. Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309, 1998.

J. R. Hollerman, L. Tremblay, and W. Schultz. Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology*, 80:947–963, 1998.

J. Horvitz, T. Stewart, and B. L. Jacobs. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research*, 759(2):251–258, 1997.

J. C. Horvitz. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96:651–656, 2000.

J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. MIT Press, Cambridge, MA, 1995.

L. M. Hurvich and D. Jameson. An opponent-process theory of color vision. *Psychological Review*, 64: 384–404, 1957.

S. Ikemoto and J. Panksepp. The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews*, 31:6–41, 1999.

H. Imai, D. A. Steindler, and S. T. Kitai. The organization of divergent axonal projections from the midbrain raphe nuclei in the rat. *Journal of Comparative Neurology*, 243:363–380, 1986.

B. L. Jacobs and C. A. Fornal. Serotonin and motor activity. *Current Opinion in Neurobiology*, 7:820–825, 1997.

B. L. Jacobs and C. A. Fornal. Activity of serotonergic neurons in behaving animals. *Neuropsychopharmacology*, 21:9S–15S, 1999.

D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15:535–547, 2002.

M. S. Jog, Y. Kubota, C. I. Connolly, V. Hillegaart, and A. M. Graybiel. Building neural representations of habits. *Science*, 286:1745–1749, 1999.

S. Jones and J. A. Kauer. Amphetamine depresses excitatory synaptic transmission via serotonin receptors in the ventral tegmental area. *Journal of Neuroscience*, 19:9780–9787, 1999.

A. Kacelnik. Normative and descriptive models of decision making: Time discounting and risk sensitivity. In G. R. Bock and G. Cardew, editors, *Characterizing Human Psychological Adaptations*, pages 51–70. John Wiley and Sons, Chichester, England, 1997.

A. Kacelnik and M. Bateson. Risky theories — the effects of variance on foraging decisions. *American Zoology*, 36:402–434, 1996.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–275, 1996.

S. Kakade, N. D. Daw, and P. Dayan. Opponent interactions between serotonin and dopamine for classical and operant conditioning. In *Society for Neuroscience Abstracts*, volume 26, page 1763, 2000.

S. Kakade and P. Dayan. Acquisition in autoshaping. In T. K. Leen S. A. Solla and K. R. Muller, editors, *Advances in Neural Information Processing Systems 12*, 2000.

S. Kakade and P. Dayan. Dopamine bonuses. In *Advances in Neural Information Processing Systems 13*, pages 131–137, 2001a.

S. Kakade and P. Dayan. Explaining away in weight space. In *Advances in Neural Information Processing Systems 13*, 2001b.

S. Kakade and P. Dayan. Acquisition and extinction in autoshaping. *Psychological Review*, 109:533–544, 2002a.

S. Kakade and P. Dayan. Dopamine: Generalization and bonuses. *Neural Networks*, 15:549–559, 2002b.

L. J. Kamin. 'attention-like' processes in classical conditioning. In M. R. Jones, editor, *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*, pages 9–31. University of Miami Press, Miami, 1968.

E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors. *Principles of Neural Science, 3rd ed.* Appleton & Lange, Norwalk, Conn., 1991.

S. Kapur and G. Remington. Serotonin-dopamine interaction and its relevance to schizophrenia. *American Journal of Psychiatry*, 153:466–476, 1996.

E. J. Kehoe, P. Graham-Clarke, and B. G. Schreurs. Temporal patterns of the rabbit's nictitating membrane response to compound and component stimuli under mixed cs-us intervals. *Behavioral Neuroscience*, 103: 283–295, 1989.

A. S. Killcross, B. J. Everitt, and T. W. Robbins. Different types of fear-related behaviour mediated by separate nuclei within amygdala. *Nature*, 388:377–380, 1997.

P. R. Killeen and J. G. Fetterman. A behavioral theory of timing. *Psychological Review*, 95:274–295, 1988.

M. R. Kilpatrick, M. B. Rooney, D. J. Michael, and R. M. Wightman. Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. *Neuroscience*, 96: 697–706, 2000.

