THE EFFECTS OF MOTIVATION ON HABITUAL INSTRUMENTAL BEHAVIOR

Thesis submitted for the degree of
"Doctor of Philosophy"

by

Yael Niv

Submitted to the Senate of the Hebrew University
July 2007

THE EFFECTS OF MOTIVATION ON HABITUAL INSTRUMENTAL BEHAVIOR

Thesis submitted for the degree of
"Doctor of Philosophy"

by

Yael Niv

Submitted to the Senate of the Hebrew University
July 2007

This work was carried out under the supervision of
Prof. Peter Dayan
Dr. Daphna Joel
Prof. Hanoch Gutfreund

Abstract

This thesis provides a normative computational analysis of how motivation affects decision making. More specifically, we provide a reinforcement learning model of optimal self-paced (free-operant) learning and behavior, and use it to address three broad classes of questions: (1) Why do animals work harder in some instrumental tasks than in others? (2) How do motivational states affect responding in such tasks, particularly in those cases in which behavior is habitual, that is, when responding is *insensitive* to changes in the specific worth of its goals, such as a higher value of food when hungry rather than sated? and (3) Why do dopaminergic manipulations cause global changes in the vigor of responding, and how is this related to prominent accounts of the role of dopamine in providing basal ganglia and frontal cortical areas with a reward prediction error signal that can be used for learning to choose between actions?

A fundamental question in behavioral neuroscience concerns the decision-making processes by which animals and humans select actions in the face of reward and punishment. In Chapter 1 we provide a brief overview of the current status of this research, focused on three themes: behavior, computation and neural substrates. In behavioral psychology, this question has been investigated through the paradigms of Pavlovian (classical) and instrumental (operant) conditioning, and much evidence has accumulated regarding the associations that control different aspects of learned behavior. The computational field of reinforcement learning has provided a normative framework within which conditioned behavior can be understood. In this, optimal action selection is based on predictions of long-run future consequences, such that decision making is aimed at maximizing rewards and minimizing punishment. Neuroscientific evidence from lesion studies, pharmacological manipulations and electrophysiological recordings in behaving animals have further provided tentative links to neural structures underlying key computational constructs in these models. Most notably, much evidence suggests that the neuromodulator dopamine provides basal ganglia target structures with a reward prediction error that can influence learning and action selection, particularly in stimulus-driven habitual instrumental behavior.

However, although reinforcement learning models have long promised to unify computational, psychological and neural accounts of appetitively conditioned behavior, we claim here that they suffer from a large theoretical oversight. While a bulk of data on animal conditioning comes from free-operant experiments measuring how fast animals will work for reinforcement, existing reinforcement learning models lack any notion of *vigor* or *response rate*, focusing instead only on competition between different responses, and so they are silent about these tasks. In Chapter 2 we first review the basic characteristics of free-operant behavior, illustrating the effects of reinforcement schedules on rates of responding. We then develop a reinforcement learning model in which vigor selection is optimized together with response selection. The model suggests that subjects choose how vigorously to perform selected actions by optimally balancing the costs and benefits of different speeds of responding. Importantly, we show that this model accounts normatively for effects of reinforcement schedules on response rates, such as the fact that responding on ratio schedules is faster than responding on interval schedules that yield the same rate of reinforcement. Finally, the model

highlights the importance of the *net rate of rewards* in quantifying the *opportunity cost* of time, and thus in determining response vigor.

In Chapter 3 we flesh out the implications of this model for the motivational control of habitual behavior. In general, understanding the effects of motivation on instrumental action selection is fundamental to the study of decision making. Recent work has shown that motivational control can be used to divide instrumental behavior into two classes: 'goal-directed' behavior is immediately sensitive to motivation-induced changes in the values of its specific consequences, while 'habitual' behavior is not. Because habitual behavior constitutes a large proportion of our daily activities, it is thus important to ask how does motivation affect habitual behavior? That is, how can habitual behavior be performed such as to achieve *motivationally relevant* goals?

We start by defining motivation as a *mapping* from outcomes to utilities. Incorporating this into the computational framework of optimal response rates, we show that in general, the *optimal* effects of motivation on behavior should be two-fold: On the one hand, motivation should affect the choice between actions such that actions leading to those outcomes that are more highly valued are more probable. This corresponds to the traditional *directing* effect of motivation. On the other hand, by influencing the opportunity cost of time, motivation should affect the response rates of *all* chosen actions, irrespective of their specific outcomes, as in the decades-old (but controversial) idea that motivation *energizes* behavior. This global effect of motivation explains not only why hungry rats work harder for food, but also sheds light on the counterintuitive observation that they will sometimes work harder for water. Based on the computational view of habitual behavior as arising from cached values summarizing long-run reward predictions, we suggest that habitual action selection can direct responding properly only in those motivational states which pertained during behavioral training. However, this does not imply insensitivity to novel motivational states. In these, we propose that the outcome-independent, global effects of motivational can 'energize' habitual actions, as a well-founded approximation to the optimal solution in a trained situation. That is, habitual response rates can be adjusted to the current motivational state, in a way that is optimal given the computational limitations of the habitual system.

Our computational framework suggests that the energizing function of motivation is mediated by the *expected net rate of rewards*. In Chapter 4, we put forth the hypothesis that this important quantity is reported by *tonic levels of dopamine*. Dopamine neurotransmission has long been known to exert a powerful influence over the vigor, strength or rate of responding. However, there exists no clear understanding of the computational foundation for this effect. Previous reinforcement learning models of habitual behavior have indeed suggested an interpretation of the function of dopaminergic signals in the brain. However, these have concentrated only on the role of precisely timed phasic dopaminergic signals in learning the predictive value of different actions, and have ignored both tonic dopamine transmission and response vigor. Our tonic dopamine hypothesis focuses on the involvement of dopamine in the control of vigor, explaining why higher levels of dopamine are associated with globally higher response rates, ie, why, like motivation, dopamine

'energizes' behavior. In this way, through the computational framework of optimal choice of response rates, we suggest an explanation of the motivational control of habitual behavior, on both the behavioral and the neural levels.

Reinforcement learning models of animal learning are appealing not only because they provide a normative basis for decision-making, but also because they show that optimal action selection can be learned through online incremental experience with the environment, using only locally available information. To complete the picture of how dopamine influences free-operant learning and behavior, in Chapter 5 we describe an online algorithm of the type usually associated with instrumental learning and decision-making, which is suitable for learning to select actions and latencies according to our new framework. There are two major differences between learning in our model and previous online reinforcement learning algorithms: First, most prior applications have dealt with discounted reinforcement learning while we use average reward reinforcement learning. Second, unlike previous models that have focused on discrete action selection, the action space in our model is inherently continuous, as it includes a choice of response latency. We thus propose a new online learning algorithm that is specifically suitable for our needs. In this, building on the experimental characteristics of response latencies, we suggest a functional parameterization of the action space that drastically reduces the complexity of learning. Moreover, we suggest a formulation of online action selection in which response rates are directly affected by the net reward rate. We show that our algorithm learns to respond appropriately, and with nearly optimal latencies, and discuss its implications for the differences between the learning of interval and ratio schedules.

In Chapter 6, the last of the main chapters, we deviate from the theoretical analysis of behavior, to describe two instrumental conditioning experiments in rats, whose goal was to clarify empirically the effects of motivation on habitual behavior. Because previous studies have only considered motivational downshifts, they cannot distinguish between 'energizing' effects of motivation and effects of generalization decrement as a result of a motivational shift. To remedy this, we investigated the effects of motivational up-shifts and side-shifts on habitual behavior. In Experiment 1, hungry rats were trained to leverpress for sucrose solution. Following a side-shift from food- to water-deprivation, rats showed less leverpressing in extinction compared to non-shifted controls, although a subsequent consumption test found no differences in sucrose consumption between thirsty and hungry groups. In Experiment 2, undeprived rats were trained to leverpress for either sucrose solution or sucrose pellets. A post-training up-shift from satiety to water-deprivation did not affect leverpressing in extinction, regardless of outcome identity, although free consumption of sucrose solution, but not of pellets, was enhanced. Together, these results provide preliminary support for the predictions of our model.

Finally, in Chapter 7 we detail the contributions of the thesis as a whole, as well as some general directions for future research.

Contents

Introduction **xix**

0.1 Themes xx

0.2 The structure of this thesis xxiii

1 Background **1**

1.1 Behavior: Conditioning, prediction and action selection 2

1.1.1 Pavlovian conditioning 2

1.1.2 Instrumental conditioning 5

1.1.3 Goal directed versus habitual instrumental behavior 6

1.1.4 Interactions between Pavlovian and instrumental conditioning 9

1.2 Computational models of animal conditioning 10

1.2.1 Optimal predictions 12

1.2.2 Optimal action selection 16

1.2.3 Computational processes underlying goal directed versus habitual control 20

1.3 Neural substrates underlying conditioned behavior 22

1.3.1 The neuromodulator dopamine and temporal difference learning 24

1.3.2 Pavlovian behavior: The 'limbic loop' 26

1.3.3 Goal directed behavior: The 'associative loop' 28

1.3.4 Habitual behavior: The 'sensorimotor loop' 30

1.4 An integrative view of conditioned behavior 31

1.5 Motivation and the organization of this thesis 33

2 A computational model of free operant behavior **35**

2.1 Introduction . . . 35

2.1.1 Free-Operant schedules . . . 36

2.1.2 Characteristics of free-operant behavior . . . 38

2.2 Methods: Modeling free-operant behavior . . . 43

2.2.1 Building blocks . . . 43

2.2.2 The model . . . 45

2.2.3 Average reward optimization . . . 47

2.2.4 Discounted reward optimization . . . 50

2.2.5 Action selection . . . 51

2.3 Results . . . 52

2.3.1 Cost/benefit tradeoffs and the opportunity cost of time . . . 52

2.3.2 Fine-scale temporal behavior . . . 53

2.3.3 Ratio versus interval schedules . . . 55

2.3.4 Effects of reward magnitude and schedule parameter on vigor . . . 58

2.3.5 Breaking-point analysis in ratio schedules . . . 60

2.3.6 Concurrent interval schedules . . . 61

2.3.7 The effect of free rewards . . . 62

2.4 Discussion: A new theory of free-operant responding . . . 63

2.4.1 Molar vs molecular explanations . . . 65

2.4.2 Relationship to previous models . . . 66

2.4.3 Limitations of the model and future directions . . . 71

2.4.4 Conclusions . . . 74

3 Motivation: The directing and the energizing **75**

3.1 Introduction . . . 75

3.1.1 Motivation: A mapping from outcomes to utilities . . . 76

3.2 Results: The effects of motivation on behavior . . . 79

3.2.1 Goal-directed behavior: A 'brute force' solution . . . 81
3.2.2 Habitual behavior: A motivation-insensitive shortcut? . . . 82
3.3 Discussion: Two sides of motivational influence . . . 85
3.3.1 Relationship to previous models . . . 87
3.3.2 Neural underpinnings of motivational control: clues from PIT . . . 87
3.3.3 Conclusions . . . 88

4 Neural substrates: Dopamine and response vigor **89**
4.1 Introduction . . . 89
4.2 Methods: Modeling dopamine in free operant response choice . . . 92
4.3 Results: The net rate of reward and tonic dopamine . . . 94
4.3.1 Tonic and phasic dopamine . . . 95
4.4 Discussion . . . 98
4.4.1 Predictions . . . 98
4.4.2 Immediate *vs.* learned effects . . . 99
4.4.3 Future directions: Reward *vs.* punishment; instrumental *vs.* Pavlovian . . . 100
4.4.4 Conclusions . . . 102

5 Online learning of optimal response rates **103**
5.1 Introduction . . . 104
5.1.1 The problems . . . 105
5.1.2 Constraints/Desiderata . . . 109
5.2 Methods . . . 111
5.2.1 An Actor/Critic policy-iteration algorithm for free operant behavior . . . 111
5.2.2 Correspondence to 'Indirect Actor' methods . . . 113
5.2.3 Incorporating the net rate of reward into the algorithm . . . 115
5.2.4 Constraining the policy parameters . . . 116
5.3 Results . . . 117

5.3.1 Learning dynamics . . . 117
5.3.2 Learned policies vs. optimal policies . . . 120
5.3.3 Action selection and changes in the expected rate of rewards . . . 120
5.4 Discussion . . . 122
5.4.1 Relationship to previous work . . . 123
5.4.2 Interval vs. ratio schedules . . . 125
5.4.3 Limitations of the learning algorithm . . . 128
5.4.4 Conclusions and future directions . . . 130

6 Motivation and habitual behavior: An empirical investigation 131
6.1 Introduction . . . 132
6.1.1 The possible effects of motivation on behavior . . . 133
6.2 Experiment 1: Motivational side-shift . . . 135
6.2.1 Materials and Methods . . . 136
6.2.2 Results . . . 138
6.3 Experiment 2: Motivational up-shift . . . 140
6.3.1 Materials and Methods . . . 140
6.3.2 Results . . . 142
6.4 Discussion . . . 143

7 Contributions and future directions 147

References 151

List of Figures

1.1 Illustrations of animal conditioning . . . 3

1.2 A scheme of the Konorskian model of Pavlovian conditioning . . . 4

1.3 Outcome revaluation teases apart habitual and goal-directed instrumental behavior . . . 6

1.4 Temporal difference prediction errors and phasic dopaminergic firing . . . 15

1.5 Actor/Critic architecture . . . 19

1.6 A cartoon of brain areas implicated in conditioned behavior . . . 23

1.7 Summary diagram of the neural control of conditioned behavior . . . 27

2.1 An illustration of the response patterns generated by different reinforcement schedules . . . 37

2.2 Responding on ratio and interval schedules . . . 39

2.3 "Matching Law" behavior for one or two actions . . . 40

2.4 Responding of well-trained, food deprived rats on a VI30 schedule of reinforcement . . . 42

2.5 Inter response distributions for consecutive actions . . . 42

2.6 Model dynamics . . . 45

2.7 Data generated by the model captures the essence of the behavioral data . . . 55

2.8 Q-values and soft-maxed Q-values . . . 56

2.9 Comparison of performance on random ratio and random interval schedules of reinforcement 57

2.10 Ratio and interval leverpress rates as a function of schedule parameter . . . 58

2.11 Probability of choosing the different available actions in an extinction test, for different ratio requirements . . . 60

2.12 Matching behavior in the model . . . 62

3.1 The effects of motivation on behavior: an illustrative problem 78

3.2 The effect of $\bar{R}$ on response rates 80

3.3 Two ways to estimate action values for action control 83

4.1 Effects of dopamine depletion on fixed ratio responding 91

4.2 The T-maze cost/benefit task 97

5.1 Estimating the net reward rate from optimal actions alone can result in biased estimates . . . 108

5.2 Learning curves for random ratio and random interval schedules 118

5.3 Policies learned in each of 20 runs of our online Actor/Critic learning algorithm on an RR10 schedule of reinforcement 119

5.4 Policies learned in each of 20 runs of our online learning Actor/Critic algorithm on an RI30 schedule of reinforcement 119

5.5 Optimal Q-values for random ratio and random interval schedules, as derived by relative value iteration 121

5.6 Steady state behavior of the Actor/Critic model trained on various fixed ratio tasks, before and after dopamine depletion 121

5.7 Between subject and within subject variability in interval schedule behavior: model and experimental results 126

6.1 Results of Experiment 1: Motivational side-shift 139

6.2 Results of Experiment 2: Motivational up-shift 142

List of Tables

6.1 Possible effects of motivation on behavior 134

6.2 Design of Experiments 1 and 2 135

Its not too far, it just seems like it is – Yogi Berra

Acknowledgments

This thesis is the direct product of four years of training and research conducted at three universities: at the Gatsby Computational Neuroscience Unit, UCL, in collaboration with and under the exemplary guidance of Peter Dayan; at Tel Aviv University with the immense help and guidance of Daphna Joel; and under the resourceful and accommodating auspices of the Interdisciplinary Center for Neural Computation (ICNC), The Hebrew University of Jerusalem, with the kind guidance of Hanoch Gutfreund. The roots of this work can be further traced to my earlier training at the ICNC and at Tel Aviv University. I am greatly indebted to many for the final result – the work contained within this thesis, as well as the immense body of knowledge and the numerous insights that I take with me to my future research.

First and foremost, I wish to thank Peter Dayan. Peter has contributed in so many ways to this work specifically and to my development as a researcher in general, so as to make any brief acknowledgment necessarily incomplete. Peter's influence on every part of this thesis can not be overestimated. Through his careful attention to detail and his unfalteringly high scientific standards he has had a major influence on the focus, substance and style of this manuscript. As a mentor, the personal example that Peter set is unparalleled: from his sesquipedalian vocabulary, through inspirational time management, and culminating in his extremely wide and diverse knowledge, which he is never unwilling to impart. I am constantly astonished at the breadth and depth of his knowledge, and at the same time, his generous attribution of credit to others for reinventing or elaborating on his ideas. I am deeply grateful to Peter for his immense patience with my slow and repetitious mode of learning, for the genuine respect and seriousness with which he has treated my meager knowledge and frequently erroneous ideas, for constantly promoting my scientific growth, and for providing me with a sound foundation on which to base my future scientific quests. It is these that make Peter a true mentor.

Daphna Joel, who has followed my scientific development from my first day as an undergraduate, has been extremely instrumental in my graduate training. In addition to friendship and personal advice, Daphy provided me with a true home in the Psychology Department – physically (whenever I was in Israel) and mentally (by teaching me to speak the psychological language correctly, and by always representing the voice of data). Despite never being published, the experiments described in Chapter 6, which were conducted in Daphy's lab and under her direct supervision, provided the scaffolding for this whole thesis. Although they

are described last, these experiments were carried out prior to the development of the theory that consists of the bulk of this thesis. Their results not only suggested a new understanding of the effects of motivation on habitual behavior, but it is the detailed analysis of the rats' behavior that gave me the intuition and insights that led to the formulation of the reinforcement learning model in terms of a choice of action and latency.

Hanoch Gutfreund, my supervisor at The Hebrew University and the one officially burdened with the responsibility of bringing me to the point of graduation, was a great facilitator of my strange trajectory from ICNC through Tel Aviv and to London. It is due to Hanoch that funding was secured for equipment, rats and my different fellowships, that all barriers in the way of research abroad with a cross-continental group of supervisors were removed, and that all this could be done under the formal auspices of ICNC. I am grateful for the attention and care that Hanoch bestowed me, for the timely meetings in London and in Israel, and for always looking out for my best interests. Despite my research being rather unrelated to his work at The Hebrew University, his genuine interest in my findings and his unfaltering commitment as my formal advisor were remarkable.