J. R. Krebs, A. Kacelnik, and P. Taylor. Test of optimal sampling by foraging great tits. *Nature*, 25:27–31, 1978.

Y. Kubota, W. E. DeCoteau, J. Liu, and A. M. Graybiel. Task-related activity in the medial striatum during performance of a conditional t-maze task. In *Society for Neuroscience Abstracts*, volume 28: 765.7, 2002.

S.K. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16:37–68, 1953.

J. Lauwreyns, K. Watanabe, B. Coe, and O. Hikosaka. A neural correlate of response bias in monkey caudate nucleus. *Nature*, 418:413–417, 2002.

D. LeBars. Serotonin and pain. In N. N. Osborne and M. Hamon, editors, *Neuronal Serotonin*, pages 171–226. Wiley, New York, 1988.

S. E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45, 1986.

M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5:356–363, 2002.

M. S. Lewicki and B. A. Olshausen. A probablistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 16:1587–1601, 1999.

T. Ljungberg, P. Apicella, and W. Schultz. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67:145–163, 1992.

P. F. Lovibond. Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 9:225–247, 1983.

R. E. Lubow. Latent inhibition. *Psychological Review*, 79:398–407, 1973.

I. Lucki and J. A. Harvey. Increased sensitivity to d- and l-amphetamine action after midbrain raphe lesions as measured by locomotor activity. *Neuropharmacology*, 18:243–249, 1976.

N. J. Mackintosh. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82:532–552, 1975.

S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms and empirical results. *Machine Learning*, 22:1–38, 1996.

S. Mahadevan, N. Marchalleck, T. Das, and A. Gosavi. Self-improving factory simulation using continuous-time average-reward reinforcement learning. In *Proceedings of the 14th International Conference on Machine Learning*, 1997.

D. Marr. *Vision: A Computational Approach*. Freeman and Co., San Francisco, Ca., 1982.

P. Mason. Physiological identification of pontomedullary serotonergic neurons in the rat. *Journal of Neurophysiology*, 77:1087–1098, 1997.

L. D. Matzel, F. P. Held, and R. R. Miller. Information and the expression of simultaneous and backward associations: Implications for contiguity theory. *Learning and Motivation*, 19:317–344, 1988.

J. E. Mazur. Tests of an equivalence rule for fixed and variable reinforcer delays. *Journal of Comparative and Physiological Psychology*, 10:426–436, 1984.

J. E. Mazur. Probability and delay of reinforcement as factors in discrete-trial choice. *Journal of the Experimental Analysis of Behavior*, 43:341–351, 1985.

J. E. Mazur. Choice between single and multiple delayed reinforcers. *Journal of the Experimental Analysis of Behavior*, 46:67–77, 1986a.

J. E. Mazur. Fixed and variable reinforcer delays: Further tests of an equivalence rule. *Journal of Experimental Psychology: Animal Behavior Processes*, 12:116–124, 1986b.

J. E. Mazur. An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, and H. Rachlin, editors, *Quantitative Analyses of Behavior, Vol. V*, pages 55–73. Lawrence Erlbaum, Hillsdale, N. J., 1987.

J. E. Mazur. Theories of probabilistic reinforcement. *Journal of the Experimental Analysis of Behavior*, 46: 67–77, 1989.

J. E. Mazur. Choice with probabilistic reinforcement: Effects of delay and conditioned reinforcers. *Journal of the Experimental Analysis of Behavior*, 55:63–77, 1991.

J. E. Mazur. Choice, delay, probability, and conditioned reinforcement. *Animal Learning and Behavior*, 25: 131–147, 1997.

A. K. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, 1995.

A.K. McCallum. Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 183–188, Menlo Park, CA, 1993. AAAI/MIT Press.

S. M. McClure, N. D. Daw, and P. R. Montague. A computational substrate for incentive salience. *Trends in Neurosciences*, 2003. (in press).

R. J. McDonald and N. M. White. A triple dissociation of memory systems: Hippocampus, amygdala and dorsal striatum. *Behavioral Neuroscience*, 107:3–22, 1993.

W. H. Meck. Neuropharmacology of timing and time perception. *Cognitive Brain Research*, 3:227–242, 1996.

R. R. Miller and R. C. Barnet. The role of time in elementary associations. *Current Directions in Psychological Science*, 2:106–111, 1993.

R. R. Miller and L. D. Matzel. The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower, editor, *The Psychology of Learning and Memory*, volume 22, pages 51–92. Academic Press, San Diego, CA, 1988.

J. Mirenowicz and W. Schultz. Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, 72:1024–1027, 1994.

J. Mirenowicz and W. Schultz. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379:449–451, 1996.

G.J. Mogenson, D. L. Jones, and C. Y. Yim. From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology*, 14:69–97, 1980.

P. R. Montague and P. R. Baldwin. Neural valuation signals and their connection to option pricing theory. 2003. (submitted).

P. R. Montague and G. S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36: 265–284, 2002.

P. R. Montague, P. Dayan, C. Person, and T. J. Sejnowski. Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, 377:725–728, 1995.

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

P. R. Montague, S. M. McClure, P. R. Baldwin, P. E. M. Philips, E. A. Budygin, , G. D. Stuber M. R. Kilpatrick, and R. M. Wightman. Dynamic gain control of dopamine delivery in freely-moving animals. *Nature Neuroscience*, 2003. (submitted).

A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13:103–130, 1993.

J. W. Moore, J.-S. Choi, and D. H. Brunzell. Predictive timing under temporal uncertainty: The td model of the conditioned response. In D. A. Rosenbaum and C.E. Collyer, editors, *Timing and Behavior: Neural, Computational, and Psychological Perspectives*, pages 3–34. MIT Press, Cambridge, MA, 1998.

B. Murray and P. Shizgal. Behavioral measures of conduction velocity and refractory period for reward-relevant axons in the anterior LH and VTA. *Physiology and Behavior*, 59:643–652, 1996a.

B. Murray and P. Shizgal. Physiological measures of conduction velocity and refractory period for putative reward-relevant MFB axons arising in the rostral MFB. *Physiology and Behavior*, 59:427–437, 1996b.

K. M. Myers, E. H. Vogel, J. Shin, and A. R. Wagner. A comparison of the rescorla-wagner and pearce models in a negative patterning and a summation problem. *Animal Learning and Behavior*, 29:36–45, 2001.

J. Myerson and L. Green. Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, 64:263–276, 1995.

A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and applications to reward shaping. In *Prceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann, San Francisco, CA, 1999.

Y. Niv, D. Joel, I. Meilijson, and E. Ruppin. Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*, 10:5–24, 2002.

R. C. O'Reilly. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. 2003. ICS Technical Report 03-03, University of Colorado.

R. Parr. *Hierarchical Control and Learning for Markov Decision Processes*. PhD thesis, University of California at Berkeley, 1998.

R. Parr and S. Russell. Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems 10*, pages 1043–1049, Cambridge, MA, 1998. MIT Press.

I. P. Pavlov. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, 1927.

G. T. Paxinos and C. Watson. *The Rat Brain in Stereotaxic Coordinates, Compact Third Edition*. Academic Press, New York, 1996.

J. M. Pearce. Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, 101:587–607, 1994.

J. M. Pearce, A. Aydin, and E. S. Redhead. Configural analysis of summation in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, 23:84–94, 1997.

J. M. Pearce and G. Hall. A model for Pavlovian learning: Variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review*, 87:532–552, 1980.

P. E. M. Phillips, G. D. Stuber, M. F. Roitman, P. A. Garris, R. M. Carelli, and R. M. Wightman. Attenuation of dopamine release during intracranial self-stimulation is mediated at the dopaminergic cell body. In *Society for Neuroscience Abstracts*, volume 28: 898.8, 2002.