In addition to my three advisors, I've had the incredible fortune of having Nathaniel Daw as a collaborator and an informal advisor (and, more recently, my husband). Nathaniel was an integral part of this work from its inception, and his ideas permeate every chapter of this thesis. His extremely useful suggestions at every step of analysis or simulation, his insightful explanations of talks, papers and concepts, and his unwavering support (not to mention great cooking) throughout the long process of writing this thesis, were invaluable. Moreover, through many a scientific conversation (or argument), Nathaniel's views and beliefs have had a major part in shaping my personal scientific agenda – it is Nathaniel who taught me the word "normative". With great love I dedicate this thesis to him.

Many others have had an essential role for which I am deeply grateful: the various people at Gatsby, in particular Arik Azran, Michael Duff, Quentin Huys and Peter Latham; Daphy's lab, especially Tom Schoenberg; the people at ICNC, especially Michal Rivlin-Etzion; and my flatmates in London, in particular Jess Gough. All of these have given me a home wherever I happened to be in the past four years.

Numerous people have contributed to my research through invaluable discussions and comments on drafts and ideas, among them Bernard Balleine, Hanna Bayer, Hagai Bergman, Peter Bossaerts, Sanne De Wit, Anthony Dickinson, Michael Duff, Zoubin Ghahramani, Aaron Gruber, Peter Holland, Quentin Huys, Simon Killcross, Peter Latham, Brian Lau, Genela Morris, Saleem Nicola, John O'Doherty, Antonio Rangel, Matthew Rushworth, Maneesh Sahani, Tom Schoenberg, Geoffrey Schoenbaum, Peter Shizgal and Mark Walton. A number of people have generously shared their experimental data with me, including Bernard Balleine, Hannah Bayer, Rudolf Cardinal, Genela Morris and Saleem Nicola. For a computational modeler such as myself, raw data is priceless.

My own experimental work was greatly facilitated by the helpful advice of Bernard Balleine and Sanne De Wit regarding experimental design and setup. I am grateful to Sharon Riwkes and Eran Katz for diligently carrying out some of the experiments, and to Shelly Inbar and Amaya De Levie for very patiently teaching

me the practicalities of working with (and looking after) rats and for supporting me through the crises of the first experiments.

I am grateful to Idan Segev and Eilon Vaadia at the ICNC for their ongoing support and for creating the opportunity for me to teach "Learning and Behavior" at the ICNC, a course from which I learned as much as I taught. I thank my thesis committee, Idan Segev, Hagai Bergman and Yoram Singer for their advice at early stages of this work, and Ruthi Suchi and Sigal Cohen at ICNC for their continuous help with the trans-continental aspects of my studies. The flexibility and openness of the ICNC are unique.

Finally, I thank my parents who made sure I did not feel alone in a foreign country, who took an active interest in all aspects of my work, and who took care of all the practical issues involved with my bi-national existence. Without them all of this would have not been possible.

Throughout my studies I was funded by the ICNC, the EC Thematic Network, a Dan David fellowship, a Hebrew University Rector Fellowship, and the Gatsby Charitable Foundation.

Last, my doctoral research began at a time of personal unbearable crisis – the inconceivable and traumatic death of my beloved Jörg Kramer. I thank from the bottom of my heart all of you above who (knowingly or unknowingly) supported me in what was an exceptionally tough time. It is due to Jörg and his love for London that I first contemplated going there. It is because of London that I managed to recover from my loss. Like with everything else I do, Jörg's influence on this thesis is ever-present.

The hallmark of the presence of and the need for motivational concepts in behavior is the energization of responses and the control of their vigor and efficiency

– Charles N. Cofer, "Motivation and Emotion" (1972), pp.34

Introduction

A fundamental and theoretically nontrivial observation is that hungry animals will work harder to get food. The challenge that motivational control poses to our current understanding of decision-making is exemplified by three related experimental observations: First, recent work in behavioral psychology has shown that a large class of behaviors (called 'habitual'), is actually *insensitive* to changes in the worth of its goals, such as the higher value of food when hungry rather than sated (Dickinson & Balleine, 2002). Second, counter-intuitively, hungry animals will not only work harder to obtain food, but in some circumstances, they will work more vigorously even for *motivationally irrelevant outcomes* such as water (Hull, 1943). Last, current accounts of the neural basis of decision-making have emphasized the role of the neuromodulator dopamine in providing a prediction error signal that can be used for learning to choose between actions. However, psychopharamacological studies show that the most prominent effects of dopamine interventions are manifest not in the choice of actions, but rather in *global changes in response vigor* (Salamone & Correa, 2002), reminiscent of the effects of motivation.

That the control of response vigor is such a puzzle is perhaps partly due to a large theoretical oversight on the part of current computational models of decision making. Although response vigor, or response rates, are by far the most widely used dependent variables in experimental investigations of behavior (Williams, 1994), the notion of vigor is wholly absent from *reinforcement learning* models, which are, to date, the most prominent normative computational models of learning and behavioral control (Sutton & Barto, 1998). In terms of decision making and action selection, response rates are in many ways inseparable from response choice: accompanying any decision of what to do is a choice of how fast (or at what instantaneous rate) to do it. It is thus surprising that normative models of responding, which have done much to explain *why* it is appropriate for animals to choose actions the way they do, have completely ignored the choice of response rates. Response rates have played a more prominent role in descriptive models of animal behavior. These models aim to quantify the relationships between different experimental variables and the rate of responding (eg. the "Matching Law", Herrnstein, 1997), but do not explain why, or in what sense these relationships are appropriate in different scenarios. In the absence of normative models, the question of *why* does motivation influence response rates, is left unanswered.

In this thesis, I aim to remedy this by extending the framework of reinforcement learning to the optimal selection of response rates, in order to investigate the central question: *how does motivation affect habitual behavior*? Habitual behavior constitutes a large proportion of our daily activities. Because our motivational state defines the current goals we should be seeking, asking how motivation controls habitual action selection is to ask whether and how can habitual behavior be performed such as to achieve *relevant* goals?

To answer this, I will show that the optimal effects of motivation on behavior should be two-fold: On the one hand, motivation should affect the choice between actions, which corresponds to the role traditionally ascribed to motivation in *directing* action selection to those outcomes that are highly valued. On the other hand, motivation should affect the response rates of *all* chosen actions, irrespective of their specific outcomes. This can be associated with the decades-old (but controversial) idea that motivation *energizes* behavior, such that responding is faster in higher deprivational states. Recent work has shown that instrumental behavior can be divided into two classes, based on its sensitivity to the directing influence of motivation: 'goal-directed' behavior is immediately sensitive to motivation-induced changes in the values of its specific consequences, while 'habitual' behavior is not. However, I will claim that habitual behavior is not wholly motivation-insensitive, but rather, it can still be sensitive to the general energizing aspects of motivation. That is, habitual response rates can be adjusted to the current motivational state, in a way that is optimal given the computational limitations of the habitual system. This global effect of motivation on habitual responding will explain not only why hungry rats work harder for food, but also why they will sometimes work harder for water. Note that while this thesis focuses on response rates in habitual behavior, the control of vigor in the goal-directed system has also been ignored. Due to its normative nature, the analysis of vigor that I will present here is, in fact, relevant to both classes of instrumental responding. However, for the sake of brevity, its implications for goal-directed responding will be only lightly touched upon.

Finally, I will suggest that the energizing function of motivation is mediated by the *expected net rate of rewards*, and I will put forth the hypothesis that this quantity is reported by *tonic levels of dopamine*. Previous reinforcement learning models of habitual behavior have indeed been strongly linked to dopaminergic signals in the brain (eg Schultz, 1998). However, these have concentrated only on the role of precisely timed phasic dopaminergic signals in learning the predictive value of different actions. In contrast, the tonic dopamine hypothesis posed here focuses on the involvement of dopamine in the control of vigor, to explain why higher levels of dopamine are associated with globally higher response rates, ie, why, like motivation, dopamine 'energizes' behavior (Salamone & Correa, 2002; Weiner & Joel, 2002). In this way, through the computational framework of optimal choice of response rates, this thesis will provide a solution to the puzzle of motivational control of habitual behavior, on both the behavioral and the neural levels.

0.1 Themes

Woven throughout this thesis are three intertwining and synergistic themes: *animal conditioned behavior* as a well-defined instantiation of action control, *computational models* of decision-making, and *neural mech-*

anisms underlying response selection in the brain. I wish here to briefly introduce each theme.

Animal behavior: instrumental conditioning

The main goal of this research is to understand the decision-making that takes place as part of day-to-day trial-and-error learning and action selection. The results of this process are most evident experimentally in instrumental conditioning paradigms, which investigate animal or human decision-making under carefully controlled conditions in which certain actions are required in order to obtain specific goal outcomes. Decades of research in instrumental responding have demonstrated that this behavior is rich and diverse, yet it obeys general characteristics that can be readily identified. I will treat these characteristics as fundamentals of instrumental conditioning, and use them to shed light on the computational and neural processes underlying instrumental control. It is these that the computational model will be expected to replicate, and therefore explain.

Computational models: on the use of normative models

The computational framework I choose to work with is that of *normative models*. In contrast to descriptive models that describe behavior as it is, normative models study behavior from the point of view of its *function*, that is, they study behavior *as it should be* if it were to accomplish specific goals in an optimal way. The appeal of normative models stems from several reasons: First, because animal behavior has been shaped and constrained by its functionality throughout evolution, the behavior we observe is, in many cases, indeed optimal in terms of solving the problems that animals face (Kacelnik, 1997). Moreover, discrepancies between observed behavior and the predictions of normative models are also illuminating: these can shed light on the neural and/or informational constraints under which animals make decisions, or suggest that animals are, in fact, optimizing something other than what the model has assumed. Normative models thus aim to derive from first principles what behavior should look like, assuming some function or goal, and by comparing the predictions of the model to the actual behavior of animals, to test what are animals optimizing, and what constraints are they subject to. The advantages of such an approach is that it not only explains the structure of behavior, but also *why* it is this way.

The normative framework best suited to model instrumental behavior, and the one which will be used in this thesis, is reinforcement learning (RL). The computational field of RL provides a host of tools with which to optimally solve a broad class of (simplified) decision making problems, namely, Markov Decision Processes. Not only does RL have a sound mathematical basis in the engineering theory of dynamic programming (Bertsekas & Tsitsiklis, 1996), but it also has long had a very close relationship with psychological accounts of decision making in animals and humans (Sutton & Barto, 1981). In the past two decades, much research has used RL as a framework for understanding how animals can predict future rewards and punishments, and choose actions that optimize those affective consequences. However, almost all existing applications

of RL have been to discrete-trial tasks, in which the only choices that subjects make are between different punctate actions (such as pressing either the left or the right lever in an operant chamber, or turning either left or right in a maze). This is clearly inadequate as a model of the effects of motivation on responding. Moreover, it does nothing to explain a large class of *free operant* tasks, in which the key dependent variable has to do with at what *rate* an animal performs an action, in the light of different schedules of reinforcement (Domjan, 2003). Here, I will suggest an extension of the standard RL model, such that along with making a choice between different possible actions, subjects also choose the latency (interpreted as response strength or vigor) with which they perform it.

The assumptions that the new model makes are the following: First, we assume a continuous-time Markov decision process (ie, that the environment comprises of a set of states (or situations) each of which is associated with a certain stationary probability of reward, and that the animal's actions can influence the transitions between these states). In this, we assume that at each state the animal can choose to perform one of a discrete set of actions. So far these assumptions are shared with previous RL models. Importantly, the two additional assumptions that we make are that the animal also chooses with what latency (or instantaneous rate) to perform its chosen action, and that behavior incurs a cost that can depend on the chosen response and its latency, as well as on the state of the environment. Finally, as befits a normative model, we assume that the goal of the animal is to *maximize the net rate of benefits* it obtains, that is, the rate of all rewards minus all costs. These assumptions will be formalized more precisely in Chapter 2, but I have explicitly stated them here because the results of this thesis all rely on, and pertain to, the validity of these assumptions.

Neural mechanisms: dopamine, learning and choice

Finally, I am interested in understanding decision making and action selection as it actually takes place in humans and animals, implemented with the neural machinery these have at their disposal. Here, as well, I am interested primarily in a systems level description of the computational roles of different neural components, rather than a detailed mechanistic or biophysical implementation. The questions I will ask are ones regarding function: *what* does an area of the brain do, rather than *how* does it do it.

As mentioned above, the main neural components that have been linked to reinforcement learning in the mammalian brain are, first and foremost, the neuromodulator dopamine, and its target areas in the basal ganglia and frontal cortex. The reward prediction error theory of phasic dopaminergic signaling (Montague et al., 1996; Schultz et al., 1997) is perhaps the best example to date of a computational theory bridging physiological and behavioral levels, and providing a unified, normative account of both the neuronal responses in dopaminergic and related brain systems and of how the computations they carry out influence learning and behavior. Here, I will extend this line of research by suggesting a normative account for the role of *tonic levels of dopamine* in the control of response vigor. In fact, I will argue that the effects of motivation on habitual behavior are mediated by tonic levels of dopamine.

0.2 The structure of this thesis

This thesis comprises of a background chapter, five main chapters that together portray a picture of the motivational control of habitual behavior (although each can be read on its own), and a conclusions chapter. The first chapter lays out the general background for the subsequent work, in terms of the three themes of behavior, computation and neural substrates. This background is meant to provide the relevant context in which this thesis should be understood, as well as to precisely define its focus: what is habitual instrumental behavior, what is already known about its computational and neural implementation, and why it is interesting and important to ask how it is affected by motivation. In this chapter I also discuss the broad view of how the basal ganglia and frontal cortex are believed to bring about conditioned behavior, within which the neural implications of this thesis should be considered.

Chapter 2 presents the new RL model of response rates. Specifically, in this chapter I describe the characteristics of free operant behavior, and use the model to replicate and explain them. The implications of the model for motivational control of responding are discussed in Chapter 3. First, I begin by defining motivation in a computationally precise way, as a mapping from outcomes to their subjective utility. I then use the model to derive the optimal effects that a change in motivational state should have on behavior, and discuss, in light of the computational constraints on habitual responding, how habitual behavior can approximate optimal sensitivity to motivation. Chapter 4 continues to flesh out the implications of the model presented in Chapter 2, but this time for the neural substrates underlying habitual control. Here, I marshal physiological, psychological and computational evidence supporting the hypothesis that the tonic level of dopamine plays a major role in determining the optimal rates of behavior, by representing the expected net rate of rewards. According to this hypothesis, the effects of motivation on habitual response rates are mediated by tonic levels of dopamine.

To complete the picture of how dopamine influences learning and behavior in a neural reinforcement learning setting, Chapter 5 describes an online RL algorithm of the type usually associated with decision-making in the basal ganglia, which is suitable for learning and selecting actions according to the new model proposed in Chapter 2. The last of the main chapters, Chapter 6 deviates from the theoretical treatment and analysis of behavior, to describe two instrumental conditioning experiments in rats, whose goal was to clarify empirically the effects of motivation on habitual behavior. In these, I trained rats to press a lever habitually in order to obtain sucrose, and examined the effects of shifts in motivational states on their behavior. The results provide preliminary support for the predictions of the model. Finally, Chapter 7 details the conclusions and contributions of the thesis as a whole, as well as directions for future research.

Theme I: Behavior
Theme II: Computation
Theme III: Neural substrates

Chapter 1

Background

Abstract: A fundamental question in behavioral neuroscience concerns the decision-making processes by which animals and humans select actions in face of reward and punishment, and their neural realization. In behavioral psychology, this question has been investigated in detail through the paradigms of Pavlovian (classical) and instrumental (operant) conditioning, and much evidence has accumulated regarding the associations that control different aspects of learned behavior. These experimental findings have been interpreted within a normative framework, in terms of the computations that must be realized in the service of appropriate action selection. Lesion studies, pharmacological manipulations and extracellular recordings in behaving animals have further provided tentative links to neural structures underlying these functions. This thesis deals with a specific aspect of such decision-making: what are the effects of motivation on habitual action selection? To provide the broader context within which this question lies, this chapter provides a brief background to conditioned behavior centered on three themes: behavior, computation, and neural substrates.

The study of instrumental conditioning is an inquiry into perhaps the most fundamental form of decision-making. It is this capacity to select actions that influence the environment to one's subjective benefit, that marks intelligent organisms. Although animals such as pigeons and rats are capable of modifying their behaviors in response to the contingencies provided by the environment, choosing those behaviors that will maximize rewards and minimize punishments in an uncertain, often changing, and computationally complex world is by no means an easy task.

In this chapter, I will first describe a long and sophisticated line of experimental psychological studies which have probed the nature of this type of decision making, and carved out its different subtypes. The second part of the chapter will then review a normative computational framework within which these experimental findings have been interpreted. According to this, optimal decision making relies on two aspects: predicting the consequences of ones actions and choosing those actions that will lead to the most favorable consequences. I will discuss these two aspects of prediction and control, providing an overview of mainstay computational

models of instrumental conditioning. Recent neuroscientific investigations have made considerable progress in elucidating the mechanisms underlying decision making in the brain: specific neural structures have been implicated in representing and learning the different components of a conditioning task, and neural signals corresponding to critical components of the computational models have been identified. In the third part of this chapter, I will very briefly discuss the current view of how the basal ganglia, the limbic cortex and the frontal cortex bring about conditioned behavior, focusing on the distinct roles of different substructures of the basal ganglia and the amygdala. The three themes of behavior, computation and neural realization will then come together in an integrative view of conditioned behavior in the fourth section. Finally, the chapter will end with a discussion of what is known about how motivation influences action selection in the service of the internal needs and goals of the decision-maker, and, importantly, the main puzzles regarding motivational control, which this thesis aims to answer.

The goals of this chapter are twofold. The first is to unfold the title of this thesis, that is, to define *instrumental habitual* behavior, and to contrast it with other types of behavior, and, by way of background, to discuss what is already known about the effects of *motivation* on these different types of behaviors. A second aim is to provide a brief tutorial of the high-level background necessary for understanding this thesis and placing it within appropriate context. Due to the multi-themed nature of this thesis, this chapter does not aim to provide a comprehensive review of all the relevant literature. Rather, each of the subsequent chapters will begin with a short review of its specific background literature.

1.1 Behavior: Conditioning, prediction and action selection

Behavioral psychologists have long made a distinction between two forms of learning (or conditioning), namely, Pavlovian (or classical) conditioning, and instrumental (or operant) conditioning. This distinction rests on whether the animal only observes the relationships between events in the world (in Pavlovian conditioning), or whether it also has some control over their occurrence (in instrumental conditioning). Operationally, in the latter motivationally significant outcomes (such as food or shocks) are contingent on the animal's behavior, whereas in the former these occur regardless of the animal's actions. However, the distinction between these two paradigms is more than technical – in Pavlovian conditioning, changes in behavior presumably reflect innately specified reactions to the prediction of the outcomes, while instrumental learning is at least potentially about maximizing rewards and minimizing punishment. Consequently, Pavlovian and instrumental conditioning can differ in the behaviors they produce, their underlying associative structures, and the role of reinforcement in establishing conditioned behavior (Rescorla & Solomon, 1967).