D. Precup. *Temporal Abstraction in Reinforcement Learning*. PhD thesis, University of Massachusetts, Amherst, MA, 2000.

D. Precup and R. S. Sutton. Multi-time models for reinforcement learning. In *Proc. ICML '97 Workshop on Modeling in Reinforcement Learning*, 1997.

D. Precup and R. S. Sutton. Multi-time models for temporally abstract planning. In *Advances in Neural Information Processing Systems 10*, pages 1050–1056, Cambridge, MA, 1998. MIT Press.

M. L. Puterman. *Markov Decision Processes: Discrete Dynamic Stochastic Programming*. John Wiley and Sons, New York, 1994.

H. Rachlin. *Judgment, Decision and Choice*. W. H. Freeman, New York, 1989.

H. Rachlin, A. W. Logue, J. Gibbon, and M. Frankel. Cognition and behavior in studies of choice. *Psychological Review*, 93:33–45, 1986.

S. Ravel, P. Sardo, E. Legallet, and P. Apicella. Reward unpredictability inside and outside of a task context as a determinant of the responses of tonically active neurons in the monkey striatum. *Journal of Neuroscience*, 21:5730–5739, 2001.

L .A. Real. Animal choice behavior and the evolution of cognitive architecture. *Science*, 253:980–986, 1991.

L. A. Real, J. Ott, and E. Silverfine. On the trade-off between mean and variance in foraging: A experimental analysis with bumblebees. *Behavioral Ecology*, 2:301–308, 1982.

P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023, 1999a.

P. Redgrave, T. J. Prescott, and K. Gurney. Is the short latency dopamine burst too short to signal reinforcement error? *Trends in Neurosciences*, 22:146–151, 1999b.

A. D. Redish, N. C. Schmitzer-Torbert, and J. C. Jackson. Classification of dorsal striatal neurons from extracellular recordings in awake behaving rats. In *Society for Neuroscience Abstracts*, volume 28: 676.3, 2002.

R. A. Rescorla. Pavlovian conditioned inhibition. *Psychological Bulletin*, 72:77–94, 1969.

R. A. Rescorla. Simultaneous second-order conditioning produces s-s learning in conditioned suppression. *Journal of Experimental Psychology: Animal Behavior Processes*, 8:23–32, 1982.

R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical Conditioning, 2: Current Research and Theory*, pages 64–69. Appleton Century-Crofts, New York, 1972.

N. R. Richardson and A. Gratton. Behavior-relevant changes in nucleus accumbens dopamine transmission elicited by food reinforcement: An electrochemical study in rat. *Journal of Neuroscience*, 16:8160–8169, 1996.

F. Rieke, D. Warland, R. de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, Mass., 1997.

R. C. Rizley and R. A. Rescorla. Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, 81:1–11, 1972.

T. W. Robbins and B. J. Everitt. Neurobehavioural mechanisms of reward and motivation. *Current Opinion in Neurobiology*, 6:228– 236, 1996.

T. W. Robbins, B. A. Watson, M. Gaskin, and C. Ennis. Contrasting interactions of pipadrol, d-amphetamine, cocaine, cocaine analogues, apomorphine and other drugs with conditioned reinforcement. *Psychopharmacology*, 80:113–119, 1983.

S. Roberts. Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, 7:242–268, 1981.

A. Rodriguez, R. Parr, and D. Koller. Reinforcement learning using approximate belief states. In *Advances in Neural Information Processing Systems 12*, 1999.

M.L. Rodriguez and A.W. Logue. Adjusting delay to reinforcement: Comparing choice in pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 14(1):105–117, 1988.

R. T. Ross and P. C. Holland. Conditioning of simultaneous and serial feature-positive discriminations. *Journal of Experimental Pscyhology: Animal Behavior Processes*, 9:349–373, 1981.

J.D. Salamone, M.S. Cousins, and B.J. Snyder. Behavioral functions of nucleus accumbens dopamine: Empirical and conceptual problems with the anhedonia hypothesis. *Neuroscience and Biobehavioral Reviews*, 21:341–359, 1997.

J. Sawynok. The role of ascending and descending noradrenergic and serotonergic pathways in opioid and non-opioid antinociception as revealed by lesion studies. *Canadian Journal of Physiology and Pharmacology*, 67:975–988, 1988.

N. A. Schmajuk and J. J. DiCarlo. Stimulus configuration, classical conditioning, and the hippocampus. *Psychological Review*, 99:268–305, 1992.

W. Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27, 1998.

W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13:900–913, 1993.

W. Schultz, P. Apicella, R. Romo, and E. Scarnati. Context-dependent activity in striatum reflecting past and future events. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 11–27. MIT Press, Cambridge, MA, 1995.

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

W. Schultz and R. Romo. Responses of nigrostriatal dopamine neurons to high intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, 57:201–217, 1987.

W. Schultz and R. Romo. Dopamine neurons of the monkey midbrain: Contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, 63:607–624, 1990.

W. Schultz, A. Rufffieux, and P. Aebischer. The activity of pars compacta dopamine neurons of the substantia nigra in relation to motor activiation. *Experimental Brain Research*, 51:377–387, 1983.

A. Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 298–305, San Mateo, Calif., 1993. Morgan Kaufmann.

M. Shidara, T. Aigner, and B. Richmond. Neuronal signals in the monkey ventral striautm related to progress through a predictable series of trials. *Journal of Neuroscience*, 18:2613–2625, 1998.

S. Singh. Reinforcement learning algorithms for average-payoff Markovian decision processes. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 202–207. MIT Press, 1994.

W. E. Skaggs, B. L. McNaughton, M. A. Wilson, and C. A. Barnes. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6:149–172, 1996.

B. F. Skinner. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, New York, 1938.

M. C. Smith. CS-US interval and US intensity in classical conditioning of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 3:679–687, 1968.

R. L. Solomon and J. D. Corbit. An opponent-process theory of motivation: I. temporal dynamics of affect. *Psychological Review*, 81:119–145, 1974.

P. Soubrié. Reconciling the role of central serotonin neurons in human and animal behavior. *Behavioral and Brain Sciences*, 9:319–335, 1986.

J. E. R. Staddon and D. T. Cerutti. Operant conditioning. *Annual Reviews of Psychology*, 54:115–144, 2003.

J. E. R. Staddon and J. J. Higa. Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior*, 71:215–251, 1999.

S. C. Stanford, editor. *Selective Serotonin Reuptake Inhibitors (SSRIs): Past, Present and Future (2nd ed.)*. R. G. Landes, Austin, TX, 1999.

D. W. Stephens. The logic of risk-sensitive foraging preferences. *Animal Behavior*, 29:628–629, 1981.

R. E. Suri. Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Experimental Brain Research*, 140:234–240, 2001.

R. E. Suri, J. Bargas, and M. A. Arbib. Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103:65–85, 2001.

R. E. Suri and W. Schultz. Learning of sequential movements with dopamine-like reinforcement signal in neural network model. *Experimental Brain Research*, 121:350–354, 1998.

R. E. Suri and W. Schultz. A neural network with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91:871–890, 1999.

R. E. Suri and W. Schultz. Temporal difference model reproduces predictive neural activity. *Neural Computation*, 13:841–862, 2001.

R. S. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts, Amherst, MA, 1984.

R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

R. S. Sutton. TD models: Modeling the world at a mixture of time scales. In *ICML 12*, pages 531–539, San Mateo, CA, 1995. Morgan Kaufmann.

R. S. Sutton and B. Pinette. The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 54–64, Irvine, CA, 1985. Larwence Erlbaum.

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

R.S. Sutton and A.G. Barto. Time-derivative models of Pavlovian reinforcement. In M. Gabriel and J. Moore, editors, *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pages 497–537. MIT Press, Cambridge, MA, 1990.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Y. Takikawa, R. Kawagoe, and O. Hikosaka. Reward-dependent spatial selectivity of anticipatory avtivity in monkey caudate neurons. *Journal of Neurophysiology*, 87:508–515, 2002.