1.1.1 Pavlovian conditioning

In a typical Pavlovian conditioning experiment (eg. Yerkes & Morgulis, 1909; Holland, 1984, Figure 1.1a) neutral stimuli (such as lights or tones; called conditional stimuli or CSs) are presented with some prede-

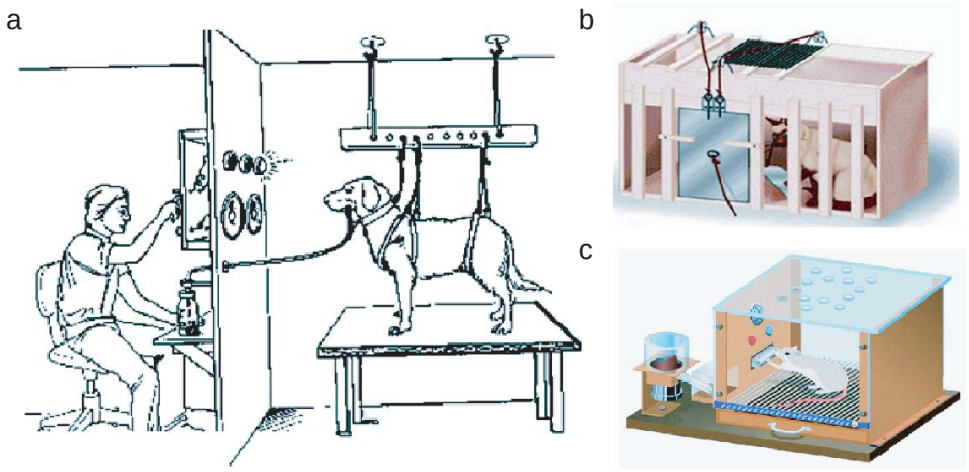

Figure 1.1: Basic animal conditioning paradigms. **a.** An illustration of Pavlovian conditioning: A dog is exposed to pairings of lights with the receipt of food. To assess conditioning, the light CS is turned on and salivation (the conditioned response) is measured in the absence of any food (ie, in extinction conditions, in order to prevent contamination of the CR with a similar UR). **b.** An illustration of a Thorndikean "puzzle box". Thorndike (1911) was the first to discover instrumental learning and describe its flexibility, based on the observation that cats can learn complex and arbitrary sequences of responses in order to escape from such boxes. **c.** An illustration of a modern instrumental learning scenario, a "Skinner box" in which a rat earns food by pressing a lever. Stimuli such as lights can be used to signal at what times leverpressing might be rewarded. Illustrations from the www.

fined probabilistic or deterministic temporal relationship to motivationally significant stimuli (eg. appetitive stimuli, such as food, or aversive stimuli, such as shocks; called unconditional stimuli or USs). Through repeated experience with this pairing between events, animals can learn to *predict* the occurrence of the US based on the occurrence of the CS, or to associate certain responses with the occurrence of the CS. That learning has occurred can be deduced from changes in the animal's behavior: as a result of this pairing, the once neutral stimuli come to elicit certain classes of (largely innate) responses. For instance, if food is always presented two seconds after a light turns on, the animal will gradually start responding to the light: approaching it and salivating when it turns on, indicating that the light has acquired a new significance (Figure 1.1a). The behaviors intrinsically elicited by the US (such as salivating for food or freezing for shock) are called 'unconditional responses' (URs), while the acquired responses to the CS are called 'conditioned responses' (CRs).

The paradigm of Pavlovian conditioning is very rich, and the learning processes involved are by no means trivial. Different types of predictive relationships (*excitatory*, in which a CS predicts the occurrence of a US, or *inhibitory*, in which the CS predicts that an otherwise available US will not appear) can be learned, using different classes of USs (broadly, *appetitive* or *aversive*), CSs that have acquired affective value through conditioning can support higher-order conditioning themselves, and the behavioral repertoire that results is diverse (eg. Holland & Rescorla, 1975a, 1975b; Rescorla, 1982). Some conditioned responses (termed 'preparatory responses') are generally responsive to the *class* of the predicted US, such as approach behavior to CSs predicting appetitive USs (or the removal of aversive USs), withdrawal from CSs predicting aversive USs (or the removal of appetitive USs), changes in heart rate and blood pressure, etc. Other conditioned responses (termed 'consummatory responses') are specific to the *identity* of the predicted US, such as blink-

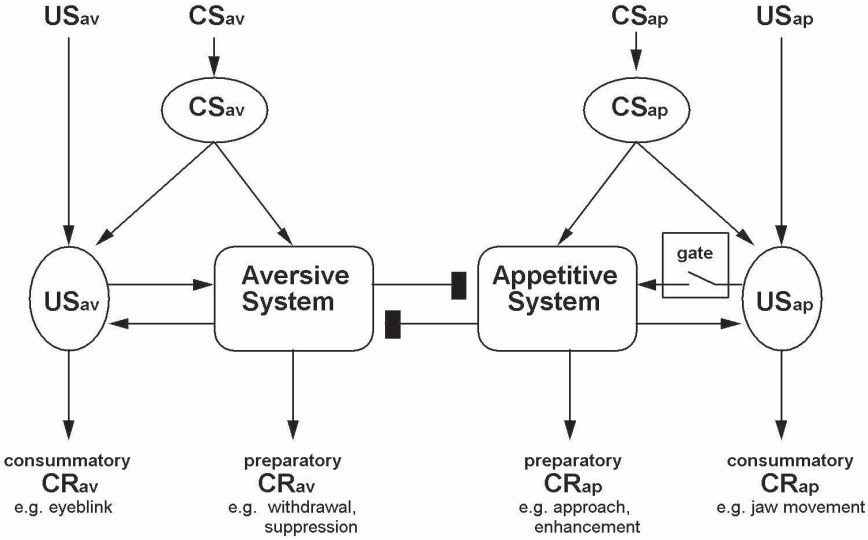

Figure 1.2: A scheme of the Konorskian model of associations underlying Pavlovian conditioning, reproduced from Dickinson and Balleine (2002). Abbreviations: ap, Appetitve; av, aversive; see text for other abbreviations and explanations.

ing to a CS predicting an air puff and salivating for a CS predicting food, but not for one predicting water (Holland, 1984). Within these US-specific responses are also 'stimulus substitution' effects in which the animal behaves toward the CS as if it was the US (for instance, licking a lighted key that predicts a water US; for some amusing examples see Breland & Breland, 1961).

The first to emphasize the theoretical implications of the distinction between consummatory and preparatory responding, was Konorski (1967). He proposed a model (Figure 1.2; reviewed in Dickinson & Balleine, 2002) according to which Pavlovian conditioning results in at least two types of independent associations. One is between a representation of the CS and a representation of the sensory properties of the US, giving rise to US-specific consummatory responses. A second association is between the CS representation and a generic motivational system (either appetitive or aversive, according to the motivational properties of the US), giving rise to preparatory responses. Other lines of work suggest additional types of associations, such as direct associations between the CS and the CR, and associations between two CSs. A detailed description of these different associations, and the data bearing on their existence, is beyond the scope of this chapter. Important for our topic, however, is the link to motivation. Several lines of evidence point to the existence of *two opponent motivational centers*, an aversive one and an appetitive one. One example for such generality (within a motivational category) and opponency (between aversive an appetitive categories) is that a CS which has been conditioned to food (an appetitive US) can easily be conditioned to a different appetitive US (such as water), but it is very difficult to condition it to an aversive US such as shock. In line with common sense, the appetitive motivational system extends to the removal of an aversive US, and the aversive system is involved in conditioning as a result of the omission of an appetitive US. Thus it is easy to condition the above mentioned (appetitive) CS to the removal of a shock.

Note that the conditioned responses that result from these different types of associations are "hard-wired"

(often highly adaptive) behaviors whose appropriateness is determined more by evolutionary processes, than by individual learning (Dayan et al., 2006). Their particular characteristic is that they are performed irrespective of their appropriateness in terms of gaining access to reward or avoiding punishment (Breland & Breland, 1961). This is most convincingly demonstrated by 'omission schedules' in which the US is explicitly *not* given if the conditioned response is emitted. For instance, in the case of pigeons trained with a pairing of a light and food, in an omission schedule the food US will be omitted whenever the pigeon pecks at the light CS (Williams & Williams, 1969). That Pavlovian behavior does not disappear despite such adverse consequences is, in fact, the hallmark of truly Pavlovian responding.

1.1.2 Instrumental conditioning

In contrast to Pavlovian responding, the hallmark of instrumental behavior is its flexibility. Instrumental learning allows animals to acquire new behavioral strategies and to choose actions so as to attain essential commodities or goals (Thorndike, 1911). While a purely Pavlovian animal must rely on evolutionary processes to ensure that the responses elicited by stimuli are appropriate for coping with the predicted events, and is thus at the mercy of an unstable environment, instrumental learning allows animals to control their environment in the service of their needs and desires (Dickinson, 1994). In an instrumental conditioning scenario, a specific action or a series of actions (denoted A for 'action' or R for 'response', for instance, pressing a lever), bring about a desirable outcome (O, for instance, food) according to some experimenter-defined schedule of reinforcement,[1] in the presence of specific stimuli (S, for instance, an experimental chamber with a lever in it; Figure 1.1b,c). An outcome that causes an increase in the performance of its preceding actions is called a 'reinforcer', as it reinforces the occurrence of those preceding behaviors.

What associative structure underlies the box-opening sequence performed by the cat in Figure 1.1b? One option, espoused by Thorndike (1911) and Skinner (1935), is that the cat has learned to associate this particular box with this sequence of actions. According to this view, the role of the desired outcome (the opening of the box's door) is to "stamp in" such stimulus-response (S-R) associations. A different option, advocated by Tolman (1948) (and later demonstrated by Dickinson and colleagues, see Dickinson, 1997), is that the cat has learned that this sequence of actions leads to the opening of the door, that is, an action-outcome (A-O) association. The critical difference between these two views is the role of the reinforcer: in the former it only has a role in learning, but once learned, the behavior is rather independent of the outcome or its value; in the latter the outcome is directly represented in the association controlling behavior, and thus behavior should be sensitive to changes in the value of the outcome. For instance, if a dog is waiting outside the box, such that opening the door is no longer a desirable outcome to the cat, according the S-R theory the cat will nevertheless perform the sequence of actions that will lead to the door opening, while A-O theory deems that the cat will refrain from this behavior. Research in the last two decades has convincingly

[1] A wide variety of schedules of reinforcement have been used in conditioning studies, such as: reinforcing every n-th response; reinforcing every response that is at least t seconds from the previous reinforcer; reinforcing every correct response in a situation in which two options are available and a stimulus (say, the color of a light) serves to signal the correct response in each trial, etc. In Chapter 2 one class of reinforcement schedules, namely, free operant schedules, will be discussed in depth.

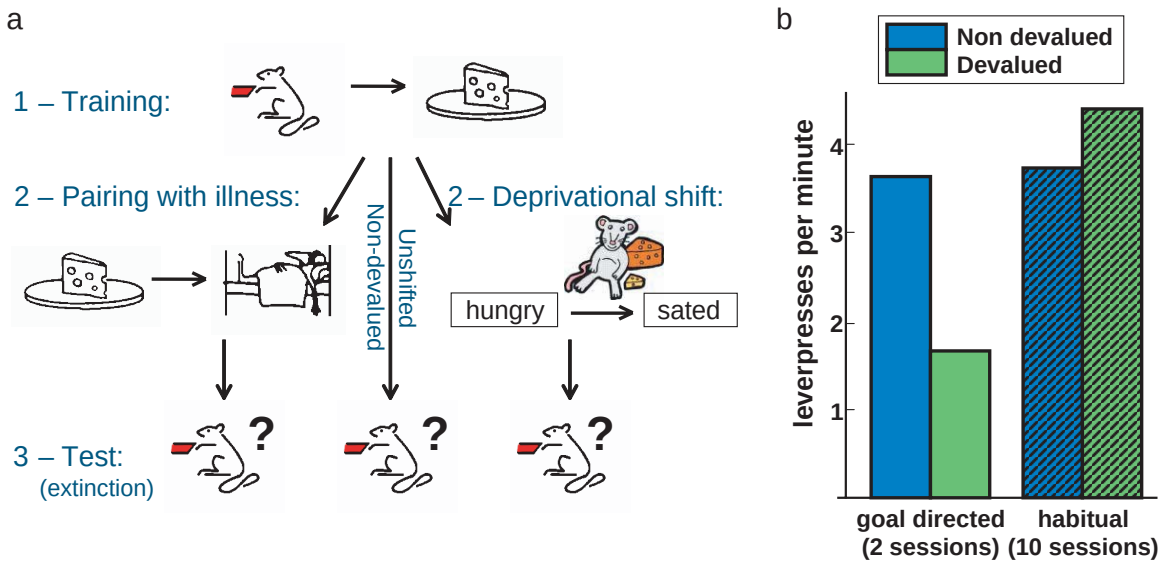

Figure 1.3: a. Experimental techniques for outcome revaluation: In a typical outcome revaluation experiment, rats are first trained to perform an instrumental action (here, pressing a lever) in order to obtain a desired outcome (phase 1: Training). In phase 2 the outcome value is manipulated by, for instance, pairing its consumption with illness (left) or inducing a shift in motivational state, such as from hunger to satiety (right). In a third phase (Test) the trained response is tested in extinction (ie, with no outcomes available), and compared to that of rats who have not undergone phase 2. Rat cartoons adapted from Dickinson and Balleine (1994). **b.** Habitual and goal-directed instrumental behavior: When hungry rats are trained to press a lever in order to obtain sucrose pellets, post training devaluation of the pellets by conditioned taste aversion (green) causes a reduction in lever-pressing (compared to rats for whom the outcome was not devalued, blue) only after moderate training, when responding is still goal-directed (left, solid). This is no longer the case after extensive training, at which point the behavior becomes habitual (right, hatched) and insensitive to changes in the utility of the outcome. In all cases behavior was tested in extinction. Adapted from Dickinson (1985).

shown that *both* types of control structures exist. In fact, instrumental behavior can be subdivided into two sub-classes, goal directed and habitual behavior, based exactly on this distinction (Dickinson & Balleine, 2002). A subdivision based on outcome sensitivity suggests that these types of instrumental responding will be differentially affected by motivation. This will be discussed in detail in Chapter 3, but first, let us define goal-directed and habitual behaviors precisely.

1.1.3 Goal directed versus habitual instrumental behavior

An important observation is that although all instrumental behavior is instrumental in achieving its contingent goals, it is not necessarily purposively *goal-directed*. Dickinson and Balleine (1994, 2002) proposed that behavior is goal-directed only if: (i) it is sensitive to the contingency between action and outcome, and (ii) the outcome is desired. Based on the second condition, outcome revaluation manipulations have been used to distinguish between two systems of action control: An action that is sensitive to revaluation is considered goal-directed, but if it persists despite the fact that the instrumental outcome is no longer a valued goal (for instance, food for a sated animal), then the action must not be goal-directed.

In a typical *outcome revaluation* experiment (Figure 1.3a), food-deprived rats are trained to perform an instrumental action (such as leverpressing) to obtain a rewarding outcome (food). A post-training stage then modifies the value of the outcome for one group of rats. The consequences of this manipulation are tested by comparing the propensity of these rats to perform the instrumental response, to that of a control group for whom the outcome has not been revalued. Importantly, this is done in extinction, ie, in the absence of rewards, to test for the effects of the revaluation on any *previously learned* associations between the action and the outcome. Of course, as a result of the extinction process, responding gradually declines, but this is true for both revalued and control groups. A significant between-group difference is thus evidence for sensitivity of the instrumental behavior to the change in the value of the outcome.

Three methods are commonly used for outcome revaluation: In a *specific satiety* procedure (Colwill & Rescorla, 1985; Balleine & Dickinson, 2000; Corbit et al., 2001; Killcross & Coutureau, 2003; Yin et al., 2005, 2005), rats are pre-fed on one outcome, such that they develop a temporary, outcome-specific satiation for this outcome. Consumption tests show that such a procedure selectively devalues only the pre-fed outcome. Another method for devaluing a specific outcome is by *conditioning taste aversion* to it (Adams & Dickinson, 1981; Adams, 1982; Colwill & Rescorla, 1985, 1988; Yin et al., 2004; Holland, 2004). In this procedure, after the rat consumes the outcome, gastric illness is induced (typically by injection of lithium chloride), rendering the food aversive to the rat. Finally, *shifts in primary motivational state* (Dickinson & Dawson, 1988, 1989; Balleine, 1992; Lopez et al., 1992; Dickinson et al., 1995; Balleine et al., 1995; Balleine & Dickinson, 2000) can either devalue or enhance the value of outcomes. Most commonly, after training rats to leverpress when hungry, their motivational state is shifted to that of satiety by allowing consumption *ad lib* of lab-chow in the home-cage. This manipulation renders the once very valuable food reward less valuable. Opposite shifts (from training when sated to testing when hungry) enhance the value of the instrumental outcome, and shifts between different motivational states (for instance, between hunger and thirst) can be used to devalue one outcome (say, food pellets) while maintaining the value of another (eg, sucrose solution).

Indeed, after moderate amounts of training, outcome revaluation brings about an appropriate change in instrumental actions (eg. Adams & Dickinson, 1981; Dickinson & Nicholas, 1983b, left bars in Figure 1.3b). This suggests that, at least during acquisition, the performance of instrumental actions depends on encoding of the relationship between actions and their consequent outcomes, and can be considered goal-directed (Dickinson, 1985; Dickinson et al., 1998, and see Dickinson & Balleine, 1994 for a review). However, this is no longer the case for extensively trained responses. When instrumental actions are over-trained, they lose sensitivity to outcome devaluation (eg. Adams, 1982, right bars in Figure 1.3b). Rather, they seem to be triggered by the stimuli in whose presence they had been repeatedly performed, autonomous of the value of the goal (Dickinson, 1985; Dickinson et al., 1998). That extensive training can render an instrumental action independent of the value of its consequent outcome, has been regarded as the experimental parallel of the folk psychology maxim that well-performed actions become habitual (Dickinson, 1985).

Interestingly, moderately trained actions can also seem indifferent to the value of their consequent actions

in a contrived setting in which the value of the outcome is altered, but without giving the animal the opportunity to experience the new value, for instance, using revaluation manipulations such as motivational shifts (in which the animal can be sated on a food different from that used in the experiment) and conditioned taste aversion (by conditioning the aversion with only one post-ingestion injection of lithium-chloride). In this case, testing the willingness of the animal to perform the instrumental task in extinction (ie, with no experience with the new value of the outcome) reveals performance that is similar to that of animals for which the outcome has not been revalued. That is, after having been trained when hungry to press a lever for a food reward, rats that have been shifted to satiety will continue to press the lever as eagerly as rats who are still hungry. However, mere exposure to the outcome and its new value (for instance, allowing the rat to consume the outcome in its home-cage, outside the experimental scenario, in a so-called 'incentive learning' stage) is enough to render the instrumental response sensitive to the new value of the outcome. For goal-directed behavior to be influenced by the new value of the outcome, this 'incentive learning' can occur at any time, either before or after instrumental training and revaluation (Dickinson & Dawson, 1988, 1989; Balleine & Dickinson, 1991; Balleine, 1992).

To summarize, goal-directed behavior can be operationally defined as such behavior whose performance is sensitive to the current value of the outcome, given appropriate incentive learning of the outcome value. In contrast, habitual behavior does not show immediate sensitivity to the new value of a revalued outcome, even in the face of incentive learning, but rather requires new (rewarded) learning in order to adjust to the new situation. Other tests for goal-directed control examine sensitivity to the contingency between an action and its outcome. For instance, in a *contingency degradation* manipulation, after the animal is trained to perform an action in order to obtain an outcome, the outcome is made available even when the action is not performed (eg. Balleine & Dickinson, 1998; Corbit et al., 2001; de Borchgrave et al., 2002; Yin et al., 2005). If performance of the action persists, it is not goal-directed. These tests also show that early in training behavior is often goal-directed, only to become habitual through over-training.

In addition to the amount of training, a number of factors have been shown to promote either habitual or goal-directed instrumental behavior. First, in tasks in which an animal can execute two different actions to obtain two different rewards, extensively trained actions remain sensitive to outcome devaluation, indicating a dominance of goal-directed control (Colwill & Rescorla, 1985, 1988; Colwill & Triola, 2002; Holland, 2004). Second, if the animal must perform a chain of actions in order to obtain a reward (for instance, press a lever and then move a flap covering the food magazine, or, pull a chain and then press a lever, Balleine et al., 1995), actions more proximal to the reward are more sensitive to devaluation than actions farther from the reward (Killcross & Coutureau, 2003; Balleine et al., 1995). Finally, behavior that is trained with an interval schedule of reinforcement (ie, the first response after a random interval of mean length t, is rewarded) habitizes more rapidly than behavior that is trained with a ratio schedule of reinforcement (ie, every n-th response, on average, is rewarded; Dickinson et al., 1983). Interestingly, Pavlovian behaviors also show similar patterns of sensitivity to outcome revaluation (eg. Holland & Rescorla, 1975b; Holland & Straub, 1979), however, few studies have explored these effects and probed the suggested parallels between Pavlovian and instrumental modes of action control.

1.1.4 Interactions between Pavlovian and instrumental conditioning

Despite their separation in operational experimental terms, Pavlovian and instrumental conditioning are tightly intertwined as two learning processes that interact, compete and rely on each other. First, every instrumental learning scenario in which an animal learns to perform an action in order to obtain an outcome, also includes Pavlovian contingencies — predictive relationships between different co-occurring cues, which give rise to Pavlovian conditioning. As a result, the animal's behavior will be comprised of both instrumental actions and Pavlovian responding. When these are mutually facilitating (such as Pavlovian approach behavior to a lever which, when pressed, predicts the occurrence of food), instrumental learning will benefit from the incidental Pavlovian relationships. In fact, experimenters routinely make use of such interactions in order to speed instrumental learning, through gradual shaping of arbitrary instrumental actions from more innate Pavlovian responses, and even in the design of the experimental apparatus (eg, placing the food magazine in close proximity to the instrumental manipulanda). One implication of such interaction is that not all responses that are instrumental in obtaining outcomes are necessarily controlled by instrumental learning mechanisms. In 'autoshaping' (Brown & Jenkins, 1968) a seemingly instrumental response such as key-pecking or leverpressing is 'automatically shaped' through the facilitation of Pavlovian responding. Only an omission schedule can reveal that, in fact, the response is controlled by a Pavlovian contingency (Williams & Williams, 1969).

However, in many situations Pavlovian responses and instrumental actions are incompatible, and so they compete with each other. For example, it is difficult to train a rat to press a lever in order to turn off an aversive loud noise, if in order to do so the rat must approach the source of the noise. Breland and Breland (1961) complain that "In our attempt to extend a behavioristically oriented approach to the engineering control of animal behavior by operant conditioning techniques, we have fought a running battle with the seditious notion of instinct." In their endeavors to train animals to perform sequences of actions for an audience, Pavlovian responses arising from appetitive stimulus substitution frequently resulted in the gradual deterioration (rather than perfection through practice) of the instrumental response.

A clear demonstration of the interaction between instrumental and Pavlovian conditioning is the phenomenon of Pavlovian-instrumental transfer (PIT). In PIT, stimuli conditioned to predict the occurrence of affectively significant outcomes influence the vigor of otherwise unrelated instrumental responding (Estes, 1948; Lovibond, 1983). The procedure involves three steps: Pavlovian training (say, pairings of a tone and sucrose), instrumental training (say, pressing the left lever for one outcome, and the right lever for another outcome), and finally, and extinction test in which leverpressing is assessed in the presence of the tone, and without it. There are two sorts of PIT (Cardinal et al., 2002; Holland, 2004): *specific*, in which the CS differentially affects instrumental responding for a similar outcome rather than for a different outcome (in the example above, if the two instrumental outcomes were sucrose and food pellets, the CS would enhance leverpressing for sucrose but not for food pellets), and *general PIT*, in which a CS influences all instrumental actions regardless of their outcome (for instance, if the instrumental training above was for food pellets and chocolate milk outcomes, which are both different from the Pavlovian US, but belong to the same appetitive

motivational class, and providing the Pavlovian US is relevant to the animal's current motivational state). The division into two opponent motivational systems holds in this case as well: as an appetitive Pavlovian CS enhances instrumental responding for appetitive outcomes, a Pavlovian CS predicting an aversive outcome suppresses such instrumental responding. In fact, the latter provides the conventional measure for the strength of aversive Pavlovian conditioning.

1.2 Computational models of animal conditioning

The last two decades have seen a proliferation of computational models that target conditioned behavior as a well-controlled instance of real-world decision-making. Within a taxonomy of computational models of learning, conditioning constitutes an example of learning with *a scalar reinforcement signal*, or *'reinforcement learning'* (RL). This is categorically different from, and lies in between, learning with an explicit teaching signal (as in 'supervised learning' models, common in artificial intelligence applications), and learning of input statistics without any supervisory signal (as in 'unsupervised learning' models, for instance of sensory processes). Thus algorithms and theory have been developed specifically for the case of RL. The core ideas of RL come from the psychological literature on Pavlovian and instrumental conditioning reviewed above (Sutton & Barto, 1990), as well as the substantial engineering (Bertsekas & Tsitsiklis, 1996), statistical and computational theory on optimizing adaptive control (Sutton & Barto, 1998). In general, the computational distinction between instrumental and Pavlovian conditioning comes down to the difference between learning to predict which events ('states') will follow one another[2], and learning to behave (learning an 'action policy') such as to influence the probability that certain events will or will not occur.

The reinforcement learning formulation is as such: let $\mathcal{S}$ be the situations or states of the animal's (external and internal) environment, $\mathcal{A}$ the possible actions (which can differ from state to state, in which case they are denoted $\mathcal{A}_S$), $p(S_{t+1}|S_0,...,S_t,a_0,...,a_t)$ the probability of transitioning from a state at time t ($S_t \in \mathcal{S}$) to a subsequent state ($S_{t+1} \in \mathcal{S}$) given the history of states and actions, and $p(U_r|S_0,...,S_t,a_0,...,a_t)$ the probability of encountering an affectively significant outcome of scalar value U_r (negative for aversive and positive for appetitive outcomes)[3] given the past (and current) states and actions. For (Pavlovian) prediction learning without action selection, $p(S_{t+1}|S_0,...,S_t)$ are the world-generated transition probabilities between states, and $p(U_r|S_0,...S_t)$ is the probability of encountering an affectively significant outcome, regardless of any action taken.

The environment is called *Markovian* if $p(S_{t+1}|S_0,...,S_t,a_0,...,a_t) = p(S_{t+1}|S_t,a_t)$, that is, if the probability of transitioning from a state $S \in \mathcal{S}$ to a subsequent state $S' \in \mathcal{S}$ (which we will shorthand as $\mathcal{T}^a_{S\to S'}$) does not depend on the history of previous actions or states.[4] In addition, for the Markov property to hold,

[2] In treating Pavlovian conditioning as prediction learning, computational models unfortunately concentrate on assumed quantities (predictions, which can not be measured directly) while offering little insight into the empirically observed Pavlovian *responses*.

[3] We denote the outcome 'r', unfortunately suggesting 'reward', because the affective-neutral 'o' is easily confused with zero.

[4] In a *stationary* Markovian environment, transition probabilities are also independent of time, as suggested by the shorthand notation.

the probabilities of obtaining different outcomes when in a certain state, shorthanded $P_r^a(S)$, must also be history-independent, that is, $p(U_r|S_0,...,S_t,a_0,...,a_t) = p(U_r|S_t,a_t)$. In this case, the setting is that of a *Markov decision process* (MDP; or *Markov process* for a Pavlovian setting), with dynamics as follows: at each discrete timestep the (simulated) animal receives information about the current state of the environment[5], decides which action to perform, and then observes the immediate reward/punishment that results and transition to the next state. A variation on this, which we will use in this thesis due to its relevance to decision-making in real-world scenarios, is a *semi-Markov decision process* (SMDP) in which transitions between states do not happen at every timestep, but rather the system dwells in each state for some (possibly stochastic and/or continuous) time before the next action and state transition. For didactic purposes, in this chapter we will discuss the simpler case of MDPs. The extension to SMDPs will be detailed in Chapter 2.

The traditional RL problem is to find an action *policy* that will maximize the sum of future rewards. This optimized sum can be defined over a finite horizon (such as a trial or an experimental session), or over an infinite horizon (all future rewards/punishments). In the latter case, in order for the sum of expected rewards to converge, it is computationally necessary to introduce *discounting* of future rewards, that is, a weighting of future rewards that decreases with temporal distance, such that the weighted sum of all future rewards in finite. The commonly used *exponential discounting* assigns exponentially diminishing weights γ^t (where t is time, and $0 < \gamma \leq 1$ is the discount factor; the closest γ is to 0 the steepest the temporal discounting)[6] to future rewards. An alternative optimization criterion, which is uniquely suitable for infinite horizon scenarios and does not require temporal discounting, is to maximize the *average rate of reward* per timestep. This method, called average-reward reinforcement learning (ARL), will be discussed in detail in Chapter 2.

In such an MDP or Markov process, finding the optimal predictions and optimal action policies constitutes of a *Dynamic Programming* problem. The solution can be derived using either *model-based* RL techniques (also called Dynamic Programming techniques) or model-free *temporal difference* methods. Below, I will first briefly describe these methods and how they are used for prediction learning, and then discuss their extension for policy learning and action selection. I will follow this with a more detailed discussion of how these two aspects of learning are combined in a specific implementation called the *Actor/Critic* model. The Actor/Critic framework will be highlighted not due to its theoretical prominence (in fact, in terms of convergence guarantees it is unfortunately the weakest of the commonly used RL algorithms), but rather because the algorithmic implementation of the model presented in this thesis, which I will describe in Chapter 5, is an instantiation of this framework, and because the Actor/Critic is the RL implementation that has been most strongly linked to prediction learning and action selection in the brain. Finally, I will turn back to the psychological distinction between different types of responding, and relate Pavlovian responding, instrumental goal-directed actions and habitual behavior, to their suggested computational counterparts.

[5] If this information is incomplete, ie, the exact state of the environment is not known and must be estimated by the animal (for instance, from noisy sensory information), the state is considered to be 'partially observable'. Otherwise it is a 'fully observable' state. Although a partially observable state space is more realistic in real-world decision-making situations, in this thesis we will deal only with observable state-spaces which render finding optimal solutions to the RL problem computationally tractable.

[6] An equivalent (but less commonly used) formulation of exponential discounting, which we will use due to its more straightforward application to SMDPs, is to weight future rewards by $e^{-\gamma t}$ (with $\gamma \geq 0$), in which higher values of γ mean steeper discounting.

1.2.1 Optimal predictions

What might an animal want to predict in a Pavlovian scenario? Given the current state (stimulus) S, it might be useful to predict what states and outcomes are likely to occur in the future (what CSs and USs are likely to appear). Alternatively, it might be useful to summarize the expectations for the future in a measure of how good or bad it is expected to be. Even a coarse predictive mechanism such as this, coupled with appropriate innate behaviors (such as approach responses to stimuli or events which predict good future outcomes and withdrawal and engagement in defense behaviors in the face of stimuli predicting bad outcomes), could confer considerable advantages to an organism attempting to survive in a frequently hostile world.

If the animal knows the transition probabilities $\mathcal{T}$ and the immediate outcome probabilities P_r, ie, if the environment dynamics are known, a simple way to create predictions for future events is by unfolding the tree of possible options in a 'simulation' of the future. That is, given the current state S, the probability of transitioning to each state S' is $\mathcal{T}_{S\to S'}$, making the expected affective value of the subsequent timestep

$$\langle r_{t=1}|S_0 = S\rangle = \sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'} \cdot U_r \cdot P_r(S'). \tag{1.1}$$

It is easy to extend this another timestep into the future, where the expected affective value would be $\langle r_{t=2}|S_0 = S\rangle = \sum_{S''\in\mathcal{S}} \sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'}\mathcal{T}_{S'\to S''} \cdot U_r \cdot P_r(S'')$, and so forth. In this way, the animal can project any number of steps into the future to predict what events will happen and what their affective value will be.

However, this method has several limitations. First, it is computationally laborious. Second, it relies on knowledge of the dynamics of the environment, which is unreasonable in animal and human prediction learning. We will now first describe how model-based RL techniques solve these problems, and then discuss model-free temporal difference methods for prediction learning without a model of the environment.

Model based RL

How can an animal make predictions if it does not know the dynamics of the environment? One possibility is to learn a *model* of the environment from experience, that is, to observe and keep track of the relationships that occur in the environment (which state follows which, and what are the outcomes in each state), and estimate the world model ($\hat{\mathcal{T}}_{S\to S'}$ and $\hat{P}_r(S)$, where the hat denotes an estimated quantity) from these. Using this estimated model, one could proceed to roll-out the potential consequences of a state (as in Equation 1.1), at least for a finite number of forward-looking steps. This type of 'forward-model' prediction has been implicated in goal-directed control (Daw, Niv, & Dayan, 2006), as will be discussed in Section 1.2.3 below.

Rather than unroll a (possibly infinitely long) forward branching tree of possible events, RL techniques provide a shortcut to computing an important quantity: the expected discounted sum of all future outcomes, given an initial state. This quantity, called the *value* of the state and denoted $V(S)$, turns out to be a very

useful quantity in terms of action selection. As will be discussed below, (instrumental) behavior based on such values can easily achieve optimality, in terms of reaping the largest possible amount of rewards. The computational shortcut to computing state values is based on the observation that correct state values satisfy a specific consistency structure. Mathematically writing the value of the current state S_t as:

$$V(S_t) = \sum_{i=t}^{\infty} \gamma^{i-t} \langle r_i | S_t \rangle = \langle r_t | S_t \rangle + \gamma \langle r_{t+1} | S_t \rangle + \gamma^2 \langle r_{t+2} | S_t \rangle + \ldots \tag{1.2}$$

we can see that, by definition,

$$V(S_t) = \langle r_t | S_t \rangle + \gamma \sum_{S_{t+1} \in S} \mathcal{T}_{S_t \to S_{t+1}} V(S_{t+1}) = \langle r_t | S_t \rangle + \gamma \langle V(S_{t+1}) \rangle_{\mathcal{T}} . \tag{1.3}$$

This consistency between state values, that is, the fact that the value of a state equals the expected immediate outcome in that state plus the expected discounted value of the next state, allows the efficient computation of values, obviating the computation of an infinite sum. For example, in *value iteration*, state values are computed by starting from zero initial values and updating each of the values iteratively according to equation (1.3), until the values converge, that is, until the consistency relations between them are fulfilled (for a more detailed exposition, see Sutton & Barto, 1998). This algorithm is proven to converge, and the resulting values are the true state values, that is, they predict the true discounted sum of future rewards when starting from a state and evolving according to the world model thereafter (Bertsekas & Tsitsiklis, 1996). Note that for a finite horizon case, this method is equivalent to the forward-model computation discussed above.

Model free RL: Temporal Difference learning

Value iteration is a 'model-based' method because the learning process relies on knowledge of $\mathcal{T}$ and P_r, the model of the environment's dynamics. A different option, is to learn the state values directly from experience with the environment, bypassing the need to estimate a world model. In these 'model free' methods, the inconsistency between consecutive values (or predictions) is used to compute a *temporal difference prediction error* according to which predictions are then updated (Barto et al., 1983; Sutton, 1988; Watkins, 1989). The consistency relationship in equation (1.3) provides the basis of this prediction error: because only the true values fulfill this relationship, for all other (potentially incorrect) estimated values (which we will denote $\hat{V}$) an inconsistency error can be computed

$$error = \langle r_t | S_t \rangle + \gamma \left\langle \hat{V}(S_{t+1}) \right\rangle_{\mathcal{T}} - \hat{V}(S_t). \tag{1.4}$$

In temporal difference (TD) learning, outcome utilities and state transitions sampled from the environment's dynamics are used to compute a stochastic approximation to the above prediction error

$$\delta(t) = U_r(t) + \gamma \hat{V}(S_{t+1}) - \hat{V}(S_t) \tag{1.5}$$

where $\hat{V}(S_t)$ is the estimated value of the current state, $U_r(t)$ is the (possibly zero) outcome sampled at this state, and $\hat{V}(S_{t+1})$ is the estimated value of the state to which the environment has just been observed to transition. This prediction error is used to update (learn) $\hat{V}(S_t)$, the prediction that led to the error, according to

$$\hat{V}(S_t) \longleftarrow \hat{V}(S_t) + \eta \cdot \delta(t) \tag{1.6}$$

where $0 < \eta \leq 1$ is the learning rate or step-size parameter. Note that in a deterministic environment, there will be no prediction errors once the correct values are achieved, and the prediction learning process will self-terminate. In stochastic environments the updates are noisy and values will not converge unless the learning rate parameter is properly decayed throughout the learning process. In both cases, to ensure convergence to the true state values, all states must be sampled infinitely often (Sutton & Barto, 1998).