S. Thrun. Monte carlo POMDPs. In *Advances in Neural Information Processing Systems 12*, 1999.

P. N. Tobler, A. Dickinson, and W. Schultz. Reward prediction versus attention: The phasic activity of dopamine neurons in a conditioned inhibition task. In *Society for Neuroscience Abstracts*, volume 27: 421.5, 2001.

P. N. Tobler, C. D. Fiorillo, and W. Schultz. Response of dopamine neurons to predicted and unpredicted variations in reward magnitude. In *Society for Neuroscience Abstracts*, volume 28: 876.1, 2002.

E. C. Tolman. *Purposive Behavior in Animals and Men*. Appleton-Century-Crofts, New York, 1932.

D. S. Touretzky, N. D. Daw, and E. J. Tira-Thompson. Combining configural and TD learning on a robot. *Proceedings of the Second International Conference on Development and Learning*, pages 271–276, 2002.

J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35:319–349, 1999.

J. N. Tsitsiklis and B. Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49:179–191, 2002.

K. D. Waddington, T. Allen, and B. Heinrich. Floral preferences of bumblebees (Bombus edwardsii) in relation to intermittent versus continuous rewards. *Animal Behavior*, 29:779–784, 1981.

P. Waelti, A. Dickinson, and W. Schultz. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412:43–48, 2001.

A. R. Wagner. SOP: A model of automatic memory processing in animal behavior. In N. E. Separ and R. R. Miller, editors, *Information Processing in Animals: Memory Mechanisms*, pages 5–47. Erlbaum, Hillsdale, NJ, 1981.

K. Watanabe and M. Kimura. Dopamine receptor-mediated mechanisms involved in the expression of learned activity of primate striatal neurons. *Journal of Neurophysiology*, 79:2568–2580, 1998.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, England, 1989.

C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8:279–292, 1992.

H. G. M. Westenberg, J. A. den Boer, and D. L. Murphy, editors. *Advances in the Neurobiology of Anxiety Disorders*. Wiley, New York, 1996.

N. M. White and R. J. McDonald. Multiple parallel memory systems in the brain of the rat. *Neurobiology of Learning and Memory*, 77:125–184, 2002.

B. Widrow and M. E. Hoff. Adaptive switching circuits. In *1960 Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, pages 96–104. 1960.

B. A. Williams and R. Dunn. Conditioned reinforcement: Neglected or outmoded explanatory construct? *Psychonomic Bulletin and Review*, 1:457–475, 1994.

D. A. Williams and J. B. Overmier. Some types of conditioned inhibitors carry collateral excitatory associations. *Learning and Motivation*, 19:345–368, 1988.

R. A. Wise. Neuroleptics and operant behavior: the anhedonia hypothesis. *Behavioral Brain Science*, 5: 39–87, 1982.

M. A. Wogar, C.M. Bradshaw, and E. Szabadi. Choice between delayed reinforcers in an adjusting-delay schedule: The effects of absolute reinforcer size and deprivation level. *Quarterly Journal of Experimental Psychology*, 45B:1–13, 1992.

M.A. Wogar, C.M. Bradshaw, and E. Szabadi. Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers. *Psychopharmacology*, 111:239–243, 1993.

C. B. Woodbury. The learning of stimulus patterns by dogs. *Journal of Comparative Psychology*, 35:29–40, 1943.

A. J. Yu and P. Dayan. Expected and unexpected uncertainty: ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems 15*, 2003. (in press).

C. L. Zimmer-Hart and R. A. Rescorla. Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, 86:837–845, 1974.

J. Zimmerman, P. V. Hanford, and W. Brown. Effects of conditioned reinforcement frequency in an intermittent free-feeding situation. *Journal of the Experimental Analysis of Behavior*, 10:331–340, 1967.