The *temporal difference prediction error* (equation 1.5) is a key theoretical construct in computational models of conditioned behavior (Barto et al., 1983; Sutton, 1988), and provides the basis for the Critic's learning rule in Actor/Critic models (Barto et al., 1983; Barto, 1995), which will be discussed below. Using this error term, the *temporal difference learning rule* (equation 1.6), originally suggested as a model of Pavlovian prediction learning on the basis of temporally complex conditioning experiments (Sutton & Barto, 1987; see Sutton & Barto, 1990, for an account of the development of TD learning), is an extension of the well-known Rescorla-Wagner (1972) learning rule for formation of associations in Pavlovian conditioning. Different from the Rescorla-Wagner rule that can only explain prediction learning on a trial-by-trial basis, and can not explain phenomena such as second-order conditioning, temporal difference learning can account for precise temporal predictions within a trial, and naturally supports high-order conditioning (Sutton & Barto, 1990).

The hallmark of temporal difference prediction errors is that they occur only when events are not predicted. For instance, in a simulated Pavlovian conditioning scenario in which a tone CS is followed two seconds later by a food US, prediction errors arise as a result of the unexpected US early in training (Figure 1.4a), but not later in training when the US is predicted (Figure 1.4b). If the CSs occur randomly, at late stages of training they themselves generate a prediction error (similar to the one that had previously accompanied the US delivery), which can support second order conditioning. In trials in which the US is not delivered, a negative prediction error occurs at the precise time of the expected US delivery (Figure 1.4c; such precise timing also necessitates a stimulus representation that can track elapsed time, as detailed in the figure caption).

Note that, different from many other theories of Pavlovian conditioning, temporal difference learning is a *normative* theory. In the words of Sutton and Barto (1990), "A distinguishing feature of the TD model is that it is based on a theory about the *function* of classical conditioning. It is based on the supposition that the goal of learning is to accurately predict at each point in time the imminence-weighted sum of future US intensity levels. The TD model thus both predicts features of classical conditioning behavior and provides an account of their function as part of a mechanism for accurate prediction" (pp. 532).

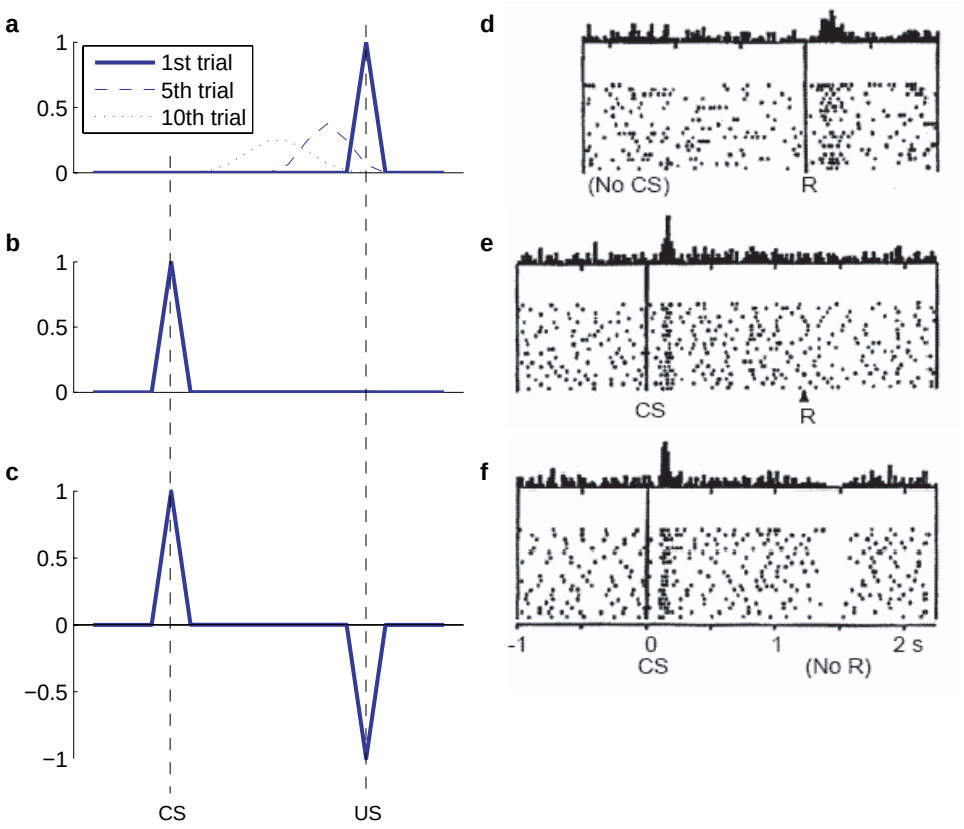

Figure 1.4: a-c. Temporal difference prediction errors in a simple Pavlovian conditioning task. A tone CS is presented at random times, followed two seconds later with a food US. In the beginning of training **(a)**, the affectively significant US is not predicted, resulting in prediction errors. As learning progresses (trials 5 and 10 are plotted as examples) the prediction error propagates back (Niv et al., 2005) as values of preceding timesteps are updated (equation 1.6). When the predictive relationships are completely learned **(b)**, the now-predicted US no longer generates a prediction error, rather, the unpredicted occurrence of the CS is accompanied by a prediction error. If the US is unexpectedly omitted **(c)**, a negative prediction error is seen at the time in which the US was expected, signaling that expectations were higher than reality. In these simulations the CS was represented over time by using the common serial compound state representation (Kehoe, 1977; Sutton & Barto, 1990), and there was no discounting ($\gamma = 1$) because the predictions were for a finite horizon (the length of a trial). **d-f.** Firing patterns of dopaminergic neurons in the ventral tegmental areas of monkeys performing an analogous instrumental conditioning task. Each raster plot shows action potentials (dots) with each row representing a trial, aligned to the time of the cue (or the reward). Bar histograms show the summed activity over the trials plotted below. When a reward is given unexpectedly, dopaminergic neurons respond with a phasic burst of firing **(d)**. However, after conditioning with a predictive visual cue (which, in this task, predicted a food reward if the animal quickly performed the correct reaching response), the predicted reward no longer elicits a burst of activity, and the phasic burst now accompanies the presentation of the predictive cue **(e)**. In 'catch' trials, in which the food reward was unexpectedly omitted, dopaminergic neurons showed a precisely-timed pause in firing, below their standard background firing rate **(f)**. Data from Schultz et al. (1997). Note that the discrepancies between the simulation and the experimental firing patterns in terms of the magnitude and spread of the prediction errors at the time of the reward likely result from the temporal noise in reward delivery in the instrumental task, and the asymmetric representation of negative and positive prediction errors around the base firing rate (Niv et al., 2005).

1.2.2 Optimal action selection

Rather than merely respond to the states of the world, actions that affect the frequency of occurrence of different states allow animals to directly influence the amount of rewards that they encounter. An action selection policy (denoted π) is the probability of choosing each action in each state $\pi(a|S) = P(a|S)$, and an optimal policy π^* is one that achieves the maximal amount of rewards possible in the environment.

The problem of learning an optimal policy is especially difficult in those (very common) cases in which actions can affect long-term outcomes, or in which an outcome depends on a series of actions. For example, when winning or losing a game of chess, it is not at all simple to infer which were the actions responsible for this outcome, in order to improve the playing policy. This is true in the animal domain as well: when reaching a dead-end in a maze, how will a rat know which of its previous actions was the erroneous one? And conversely, when it finds the cheese in the maze, how will it know which actions should be credited with the success? This is the (in)famous *credit assignment problem* (Barto et al., 1983; Sutton & Barto, 1998). RL methods solve the credit assignment problem by basing action selection not only on immediate outcomes, but also on value predictions, which embody long-term predictions of outcomes.

This suggests a straightforward solution: a good proxy for the long-term rewards contingent on performing an action, is the immediate reward plus the value of the successor state, so if the values of different states of the environment are known (for instance, they were learned through one of the methods described in the previous section), one should choose actions that will most likely result in the largest immediate reward plus successor state value. That is, if a model of the environment is known (or can be estimated), a simulation of the immediate consequences of an action (in terms of the expected immediate reward and the expected value of the successor state) can be used in order to estimate the value of the action, in a form of 'model-based' action selection. The wrinkle in this is that the state values themselves are dependent on the expected state transitions, that is, on the action selection policy, and thus they must be updated concurrently with the policy.

A variety of RL algorithms have been proposed that solve this problem. These are based on iteratively learning the values V^π corresponding to a policy π, and then updating the policy such that it chooses actions greedily according to these values (ie, chooses in each state the action that will lead to the highest expected value). This is repeated until the policy can no longer be improved, that is, until the policy is greedy with respect to the state values it engenders. Such consistent policy and state values are only possible for an optimal policy π^*, and the corresponding state values are the optimal state values V^* (Bertsekas & Tsitsiklis, 1996). If a model of the environment is available, such a method can be used to learn the optimal state values and the corresponding optimal policy directly, by iterating[7] on the consistency equation for the optimal values, called the "*Bellman equation*" of the MDP:

$$V^*(S) = \max_{a\in\mathcal{A}} \left\{ \sum_{U_r} P_r^a(S)\cdot U_r + \gamma \sum_{S'\in S} \mathcal{T}^a_{S\to S'} V^*(S') \right\} = \max_{a\in\mathcal{A}} \left\{ \langle r|S\rangle_{P_r^a} + \gamma \langle V^*(S')|S\rangle_{\mathcal{T}^a} \right\}. \tag{1.7}$$

[7] To iterate, state values are intialized at zero, and then updated one by one according to the right side of the equation, to convergence.

A rapidly converging variant called 'policy iteration' uses a similar policy-dependent consistency equation

$$V^{\pi}(S) = \sum_{a \in \mathcal{A}} \pi(a|S) \left[\sum_{U_r} P_r^a(S) \cdot U_r + \gamma \sum_{S' \in S} \mathcal{T}^a_{S \to S'} V^{\pi}(S') \right] \tag{1.8}$$

to iteratively update values according to the current policy (typically using one update iteration rather than letting the values converge), and then update the policy to the greedy one corresponding to these values.

State-action values

An alternative is to combine the two stages of the learning process, and instead of learning state values and choosing actions based on these, learn *state-action values* (called Q-values; Watkins, 1989). A $Q^{\pi}(S,a)$ value is defined as the expected cumulative future reward when taking action a at state S and from then on selecting actions according to policy π. An optimal $Q^*(S,a)$ value is the value of taking action a at S and behaving according to the optimal policy thereafter. Such values can be learned in a very similar way to state values, through iterations on a policy-dependent consistency equation, or on the Bellman equation corresponding to the optimal Q-values (see Chapter 2 for more details). Note that, in contrast to V-values which do not directly specify a behavioral policy, given Q-values action selection is trivial: at each state S, the action a with the highest $Q(S,a)$ value is the best action. State values are also easy to derive from Q-values: the value of state S is simply its policy-weighted (for $V^{\pi}(S)$; or maximal, for $V^*(S)$) Q-values.

The main advantage of Q-learning over V-learning is the ease of action selection. Recall that action selection using state values required a one-step lookahead process of estimating the expected successor state when taking each action. This means that although model-free methods for learning state values exist, it is impossibly to explicitly derive a policy from these in the fully model-free case (Actor/Critic algorithms are one way to learn a model-free policy in this case – see below). However, there are several model-free temporal-difference methods for Q-learning, from which a policy can be derived without necessitating a model of the environment. An example is the SARSA algorithm, for which the prediction error and the update rule are:

$$\delta(t) = U_r(t) + \gamma \hat{Q}(S',a') - \hat{Q}(S,a) \tag{1.9}$$

$$\hat{Q}(S,a) \longleftarrow \hat{Q}(S,a) + \eta \delta(t) \tag{1.10}$$

The algorithm is so called because it uses the current State (S), Action (a) and Reward (U_r), and the subsequent State (S') and Action (a') in each update step. Such a learning rule is called "on-policy" as it is based on the consistency between successive Q-values of the actions *actually* chosen. This is in contrast to "off-policy" variants (such as the originally proposed "Q-learning" algorithm; Watkins, 1989), which update the value of the current state-action pair *as if* the optimal action will be taken at the next state, even if this is not

the case. The prediction error in this case is

$$\delta(t) = U_r(t) + \max_{a' \in \mathcal{A}} \{\gamma \hat{Q}(S', a')\} - \hat{Q}(S, a) \tag{1.11}$$

(with the same update rule as in SARSA), and the Q-values converge to the optimal $Q^*(S,a)$ values regardless of the policy used in the learning process.

Another related model-free temporal difference method, which assigns numerical values to both states and state-action pairs, is *advantage learning* (Baird, 1993). The advantage $A(S,a)$ of action a at state S

$$A(S,a) = Q(S,a) - \max_{a \in \mathcal{A}} \{Q(S,a)\} = Q(S,a) - V^*(S), \tag{1.12}$$

essentially quantifies how much better (or worse) this action is, compared to the best action in this state. Advantages can be learned directly (without estimation of the Q-values) through incremental stochastic iterations on a related Bellman equation. After converging, the optimal actions at each state have an advantage of zero, and all other actions have negative advantages. Though perhaps superior to Q-learning in terms of convergence rate and in numerical stability, this method is, in many ways, similar to Q-learning, and can be learned either "on-policy" or "off-policy".

The Actor/Critic Model

Another method of action selection in the model-free case is the *Actor/Critic* framework (first suggested by Barto et al., 1983), which is, to date, the framework most widely used as a model of neural reinforcement learning and action selection in the basal ganglia (Barto et al., 1983; Barto, 1995; Houk et al., 1995; Suri & Schultz, 1999; Suri, 2002; O'Doherty et al., 2004; see Joel et al., 2002, for a review). In this architecture (depicted in Figure 1.5), a *Critic* module learns state values using temporal difference prediction errors, while an *Actor* module represents and directly learns a policy π (rather than action values) using this same prediction error as a training signal. The Critic has been suggested as a model of Pavlovian conditioning, and the combined Actor and Critic as a model of instrumental conditioning (Barto et al., 1983; Sutton & Barto, 1990, but see Dayan & Balleine, 2002, for caveats of this tight link between Pavlovian and instrumental processes).

Note that the Actor does not have access to either $U_r(t)$, the immediate outcomes of the actions it chooses, or the values of the successor states $V(S')$. Instead, the Actor/Critic framework solves the credit assignment problem by using the temporal difference error computed by the Critic (equation 1.5) as an immediate proxy for the outcome of the chosen action (Barto et al., 1983). The temporal difference error does not provide the Actor with information regarding the value of an action. Rather, it attests to whether the chosen action has led to an improvement in the situation ($\delta(t) > 0$, hence $r(t) + \hat{V}(S_{t+1}) > \hat{V}(S_t)$, and the current prediction of cumulative reward is better than that in the previous state), or not ($\delta(t) < 0$). The policy is then adjusted

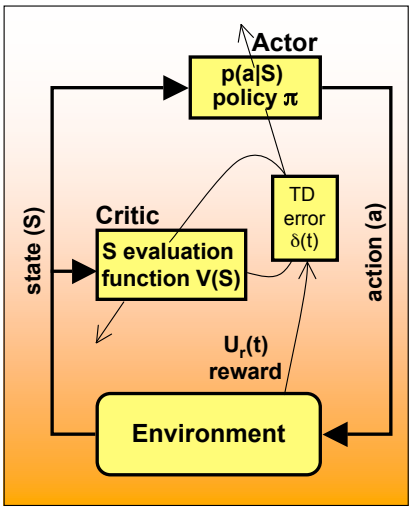

Figure 1.5: Actor/Critic architecture: The environment provides a state S and a reinforcement signal $U_r(t)$ to the Critic, who uses these to compute a temporal difference prediction error (equation 1.5). This prediction error is used to train the state value predictions $V(S)$ in the Critic (equation 1.6), as well as the policy $\pi(a|S)$ in the Actor. Note that the Actor does not receive information regarding the actual outcomes of its actions. Rather, the prediction error signal is a proxy to these, telling the Actor whether the outcomes are better or worse than expected.

according to

$$\pi(a|S) \longleftarrow \pi(a|S) + \eta_\pi \delta(t) \tag{1.13}$$

(where η_π is the Actor's learning rate), such that an action that led to an improvement will be chosen more often in the future, and actions that led to a worse situation will be performed more rarely (Barto et al., 1983; Barto, 1995; Montague et al., 1996). This learning rule can be interpreted as a gradient-climbing rule that attempts to maximize the (long-term) expected rewards in each state (Dayan & Abbott, 2001), and as such, if it converges (for which there is no guarantee), it is to the optimal greedy policy.

The main difference between this Actor/Critic method and methods such as Q-learning is that while the learned action values in the latter methods represent a real property of the environment (the sum of future rewards expected when taking this action), the Actor's policy involves only an ordering of actions, with the relative worth of different actions lost. This means that whereas two actions that lead to very similar expected outcomes will have similar Q-values (and so will have similar probabilities of being chosen in a soft-max scheme, for instance), Actor/Critic learning will converge to the policy that exclusively chooses the better of the two. Note, also, that although the coupling between prediction learning and action selection in the Actor/Critic is not as direct as in Q-learning methods, policy learning nevertheless builds strongly on value learning in this framework as well.

Exploratory action selection

One important aspect in which the above algorithms differ, is in their treatment of exploration. Exploratory action selection involves choosing an action that does not seem best in the current state, to explore its consequences and find out whether they might actually be better than currently believed. Exploration is important for two reasons: First, premature fixation on one action will prevent the animal from learning

about the true values of alternative actions, possibly preventing it from learning the optimal policy. In fact, the convergence (in the limit) of on-line stochastic learning algorithms depends on sampling each state and action pair infinitely often. Second, even after the values have converged, if the dynamics or outcome contingencies in the environment change such that there is now a better policy, lack of exploration can prevent the animal from detecting this change and adapting its behavior appropriately.

Exploration is of course potentially costly (in terms of forfeited rewards), or even dangerous in real-world situations, and it is usually desirable to balance exploration and exploitation. Unfortunately, optimal solutions to the *exploration/exploitation tradeoff* are only available for a restricted class of Markov decision processes. For the general case several heuristics for exploration have been proposed, the two most commonly used being 'ε-greedy' and 'soft-max' exploration. In the former, the optimal action is chosen most of the time and in a small (ε) percent of the trials a random action is chosen uniformly. Exploration is thus *undirected*, but it ensures that all actions are sampled infinitely often. In contrast, soft-max action selection chooses actions with a probability that is proportional to their (exponentiated) values

$$p(a|S) = \frac{e^{\beta \hat{Q}(S,a)}}{\sum_{a \in \mathcal{A}} e^{\beta \hat{Q}(S,a)}}, \tag{1.14}$$

(with $\beta \geq 0$), so exploratory action selection is directed to those actions that are expected to be better. The name 'soft-max' refers to the soft-maximum properties of this probability distribution over actions: in the two extremes of this family of distributions, when the inverse-temperature parameter β is very large, the maximum-valued action is chosen with probability close to one, while for $\beta = 0$ all actions are chosen uniformly. Thus different β values span the range from complete maximizing to uniform action selection.

The difference between "on-policy" and "off-policy" methods is in the effect of suboptimal action selection on the learned Q-values (or advantages). While "on-policy" methods learn Q^{π}-values which reflect the *actual* action policy, whether this be optimal or not, this is not the case in "off-policy" methods which always learn the optimal Q^{*}-values. "Off-policy" Q-learning is thus especially useful because high rates of exploration (or even random action selection) during the learning process will not prevent the animal from learning optimal values, and hence the optimal policy. Actor/Critic methods, which are also "on-policy" but search in policy space rather than in action value space, do not allow any explicit control of the amount of exploration. This is a major disadvantage of Actor/Critic methods: not only do they frequently suffer from premature fixation on suboptimal policies, leading to slow convergence on the optimal policy, but convergence on a maximizing policy is problematic in terms of dealing with changes in the environment, and in terms of modeling animal behavior (which, for the major part, shows sustained levels of exploration).

1.2.3 Computational processes underlying goal directed versus habitual control

The reinforcement learning methods described above can be seen as lying on a spectrum, on the one end of which are model-based *tree-search* methods that use a model of the environment to roll-out the future

consequences of actions, and on the other end are model-free *caching* methods that summarize experience in values, which are a scalar measure of long-term expectations. Both these methods assess actions or states by computing their values, ie, the long-term outcomes they are expected to bring about, however, they differ in how they trade off flexibility and the statistically efficient use of experience, against computational complexity of this computation (Daw et al., 2005).

Tree-based computation makes efficient use of every piece of information regarding observed state transitions and outcomes by building a directly corresponding model of the environment. Long-run values of actions are computed by chaining together these short-term predictions about the immediate consequences of each action in a sequence, using a so called *forward-model* simulation of the task. Searching for the optimal action in deep or wide trees can be expensive in terms of memory and time, however, that the predictions are constructed on the fly means that they are as accurate as could be given the previous experience. In particular, this allows them to correctly take into account the motivation-dependent utility of outcomes, and to react quickly to changed circumstances, such as when outcomes are revalued. Because the behavioral hallmark of goal-directed behavior is the immediate sensitivity of action selection to changes in circumstance, Daw et al. (2005) have suggested that goal-directed control might rely on such a tree-based method for value estimation.

Caching methods, in contrast, combine experience in a scalar storage which does not preserve the identity of the different events and outcomes from which it is comprised. Such a strategy does well to reduce the complexity of computing the long-term values of actions and to allow the rapid comparison of available actions in order to choose between them. However, this comes at the cost of inflexibility: the values are divorced from the outcomes themselves, and so do not immediately change with revaluation of the outcome. For instance, imagine a rat that has learned to traverse a maze with different bits of food at different locations, and thus knows the long-term values of taking left and right turns at each of the junction points. If it suddenly encounters a newly blocked pathway, or if one of the food bits is found in a new location, the rat can not immediately update all the different action values accordingly. Rather, these will be updated slowly through learning from temporal difference prediction errors encountered during repeated experience with the new situation. Because this inflexibility in face of outcome revaluation and changing contingencies is exactly the defining behavioral characteristic of habitual behavior, Daw et al. (2005) suggest that habitual control may be based on a caching mechanism.

Daw et al. (2005) thus argue that habitual and goal-directed control arise from two different computational mechanisms. Specifically, they suggest that habitual behavior might arise from a caching mechanism in the basal-ganglia, and goal-directed behavior may result from a forward-model tree-search mechanism in the frontal cortex. From a normative standpoint, they argue that these two strategies (and the different approximations that necessarily accompany each of them) will be differentially accurate in different situations, and thus there is a benefit to employing both types of decision-making systems in parallel and using each when it is most appropriate. They further suggest a Bayesian principle of arbitration between the two controllers, according to the uncertainty each system places in its value predictions. In this scheme, the system that esti-

mates the action value with least uncertainty is the one whose estimation is used in action selection, such that each system is deployed in situations in which it should be most accurate. This principle can account well for some of the experimental circumstances in which habitual control or goal-directed behavior are typically seen[8]. For instance, early in training when experience with the task is still limited, tree-search methods are more reliable as they distribute every bit of experience optimally. Later in training, caching methods have gained accuracy, and compute long term values more efficiently, thus suffering from less computational noise than the computation-intensive (and thus error-prone) tree-search methods. This could explain why goal-directed behavior succumbs to habitual control with over-training (see Daw et al., 2005, for details).

Where do Actor/Critic methods, those methods that were previously associated with habitual control in the basal ganglia, fall within this taxonomy? In the Actor/Critic framework, action values are not represented at all. However, the Critic learns cached values of states and the Actor's representation of a policy certainly does not maintain information regarding the specific outcomes of different actions (in fact, it does not even maintain summary information regarding the long-run value of actions), placing the Actor/Critic firmly within the category of caching algorithms. In fact, one can argue that these methods are an extreme version of stimulus-response habitual type of control, as they base action selection on an explicit mapping between states and actions, which is even further removed from the specific outcomes of different actions than are action values. It is unclear how the normative uncertainty-based arbitration scheme suggested by Daw et al. (2005) could be extended to an Actor/Critic caching method for habitual behavior (it is unreasonable to think that although action values are not represented, uncertainty in action values would be). However, heuristics can be used to approximate how the Actor/Critic's uncertainty in its policy should change with experience. In the next section, we review evidence for dissociable neural substrates underlying these modes of computation in habitual and goal-directed instrumental control, as well as substrates for Pavlovian learning. We further dissect the representational roles of different brain structures within each decision-making system.

1.3 Neural substrates underlying conditioned behavior

The last two decades have seen a wealth of investigations into the neural basis of decision-making and conditioned behavior, utilizing diverse methods: electrophysiological recording in behaving rats and monkeys have shed light on what components of the learning process may be encoded by different brain areas and cell types, microdialysis and fast scan voltammetry studies have been used to measure task-related fluctuations in local extracellular concentrations of neuromodulators, and lesions and pharmacological interventions have been employed in order to reveal causal relationships and dependencies between different areas of the brain and different aspects of learning and behavior.

The main neural structures implicated in conditioned behavior are the basal ganglia (in particular the striatum), limbic subcortical structures (the amygdala and the hippocampus), and prefrontal cortical areas. Gen-

[8]The one observation that is not explained by this framework is the tendency of behavior on interval schedules of reinforcement to habitize more rapidly than responding on ratio schedules, to which we will return in Chapter 5.

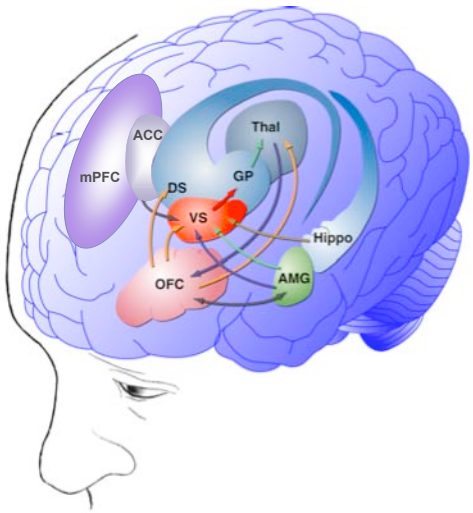

Figure 1.6: A cartoon of the human brain highlighting the major brain areas implicated in conditioned behavior and their connectivity (adapted from Zald & Kim, 1996). For visual clarity, the midbrain and its dopaminergic projections are not depicted. Abbreviations: VS, ventral striatum; DS, dorsal striatum; GP, globus pallidus; Thal, thalamus; AMG, amygdala; Hippo, hippocampus; mPFC, medial prefrontal cortex; ACC, anterior cingulate cortex; OFC, orbitofrontal cortex.

erally speaking, these can be viewed as the decision-making interface between sensory (input) and motor (output) areas of the brain. The broad anatomical organization of this circuitry (Figure 1.6) is as follows: the striatum, which is the input structure of the basal ganglia, receives convergent topographically organized inputs from motor, sensory and prefrontal cortical areas, as well as from the basolateral nuclei of the amygdala, the hippocampus, and the sensory thalamus. The basal ganglia then provide a positive feedback loop to the frontal cortex through cortico-striatal-pallido-thalamo-cortical loops ('cortico-basal-ganglia loops' for short; Parent & Hazrati, 1993; Joel & Weiner, 1994). Information from all sensory modalities also converges in the prefrontal cortex (Sah et al., 2003), whose substructures include the medial prefrontal cortex (mPFC) and the orbitofrontal cortex (OFC). These areas are tightly interconnected, and also share reciprocal connections with subcortical structures. Both subareas project to the amygdala and to the striatum, and the OFC also projects directly to the thalamus. The amygdala projects back to the OFC as well as to the striatum, specifically, to its ventral part, which is functionally and anatomically separable from its more dorsal aspects (see below). To further complicate the picture, neural activity and synaptic plasticity in both the striatum and the prefrontal cortex are modulated by dopamine.

Recent comprehensive reviews (eg. Cardinal et al., 2002; Balleine, 2005) portray an intricate but compelling picture of the roles of these different brain areas in conditioned behavior. In this section, I will very briefly describe this picture, starting from the converging evidence for the involvement of dopamine in signaling temporal difference prediction errors, and continuing with an account of the contributions of different neural structures to Pavlovian, goal-directed, and habitual conditioned behaviors, stressing the specific contributions of subparts of the striatum, amygdala and prefrontal cortex to prediction learning and action selection. Section 1.4 will then summarize the emerging picture of the computational function and neural circuitry underlying different types of conditioned behavior. Of course, books can, and have been written regarding the involvement of each brain area in conditioning. For the sake of brevity, and as this chapter is only meant to provide a backdrop for the models and functions to be discussed in the rest of the thesis, the account below will concentrate on functional and computational roles, suppressing much relevant anatomical and physiological detail.

1.3.1 The neuromodulator dopamine and temporal difference learning

The neuromodulator dopamine plays a major role in learning and action selection, and has been implicated in a variety of disorders, including Parkinson's disease, schizophrenia, obsessive-compulsive disorder, and drug addiction. The majority of dopaminergic neurons are found in the midbrain, in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA), from which they project to virtually the whole brain. The areas receiving the largest projections are the striatum and the prefrontal cortex (Parent & Hazrati, 1993; Schultz, 1998), with the SNc projecting to the dorsal parts of the striatum, and the VTA projecting to the ventral striatum and the prefrontal cortex (Heimer & Wilson, 1975). An impressively large body of data regarding the phasic activity of these cells in Pavlovian and instrumental appetitive conditioning tasks (eg. Ljungberg et al., 1992; Schultz, 1998; Montague et al., 2004) has been taken to suggest that the phasic activity of dopaminergic neurons represents temporal difference errors in the predictions of future reward[9] (henceforth the *prediction error theory of dopamine*; Montague et al., 1996; Schultz et al., 1997).

Some of the most compelling evidence comes from studies investigating the activation of these neurons when a monkey is presented with arbitrary visual stimuli that predict the proximate availability of juice rewards. The basic finding (Schultz et al., 1997) is that whereas dopamine neurons show a strong phasic response to unexpected rewards (such as rewards given in early trials of conditioning, or unsignalled rewards; Figure 1.4d), this response disappears when the reward is predicted and instead the neurons respond to the predictor (Figure 1.4e). If a predicted reward is unexpectedly omitted, the neurons show a precisely timed pause in firing exactly at the time of the previously expected reward (Figure 1.4f; Hollerman & Schultz, 1998; Tobler et al., 2003). This shift in activity from the time of reward to the time of the predictor (Takikawa et al., 2004) resembles the shift of behavioral responses from the time of the US to that of the CS in Pavlovian conditioning experiments (Schultz et al., 1997; Hollerman & Schultz, 1998). These basic characteristics of dopaminergic phasic responsding have been replicated in many variants (Hollerman & Schultz, 1998; Romo & Schultz, 1990; Ljungberg et al., 1992; Schultz et al., 1992, 1993; Schultz, 1998; Tobler et al., 2003; Takikawa et al., 2004; Bayer & Glimcher, 2005) and are exactly as expected for a temporal-difference based prediction error. Further lines of evidence are that the phasic dopaminergic response at the time of a CS predicting reward is proportional to the magnitude and/or probability of the predicted reward (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005), that dopaminergic activity in sophisticated conditioning tasks such as appetitive conditioned inhibition is in line with the predictions of temporal difference learning theory (Tobler et al., 2003), and that firing patterns in tasks involving probabilistic rewards are in accord with a constantly back-propagating error signal (Niv et al., 2005).

The prediction error theory of dopamine is a *computationally precise* theory of the *generation* of phasic dopaminergic firing patterns. It suggests that dopaminergic neurons combine their diverse afferents (which include the mPFC, the nucleus accumbens shell, the ventral pallidum, the central nucleus of the amygdala,

[9]Interestingly, dopaminergic neurons do not seem to be involved in the signaling or prediction of aversive stimuli (Mirenowicz & Schultz, 1996; Tobler et al., 2003; Ungless et al., 2004), in which the neuromodulator serotonin has been implicated instead (Daw et al., 2002).

the lateral hypothalamus, the habenula, the cholinergic pedunculopontine nucleus, the serotoninergic raphe and the noradrenergic locus coeruleus; see Figure 1.7a; Christoph et al., 1986; Floresco et al., 2003; Geisler & Zahm, 2005; Matsumoto & Hikosaka, 2007; Kobayashi & Okada, 2007) to compute a reward prediction error. Moreover, it suggests that dopamine provides target areas with a neural signal that is theoretically appropriate for controlling learning of both predictions and reward-optimizing actions. Following the analogy between the dopamine signal and the temporal difference prediction error signal in Actor/Critic models (Joel et al., 2002), it has been suggested that the signal reported by VTA neurons to ventral striatal and frontal target areas, is used to train predictions (as in the Critic; Barto, 1995; Waelti et al., 2001), while a similar signal reported by the SNc to dorsal striatal target areas, is used to learn an action-selection policy (as in the Actor; Miller & Wickens, 1991; Wickens & Kötter, 1995; Houk et al., 1995; Joel & Weiner, 1999; O'Doherty et al., 2004). Current views on how these areas contribute to the processes of learning and action selection in Pavlovian and instrumental conditioning will be detailed shortly.

Before we continue, it should be mentioned that there are alternative theories regarding the role of dopamine in conditioned behavior (for a recent debate-style review, see Berridge, 2007). These include Wise's 'anhedonia hypothesis' (eg. Wise et al., 1978), Redgrave and colleagues' 'incentive salience' (eg. Redgrave et al., 1999; Horvitz, 2000; Redgrave & Gurney, 2006), Berridge and Robbinson's 'wanting' versus 'liking' (eg. Berridge & Robinson, 1998; Berridge, 2007), and ideas about dopamine signaling uncertainty (Fiorillo et al., 2003). A discussion of the merits and pitfalls of these theories is beyond the scope of this chapter. Nevertheless, in as far as these theories are indeed fundamentally different from the prediction error theory (which is not always clear), it is my belief that, to date, no alternative has mustered as convincing and multidirectional experimental support as the prediction error theory of dopamine.

This having been said, the prediction error theory concentrates on only one aspect of dopaminergic activity and influence: the effect of *phasic* dopaminergic signaling on learning and plasticity. However, dopamine neurons operate in both a phasic and a tonic mode (Grace, 1991; Schultz, 2002; Weiner & Joel, 2002; Floresco et al., 2003; Bergstrom & Garris, 2003; Goto & Grace, 2005), and affect not only synaptic plasticity, but also membrane potentials and neural excitability[10] (Nicola et al., 2000; Schultz, 2002). The base level of dopamine in the striatum (and more specifically, in its ventral part – the nucleus accumbens) has been linked to invigorating preparatory (approach) Pavlovian responses (Ikemoto & Panksepp, 1999), and a multitude of studies have shown that 6-hydroxydopamine lesions of the nucleus accumbens (which cause the death of dopaminergic cells projecting to it) profoundly reduce the rate of instrumental responding (for a review see Salamone & Correa, 2002). To date, there exist no computational theory of the role of tonic levels of dopamine in conditioned behavior. In Chapter 4 of this thesis, I will attempt to remedy exactly this.

By nature, the influence of dopamine on learning and action selection is a modulatory one. What neural structures does it modulate and how do these bring about conditioned behavior? Balleine (2005), describes three functional cortico-basal-ganglia loops (sensorimotor, associative and limbic) in the rat brain, each fulfilling a specific role in conditioned behavior. Within these, the respective subparts of the striatum (the

[10]These influences on neural excitability have been suggested to mediate a role for phasic dopamine signals to the prefrontal cortex in attentional switches (Cohen et al., 2002).

dorsolateral striatum (DLS; or its primate homologue, the putamen), the dorsomedial striatum (DMS; or its primate homologue, the caudate), and the nucleus accumbens (NAC, also called the ventral striatum), further subdivided into core and shell) are the main structures implicated in habitual, goal-directed and Pavlovian control. Cortical and limbic areas projecting to these are thought to represent the information required by each controller, and dopaminergic projections arising from the VTA ('mesolimbic dopamine', which targets the ventral striatum) and the SNc ('nigrostriatal' dopamine, which targets the dorsal striatum) influence both learning and action selection. In the following, each of these three loops is described, starting from the limbic loop associated with Pavlovian behavior, continuing to the associative loop thought to underly goal-directed behavior, and finishing with the habitual sensorimotor loop.

1.3.2 Pavlovian behavior: The 'limbic loop'

Although reflexive Pavlovian behavior is seemingly simpler than its flexible instrumental counterpart, elucidating the functional, computational and neural control of Pavlovian *responding* (rather than Pavlovian prediction learning) has proved difficult. Empirically, because the nature of the Pavlovian response is not under the control of the experimenter, it is frequently harder to measure (eg, salivation and orienting responses are harder to measure than leverpressing). Computationally, because Pavlovian responding is not necessarily normative (ie, situations can be contrived in which Pavlovian responding is strictly suboptimal in terms of obtaining rewards), the framework of RL which has been used to explain Pavlovian predictions, stops short of explaining their associated responses. Nevertheless, considerable knowledge has amassed regarding the neural basis of Pavlovian conditioning.

At the level of the striatum, the NAC has been strongly implicated in attributing value to Pavlovian CSs and in mediating the ability of the affective value of anticipated outcomes to affect instrumental performance, as in PIT and conditioned reinforcement[11] (Cardinal et al., 2002). Within the NAC, the core is more similar to the dorsal striatum in its afferent and efferent connections to general striatal targets, while the shell receives additional inputs from the hippocampus, and also projects to the lateral hypothalamus and to dopaminergic neurons in the VTA (through which it can indirectly affect processing in the core (Cardinal et al., 2002); Figure 1.7a). Lesion studies suggest that the NAC core (but not the shell) is necessary for the general form of PIT (Hall et al., 2001) and for Pavlovian preparatory behavior (Cardinal et al., 2002). In contrast, the NAC shell (but not the core) is implicated in the outcome-specific form of PIT (Corbit et al., 2001), as well as in US-specific consummatory responses (Cardinal et al., 2002). Accumbal dopamine arising from the VTA is critical for many of these functions, especially those that are of an outcome-general nature and are mediated by the NAC core (Ikemoto & Panksepp, 1999; Dickinson et al., 2000; Sellings & Clarke, 2003).

Another structure implicated in Pavlovian learning is the amygdala. The amygdala is a limbic structure in the mediotemporal lobe comprising of a group of nuclei that is commonly divided into two subgroups (Killcross & Blundell, 2002; Swanson, 2003): the more cortex-like basolateral and basomedial nuclei,

[11] In conditioned reinforcement a Pavlovian CS is used as a reinforcer for instrumental responding. This paradigm is related to PIT in that learned Pavlovian values are used to modify instrumental responding.

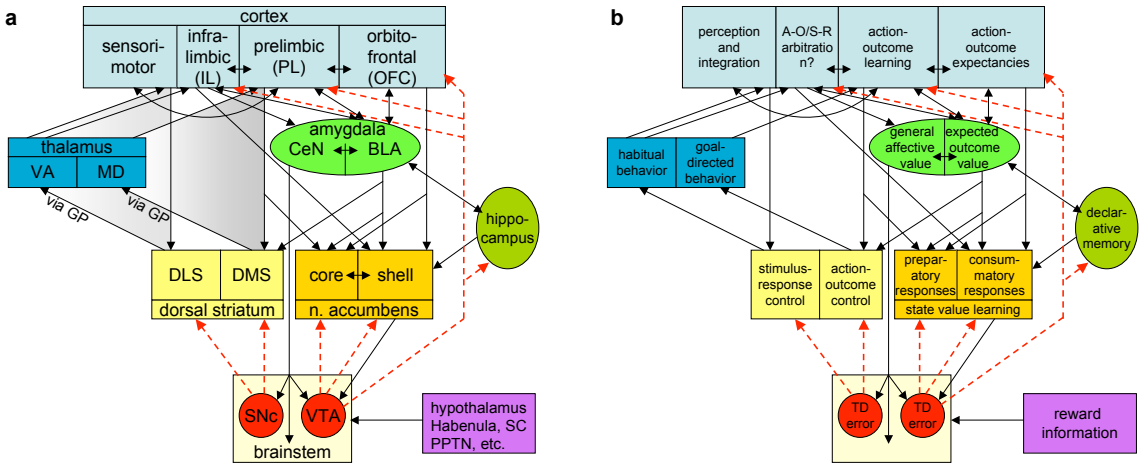

Figure 1.7: An overall picture of the neural control of conditioned behavior. **a.** A (partial) diagram of the neural structures involved in conditioned behavior and their relevant afferent and efferent pathways. In red dashed lines are dopamine projections, and the 'sensorimotor' and 'associative' cortico-basal-ganglia loops are gradient-shaded for clarity. Due to the high complexity of the neural connectivity patterns, not all projections are shown, the ACC is not depicted, and the traditional basal-ganglia 'limbic loop' is not shown. Abbreviations: MD mediodorsal nucleus of the thalamus; VA, ventral anterior nucleus of the thalamus; GP, globus pallidus; DLS, dorsolateral striatum; DMS, dorsomedial striatum; CeN, central nuclei of the amygdala; BLA, basolateral nuclei of the amygdala; SNc, substantia nigra pars compacta; VTA, ventral tegmental area; PPTN, pedunculopontine nucleus; SC, superior colliculus. **b.** The suggested functional contributions of each of the areas in (a) to conditioned behavior, overlaid on the same circuitry diagram. The anatomical connectivity patterns are in line with the implied functional dependencies. Abbreviations: A-O, action-outcome associations; S-R, stimulus-response associations; TD, temporal difference.

together forming the basolateral amygdala (BLA), and the more striatum-like medial and central nuclei, together forming the central amygdala (CeN). Traditionally associated with emotional learning, particularly in aversive situations (Sah et al., 2003), the amygdala has recently been ascribed a more general role in conditioning, which encompasses appetitive situations as well (Holland & Gallagher, 1999; Balleine & Killcross, 2006). In Pavlovian conditioning, Balleine and Killcross (2006) argue that the BLA mediates the associations that support consummatory responding, while the CeN is involved in those associations that bring about preparatory responses (see Konorski's model, Figure 1.2). Importantly, they argue against the classical attribution of fear acquisition to the BLA and expression of fear responses to the CeN, maintaining that the consummatory/preparatory distinction holds for both appetitive and aversive Pavlovian conditioning.

More specifically, the BLA, together with its connections to the NAC and OFC, has been suggested to be involved in the learning processes by which initially neutral stimuli become associated with the motivational value (sometimes called 'incentive value') of the outcomes they predict (eg. Holland & Gallagher, 1999; Gewirtz & Davis, 2000; Killcross & Blundell, 2002; Cardinal et al., 2002). This role in endowing stimuli with an affective value associated with the specific sensory properties of their predicted outcomes (rather than a general motivational value), corresponds to the computational role of learning one-step state-dependent outcome predictions within a world model (although bar Dayan and Balleine (2002), current computational theories do not distinguish between the specific sensory and general motivational components of one-step predictions). Empirically, lesions of the BLA (but not the CeN) impair the sensitivity

of Pavlovian responding to outcome-specific devaluation and disrupt outcome-specific PIT (Blundell et al., 2001, 2003; Corbit & Balleine, 2005). Furthermore, the ability of a conditioned stimulus to support new learning in second-order Pavlovian conditioning, or as a conditioned reinforcer in instrumental conditioning, depends on the integrity of the BLA (the latter specifically on BLA afferents to the NAC; Everitt & Robbins, 1992; Hatfield et al., 1996). Anatomically, the BLA is well situated for this role of integrating stimulus representations with outcome-specific affective value, as it receives diverse sensory cortical and thalamic inputs, has strong reciprocal connections with prefrontal cortical areas (specifically the OFC, the gustatory insular cortex, and both the infralimbic and prelimbic subdivisions of the mPFC) and the hippocampus and perirhinal cortex, and receives inputs from the hypothalamus (Krettek & Price, 1977; Kita & Kitai, 1990; Sah et al., 2003). Apart from the OFC and mPFC, its major efferent pathways target the CeN, the NAC (with topographically organized inputs to both core and shell; Voorn et al., 2004), the DMS (but not DLS) and the mediodorsal thalamus (Krettek & Price, 1977; Kita & Kitai, 1990; Sah et al., 2003; McDonald, 2003, see Figure 1.7a for a partial connectivity diagram), also consonant with a role in learning a world-model which is relevant not only to Pavlovian predictions, but also to goal-directed behavior (see Section 1.3.3).

The CeN, which has strong reciprocal connections to the BLA, also receives direct sensory cortical and thalamic inputs, and inputs from the perirhinal cortex and the hypothalamus. However, different from the BLA, it receives afferents from brainstem areas, its only prefrontal input is from the infralimbic cortex (Killcross & Blundell, 2002), and it influences hypothalamic, midbrain and brainstem areas including those involved in projecting neuromodulators throughout the brain: the dopmainergic SNc and VTA, the noradrenergic locus coeruleus, the serotonergic raphe and the basal forebrain cholinergic nuclei (Cardinal et al., 2002; Sah et al., 2003). In contrast to the BLA, the CeN is thought to be important for endowing stimuli with an *outcome-general* motivational value (appetitive/aversive). For example, the CeN is key in conditioned approach and withdrawal, and is involved in general contextual fear conditioning, whereas the BLA is necessary for outcome-specific fear-induced freezing responses (Cardinal et al., 2002). In a recent study, Corbit and Balleine (2005) elegantly demonstrated this distinction between outcome-specific involvement of the BLA and outcome-general role of the CeN, by showing that the former is necessary for outcome-specific PIT while the latter is necessary for general PIT (see also Blundell et al., 2001; Holland & Gallagher, 2003). The involvement of the CeN in producing outcome-general conditioned orienting responses to stimuli which have acquired conditioned (predictive) value, has also suggested an attentional function (Holland & Gallagher, 1999; Gewirtz & Davis, 2000). In this, the CeN has been implicated in the upregulation of the associability of stimuli that are unreliable predictors of outcomes through its projections to cholinergic brainstem nuclei (Pearce & Hall, 1980; Holland, 1997; Holland & Gallagher, 1999, 2006; Lee et al., 2006).

1.3.3 Goal directed behavior: The ‘associative loop’

Relatively little is known about the control structure underlying goal-directed behavior. The flexible, forward-model based action selection system is, as of yet, under-constrained both computationally and behaviorally. However, recent work has identified several neural components of the circuit which beings about goal-directed behavior. These studies assess the contribution of different brain areas to the hallmark sensitivity of

goal-directed behavior to changes in the forward model. Typically, the one-step model of the environment is changed post-training (by re-valuing an outcome, or by changing the instrumental relationships between actions and outcomes or the Pavlovian relationships between stimuli and outcomes), and the impact of this is observed in healthy rats and in rats whose normal neural function has been compromised (pharmacologically or through lesions).

Convergent evidence from such studies from the lab of Bernard Balleine has implicated the rat posterior DMS in the acquisition and expression of so-called 'action-outcome' (forward prediction) associations. Rats whose posterior DMS is lesioned either before or after instrumental training, do not show outcome-specific sensitivity in a devaluation test or in a contingency degradation test (Yin et al., 2005). Moreover, Yin et al. (2005) demonstrated impaired acquisition of action-outcome associations as a result of infusion of an NMDA antagonist into the posterior DMS before (but not after) training: rats whose posterior DMS was infused with antagonist did not show outcome-specific sensitivity to devaluation, while infusion into the DLS had no effect[12]. In primates, electrophysiological recordings in the caudate (the homologue of the rat DMS) have shown modulation of single cell activity by outcome expectancy (Kawagoe et al., 1998).

The rat DMS receives inputs from premotor and prelimbic association cortices (PL, part of the rat medial prefrontal cortex, and the homologue of the dorsolateral prefrontal cortex in primates), as well as from the basolateral amygdala, and projects to the mediodorsal thalamus (Yin et al., 2005, see also Figure 1.7a). (Balleine, 2005) suggests that this whole so-called 'associative cortico-basal-ganglia loop' is involved in the control of goal-directed instrumental behavior based on action-outcome (forward model) relationships, because, like lesions of the DMS, lesions to other components of the loop (specifically, the PL or the mediodorsal thalamus), result in a deficit in the sensitivity of behavior to the utility or contingency of its outcomes. Balleine and Dickinson (1998) were the first to show that lesions of the PL impaired rats' sensitivity to outcome-specific contingency degradation[13]. In an elegant within-animal comparison of goal-directed and habitual responding, Killcross and Coutureau (2003) later showed that PL lesions cause behavior that would otherwise be goal-directed (ie, sensitive to outcome devaluation) to come under habitual control. Interestingly, different from lesions of the DMS, post-training PL lesions do not affect sensitivity to outcome devaluation (Yin et al., 2005), suggesting that the PL is involved in the *acquisition* of action-outcome associations (Cardinal et al., 2002), but is not required for these to control behavior.

Another area that has been implicated in learning of a world model is the BLA (see Section 1.3.2). The integrity of the BLA is important mostly at the beginning of training (Cardinal et al., 2002), and BLA lesions also affect the sensitivity of instrumental responding to outcome devaluation and contingency degradation (Balleine et al., 2003; Corbit & Balleine, 2005). This could be due to the effects of Pavlovian predictions on instrumental action selection (for example, as in the Actor/Critic framework), or it might indicate a

[12]This study provides the first direct evidence for the localization of forward-prediction action-outcome learning in a specialized brain area, and for the dependence of such learning on NMDA receptor dependent plasticity.

[13]Lesions to a different frontal cortical area, the gustatory insular cortex, produced similar results. However, only rats with PL lesions showed this impairment in a *rewarded* test, suggesting that lesions of the PL impair the immediate sensitivity of the rats' behavior to specific outcome contingencies (that is, absent a PL, the behavior comes under habitual control and can only incorporate changes in contingencies through slow learning), while the gustatory insular cortex is perhaps only the site of *memory* for the action-outcome contingency ('incentive memory'; Balleine & Dickinson, 2000).

more general role of the BLA in learning the one-step outcome predictions contingent on both states and state-action pairs. Balleine et al. (2003) showed that the BLA is essential not only for stimulus-outcome learning, but also for these relationships to control instrumental responding, as lesions to the BLA affect outcome-specific responding even when the outcomes are available at test.

Finally, the orbitofrontal cortex (OFC) has also been tightly linked to learning of outcome expectations. The hallmark deficit associated with OFC damage is in reversal learning: while animals with OFC lesions will acquire a task quite readily, they show abnormal perseveration when task contingencies are reversed (eg. Schoenbaum et al., 2002; Fellows & Farah, 2003; Kim & Ragozzino, 2005), and these deficits persist even when returned to the original task (Schoenbaum et al., 2003). Single cell recordings and neural imaging studies have shown that activity in the OFC is related to the value of anticipated outcomes, and tracks changes in these values (Schoenbaum et al., 1998; Tremblay & Schultz, 1999, 2000; O'Doherty et al., 2002; Roesch & Olson, 2004; Valentin et al., 2007). Accordingly, damage to the OFC either before or after training impairs sensitivity to outcome devaluation (Gallagher et al., 1999). These results suggest that the OFC has a role in representing expected outcomes, or more specifically, the immediate outcomes expected when performing different responses (so called 'outcome expectancies'; Schoenbaum & Roesch, 2005). This role is similar to the one ascribed to the BLA, which is indeed extensively reciprocally connected to the OFC. However, different from the OFC, post-conditioning lesions of the BLA do not abolish devaluation effects, suggesting that while the BLA is necessary for learning stimulus-outcome expectations, the OFC is sufficient in order to use these to control responding post-conditioning (Holland & Gallagher, 2004).

1.3.4 Habitual behavior: The 'sensorimotor loop'

A similar line of studies examining the effects of outcome revaluation on rats with selective lesions of sub-parts of the striatum, has shown that habitual behavior, previously associated with the striatum in general, is more specifically linked to the DLS (Balleine, 2005). This area receives afferents from sensorimotor cortices, and projects to the ventroanterior thalamus, in what is considered the 'sensorimotor loop' (Figure 1.7a). Yin et al. (2004) showed that lesions of the rat DLS cause behavior to be sensitive to outcome devaluation, even in conditions that promote outcome-insensitive habitual responding in non-lesioned rats.

Also implicated in habitual control is the infralimbic cortex (IL; which, like the PL, is a part of the rat mPFC). In contrast to lesions of the adjacent PL area, lesioning the IL prevents the habitizing of behavior even after extensive training (that is, extensively trained behavior remains sensitive to devaluation), and inactivation of the IL after extensive training reinstates goal-directed behavior, as detected by the outcome-sensitivity of responding (Killcross & Coutureau, 2003; Coutureau & Killcross, 2003). One important implication of this, is that if habitual control is disrupted, the goal-directed system can take control of responding. This indicates that the transition from goal-directed to habitual responding that accompanies overtraining is not the result of gradual deterioration of goal-directed control and a loss of the learned forward-model of the task, but rather results perhaps from inhibition of goal-directed control mechanisms. Together with the results described above, showing that the disruption of goal-directed control results in habitual behavior,

these studies suggest that both the habitual and the goal-directed control mechanisms are available at any stage of training. Indeed, because the IL is not traditionally part of the cortico-basal-ganglia sensorimotor loop, Killcross and Coutureau (2003) suggested that it might play a role in inhibiting habitual behavior early in training, to allow goal-directed control. Anatomically, the IL cortex projects to limbic regions including the amygdala and the NAC shell, and has reciprocal connectivity with the PL (see Figure 1.7a), placing it in a suitable position to influence the arbitration between habitual and goal-directed control.

1.4 An integrative view of conditioned behavior

The overall picture that emerges is intricate but compelling: two types of instrumental behavior result from two computational strategies for learning and action selection, which are realized by dissociable neural pathways. On the one hand is goal-directed behavior, which is characterized by its sensitivity to the value of the specific outcome contingent upon each possible action. This type of behavior involves a forward-model type of controller, which seems to be realized by the 'associative' basal ganglia loop, from the prelimbic part of the medial prefrontal cortex, to the dorsomedial striatum, through the mediodorsal thalamic nucleus, and back to the prelimbic cortex. The PL cortex is likely involved in learning of forward-model action-outcome relationships, while action selection based on these is performed by the DMS (Figure 1.7b; Yin et al., 2005, 2005; Balleine, 2005). On the other hand is habitual behavior, defined by its lack of immediate sensitivity to manipulations of outcome value or contingency (of course, through relearning, habitual behavior can gradually respond to these manipulations). This response strategy involves a cache-based controller, as in the Actor/Critic model, and is dependent on the integrity of the 'sensorimotor' basal ganglia loop (from sensorimotor cortical areas, to the dorsolateral striatum, through the ventral anterior nucleus of the thalamus and back to the originating cortical areas) and its dopaminergic afferents (Figure 1.7b; Packard & White, 1991; Yin et al., 2004; Balleine, 2005; Faure et al., 2005; Nelson & Killcross, 2006).

Importantly, studies in which damage to the circuitry underlying one type of behavior reverts instrumental control to the alternative type (eg. Balleine & Dickinson, 1998; Killcross & Coutureau, 2003; Coutureau & Killcross, 2003; Balleine, 2005), demonstrate that both action control mechanisms can, in principle, control behavior at any point in time. In fact, initial learning of a task can apparently use either system, and thus is robust to lesions of some[14] of the neural structures involved in one system but not the other (eg. Holland & Gallagher, 1999; Dickinson et al., 2000; Hall et al., 2001; Blundell et al., 2003; Balleine et al., 2003; Killcross & Coutureau, 2003). Indeed, Daw et al. (2005) suggested that, in the intact animal, behavioral control is determined on an action-by-action basis, with the controller that produces the most reliable prediction the value of each action being the one influencing that action. Studies of instrumental chains of actions (Balleine et al., 1995; Corbit & Balleine, 2003) show exactly this: at the same time that one action in a chain is susceptible to outcome devaluation, another action can be insensitive to this manipulation.

[14] An exception to this is lesions of the CeN, which do impair initial performance vigor (Corbit et al., 2001; Corbit & Balleine, 2005), perhaps through the the influence of the CeN on dopaminergic afferents to the NAC. In Chapter 4, I will propose a computational account for exactly this dopamine-dependent function of the CeN and NAC core, in regulating response vigor.

The psychological and computational distinction between Pavlovian and instrumental conditioning is also somewhat mirrored by the underlying neural structures, with Pavlovian prediction learning, as in the Critic, relying on the NAC and the amygdala, while action selection, as in the Actor, is subserved by the dorsal striatum (DMS and DLS). Interestingly, the distinction between outcome-general and outcome-specific aspects of behavior is also useful in interpreting the involvement of different neural structures in Pavlovian behavior, and in the relationships between Pavlovian and instrumental behavior. In the Pavlovian conditioning literature, this parallels the distinction between consummatory (outcome-specific) and preparatory (outcome-general) responding, with the former based on CS-US associations (Colwill & Motzkin, 1994) and the latter on associations between the CS and the relevant motivational system ("CS-affect" associations).

This suggests *two Pavlovian prediction learning systems*: the NAC shell and the BLA for outcome-sensitive (CS-US, forward-model based) predictions, and the NAC core and the CeN for outcome-insensitive (CS-affective value, cache based) predictions (see Figure 1.7b). Assuming that Pavlovian responding is a direct consequence of Pavlovian predictions (as in learned state values; Dayan et al., 2006), this distinction is completely in accord with the dependence of (outcome-sensitive) consummatory Pavlovian responses on the BLA and the NAC shell, and of the dependence of (general, affect-based) preparatory Pavlovian responding on the CeN and NAC core. For example, amphetamine-induced (outcome-general) locomotion is dependent on NAC core, while outcome-specific conditioned place preference and potentiated feeding depend on NAC shell and on the BLA (Sellings & Clarke, 2003; Holland & Gallagher, 2003). This is also in line with the effects of Pavlovian predictions on instrumental responding, as general PIT is abolished by lesions of the NAC core or the CeN, but spared by lesions of the NAC core or the BLA (Hall et al., 2001; Holland & Gallagher, 2003), and outcome-specific PIT depends on the integrity of the BLA and the NAC shell, and is independent of the CeN and NAC core (Corbit et al., 2001; Killcross & Blundell, 2002).

I would thus argue that, in addition to two different action controllers (Daw et al., 2005), the evidence points to two separate evaluators, and, in general, two parallel Actor/Critic mechanisms can be identified in the brain[15]. One of these is the traditional Actor/Critic, with model-free temporal difference learning in the Critic[16] (CeN and NAC core), and policy learning in the Actor (DLS). In agreement with the computational model, and with the identification of phasic dopaminergic signals as carrying a temporal difference reward prediction error, the normal function of this system is dependent on intact dopaminergic afferents. The second system involves a forward model evaluator that computes state values based on a lookahead process in a learned model of the task (putatively involving working memory in the prefrontal cortex, and a world model learned by the OFC, BLA and NAC shell)[17], accompanied by a goal-directed action evaluation and selection system (in the DMS). This goal-directed system might be realized through a variety of computational

[15]This proposed division into two actors and two critics builds on a model recently suggested by Balleine (2005), but differs from it in the attribution of roles to amygdala and NAC substructures.

[16]Recent results from recordings of dopamine neuron activity in monkeys performing a choice task suggest that the prediction error conveyed by midbrain dopamine is, in fact, a SARSA-type of prediction error, which takes into account not only the current and previous states, but also the actions chosen at these states (equation (1.9); Morris et al., 2006). An Actor/Critic model using this type of error has yet to be proposed, however, new evidence from VTA recordings in rats questions the generality of this finding (Roesch et al., Submitted).

[17]This is in line with suggestions that the hippocampus (which projects to the NAC shell, but not core) is part of the goal-directed system (Corbit & Balleine, 2000; Daw et al., 2005).

schemes, and is relatively underconstrained at the present. It seems, however, that despite the dopaminergic projections to the DMS, goal-directed control is dopamine-independent.

1.5 Motivation and the organization of this thesis

How does motivation fit into this picture of outcome-specific goal-directed and outcome-general habitual control? Motivational control is vital to instrumental behavior, because the adaptive advantage of instrumental control is only inasmuch as the achieved outcomes are relevant to the state of the animal — its needs, deprivations and desires. Much of the research reviewed in this chapter has relied on motivational effects to define and dissect different action selection mechanism. However, this research has not explicitly attempted to define and dissect motivational control. In this thesis I will attempt to do exactly this.

In delineating motivational control, the effects of motivation on goal-directed behavior seem quite straightforward and adequately described in the literature. If motivational states define the subjective values of outcomes that are contingent on different actions (a role which we will formalize more precisely in Chapter 3), then because goal-directed control plans actions based on these outcome values, it is directly influenced by motivation. But what about motivational control of habitual responding? Because most of day-to-day responding is habitual, this question is of crucial importance: can our daily habitual behavior be flexibly adjusted to our needs? On the one hand, it seems that the answer to this is 'no' — habitual behavior is, by definition, not flexibly sensitive to the value of its consequent outcomes, and so it is motivation-insensitive (Robbins & Everitt, 1996). This answer makes habitual behavior seems grossly maladaptive. On the other hand, Dickinson and Balleine (2002) postulate that, like Pavlovian US-specific responding, habitual behavior *is directly sensitive* to outcome-specific motivation, in fact, more sensitive than goal-directed responding, as the latter but not the former requires incentive learning for the effects of motivational shifts to be manifest[18] (see also Dickinson et al., 1995). This account is also counterintuitive: if the Pavlovian prediction system knows the current motivation-dependent value of the outcome, why would this information be used by the habitual system, but not by the computationally more complex and behaviorally more advanced goal-directed system? Moreover, if this account is indeed true and habitual behavior is sensitive to motivation in an outcome-specific way, then habitual control *cannot be realized by temporal difference learning and an Actor/Critic framework* as is currently believed, because cached values in the Actor/Critic are inherently insensitive to changes in the value of a specific predicted outcome.

I will argue that the resolution to this conundrum lies in the two different effects of motivation on responding (Bolles, 1967; Bindra, 1974): a 'directing' effect that affects only goal-directed responding, and an 'energizing' effect that can affect habitual behavior as well. The latter is perhaps at the source of this confusion: while a hungry rat navigating a food-containing maze will run faster than a sated rat, current computational

[18] Dickinson and Balleine (2002) assume that Pavlovian responding is sensitive to motivational manipulations without necessitating incentive learning, based on a study by Balleine (1994) that showed that PIT is directly sensitive to motivational shifts. The hypothesis this thesis will lay out regarding the effects of motivation on outcome-general habitual responding, will also explain the effect of motivation on general PIT which was seen in this study.

models of action selection completely lack a notion of *response vigor*. In the absence of this, the effects of motivation on instrumental (or Pavlovian) responding cannot be fully explained.

In this thesis, I will develop a computational, behavioral and neural account of the motivational control of response vigor (or response rates) in habitual behavior. The next chapter will lay out a reinforcement learning model in which animals not only choose which action to perform, but also with what vigor, or at what instantaneous rate, to perform it. The critical tradeoff that will determine optimal response rates is that between the high costs of fast responding, versus the potential loss of rewards if responding is too sluggish. Optimal behavior in the model will maximize the net rate of rewards, ie, the amount of rewards earned minus the costs of responding, per unit time. I will show that a critical determinant of the optimal rate of responding is the *net rate of rewards*, which acts as the 'opportunity cost' of time. In Chapter 3, I will lay out the implications of this model for the effects of motivation on response rate and response selection. Importantly, I will claim that the effects of motivation on response rates (ie, the 'energizing' effects of motivation), should *normatively* be outcome-general, and thus can be manifest in habitual responding as well as in goal-directed behavior. This suggests a solution to the puzzle of habitual sensitivity to motivation: habitual behavior is indeed directly sensitive to the 'energizing' effects of motivation, but, at the same time, is insensitive to outcome-specific 'directing' effects of motivation. Thus habitual responding is not grossly maladaptive, but can still be realized by a computational strategy which relies on caching.

In Chapter 4, I will relate these results to the neural substrates of habitual behavior and the control of response vigor. On psychological, computational and neural grounds, I will argue that the net rate of rewards is reported by tonic levels of dopamine. This hypothesis provides a normative explanation to the pronounced effects of dopaminergic manipulations on response vigor, and completes the computational account of the role of dopamine in conditioned behavior, providing a role for both the phasic and the tonic modes of dopamine transmission. To flesh out the relationship between the new model and previous Actor/Critic accounts of instrumental learning in the brain, Chapter 5 will follow with an online Actor/Critic algorithm implementing learning of optimal response rates. This model is easily mapped onto the same neural substrates which we associated with the habitual Actor and Critic in this chapter, with the additional role of tonic dopamine as the critical net rate of rewards. Finally, Chapter 6 will describe two experiments on the effects of shifts in motivational states on leverpress rates in rats, whose results support the idea that habitual behavior is affected only by general 'energizing' aspects of motivation.

As a final introductory word, although this thesis focuses on instrumental behavior, its results can, for the most part, be easily extended to the domain of Pavlovian conditioning. In fact, I will postulate that, with the possible exception of PIT[19], the motivational sensitivities of US-specific consummatory and general preparatory Pavlovian responses parallel those of goal-directed and habitual instrumental behaviors, respectively, due to the similarity in the computational strategies underlying these forms of outcome-specific and outcome-general control (for instance, see Holland, 1998).

[19]Both outcome-specific and general PIT have been shown to be insensitive to outcome devaluation using conditioned taste aversion (Holland, 2004), suggesting that PIT is perhaps only sensitive to outcome-general aspects of motivational control.

The model presented in this chapter was first published, in shorter format, in: Niv, Y., Daw, N.D. and Dayan, P., *How fast to work: Response vigor, motivation and tonic dopamine*, **NIPS**, 2005.

All models are wrong, some models are useful – George Box

Chapter 2

A computational model of free operant behavior

Abstract: Reinforcement learning models have long promised to unify computational, psychological and neural accounts of appetitively conditioned behavior. However, although the bulk of data on animal conditioning comes from free-operant experiments measuring how fast animals will work for reinforcement, existing reinforcement learning models lack any notion of *vigor* or *response rate*, and so are silent about these tasks. In this chapter, I review the basic characteristics of free-operant behavior, illustrating the effects of reinforcement schedules on rates of responding. I then develop a reinforcement learning model in which both response selection and vigor selection are optimized. The model suggests that subjects choose how vigorously to perform selected actions by optimally balancing the costs and benefits of different speeds of responding. This model accounts normatively for effects of reinforcement schedules on response rates, such as the fact that responding on ratio schedules is faster than that on yoked interval schedules. Finally, the model highlights the importance of the *net rate of rewards*, which acts as the *opportunity cost* of time, quantifying the cost of sloth. The implications of the model for motivational control of habitual behavior, and for the effects of tonic levels of dopamine on response rates, will be discussed in turn in the following two chapters.

2.1 Introduction

Choice behavior has most frequently been studied in instrumental (or 'operant') conditioning paradigms. In the commonly used *free-operant* form of these, animals[1] (typically rats, mice or pigeons) perform an action (pressing a lever, pulling a chain, pecking a key etc.) in order to obtain some coveted reinforcement (such as food for a hungry animal). Skinner, who pioneered these methods, was interested in revealing how

[1] Although instrumental conditioning paradigms have also been tested extensively on humans and the results conform to those seen in animal behavior, we will concentrate our discussion on results from animal experimentation.

animal behavior is controlled (or shaped) by the experimenter-specified schedule of reinforcement (Ferster & Skinner, 1957). Importantly, rather than performing actions at discrete, predefined time-points (as is typically modeled in reinforcement learning; Sutton & Barto, 1998), free-operant responding is *self paced*, and the most common dependent variable in such experiments is the *rate* of responding (Williams, 1994).

2.1.1 Free-Operant schedules

The reinforcement schedules most frequently used are *ratio* schedules and *interval* schedules (Baum, 1993; Williams, 1994; Domjan, 2003). In interval schedules the first response after an unsignalled predetermined interval has elapsed, is rewarded. The interval duration can be fixed (say, 30 seconds; FI30) or randomly drawn from a distribution with a given mean. Traditionally, if this distribution is the memoryless exponential distribution, the schedule is called a random interval (RI) one, else it is a variable interval (VI) schedule. The first interval in an experimental session is timed from the start of the session, and subsequent intervals are timed from the previous reward. In ratio schedules reinforcement is given after a predefined number of actions have been emitted. The required number of responses can be fixed (FR) or drawn randomly from some distribution (VR; or RR if drawn from a Geometric distribution). Schedules are often labeled by their type and the schedule parameter (the mean length of the interval or the mean ratio requirement). For instance, an RI30 schedule is a random interval schedule with the exponential waiting time having a mean of 30 seconds, and an FR5 schedule is a ratio schedule requiring a fixed number of five responses per reward.

Note that ratio and interval schedules imply very different functional relationships (sometimes called 'molar feedback functions'[2]) between response rate and reinforcement rate. In ratio schedules the rate of reinforcement is linearly related to the rate of responding. In interval schedules, by contrast, any responding prior to the termination of the interval is not rewarded, and the rate of reinforcement is limited by the interval duration. This results in a nonlinear saturating relationship between responses and reinforcements.

Numerous experiments have shown that the schedule of reinforcement, the nature or amount of the rewards used, and the motivational state of the animal, profoundly affect the rate of instrumental responding (Williams, 1994). Generally speaking, responding is slower in longer interval schedules or higher ratio schedules (Herrnstein, 1970; Barrett & Stanley, 1980; Mazur, 1983; Baum, 1993; Killeen, 1995; Foster et al., 1997), and faster for higher magnitude or more desirable reinforcers (Bradshaw et al., 1978, 1981). More specifically, each type of schedule brings about a characteristic response pattern (illustrated in Figure 2.1). Response rates on random ratio and random interval schedules are relatively constant throughout the session (bar the occasional pauses in order to consume the rewards earned, not illustrated in the figure), however, these rates are higher for ratio schedules compared to yoked interval schedules. That is, an animal working for rewards on an interval schedule in which the intervals match (are 'yoked' to) the temporal intervals between successive rewards earned by a second animal who is responding on a ratio schedule, will

[2]In the psychological literature, '*molar*' refers to large units of behavior, such as overall response rates or holistic aspects of reinforcement schedules. In contrast, '*molecular*' analyses and theories pertain to single actions or local aspects of responding (Reber, 1985).

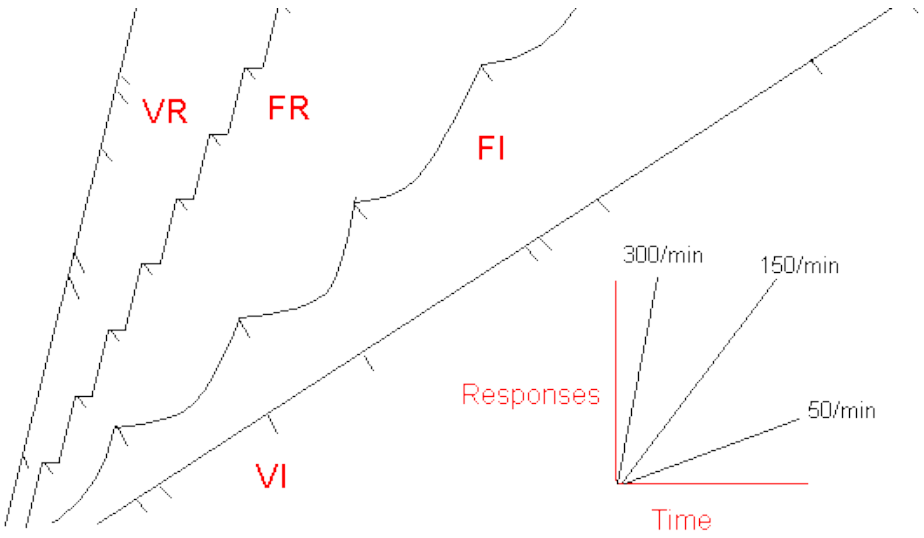

Figure 2.1: An illustration of the response patterns generated by different reinforcement schedules. As was the tradition at the time when Skinner conducted his experiments, responses (y axis) over time (x axis) are marked by a moving pen such that the slope of each trace represents the rate of responding (box inset in bottom right). Large pen displacements represent rewards. Note the constant response rates on variable interval and ratio schedules, and, in contrast, the 'scalloping' response pattern in fixed interval schedules (but see Gallistel et al., 2004; Taylor et al., 2006), and the post-reinforcement pauses in fixed ratio schedules. FR = fixed ratio; VR = variable ratio; FI = fixed interval; VI = variable interval.

show considerably slower overall response rates compared with that second animal (Zuriff, 1970; Catania et al., 1977; Dawson & Dickinson, 1990; Baum, 1993; Cole, 1999).

Responding on fixed interval schedules, by contrast, provides evidence that the animals time the length of the interval, and given the statistics of their timing errors (which are known to scale with the length of the interval; Gibbon, 1977, 1992; Gallistel & Gibbon, 2000), try to respond only at its end. Response rates within an interval are therefore not constant: after a reward is obtained there is a pause in responding (post reinforcement pause; PRP), which is not solely the result of reward consumption, as its length is correlated with the mean schedule interval. When averaging response rates over animals and intervals, as has been done traditionally, responding on fixed interval schedules seems to follow a 'scalloping' pattern with a low rate of responding after the PRP, and accelerating to a higher response rate up to the end of the interval (see Figure 2.1). However, recent molecular-level analyses of response rates of individual animals in individual intervals suggests that once they have started responding, response rates within an interval are in fact constant, and the molar scalloping pattern is actually an artifact of the averaging over different starting times (eg. Gallistel et al., 2004; Taylor et al., 2006). Finally, fixed ratio schedules tend to generate responding which is quite regular, bar a paradoxical PRP which appears in high ratio schedules (Felton & Lyon, 1966). That this PRP is relatively matched in length to that on a yoked interval schedules, suggests that it arises because animals confuse the long inter-reinforcer intervals due to the time it takes them to complete the ratio requirement, with an interval schedule.

The fact that response rates are affected by manipulations of the schedule of reinforcement suggests that animals choose with the rate with to perform actions. Furthermore, in most cases, animals' behavior in such schedules is well below ceiling rates, evidence that free-operant response latencies are not constrained by decision times, motor or perceptual requirements. We can therefore assume with some confidence that animal are *selecting the particular response latencies* (or the overall response rate) as an adaptation to the reinforcement schedule.

In the following I will assume that this choice is the result of an optimization process that is influenced by two opposing goals: the desire to acquire rewards quickly and the wish to minimize effort costs. Negotiating this tradeoff leads to optimal behavior, with respect to the obtained net rate of reinforcements minus costs. In this chapter, I will propose a normative model of rate selection, and use it to investigate this optimum and the characteristics of the behavior it entails (the online learning algorithm by which animals can acquire this optimal behavior is developed later, in Chapter 5). But first, I review in some detail several experimental results characterizing free-operant behavior, which the model will need to reproduce and explain. I will start with molar characteristics, namely, the difference in response rates between ratio and interval schedules (Dawson & Dickinson, 1990; Baum, 1993), the hyperbolic relationship between response rate and payoff in simple interval schedules and the well-known matching behavior in concurrent interval schedules (Herrnstein, 1961, 1970, 1997). I will then continue to the molecular aspects of the fine-scale temporal structure of free-operant responding, as measured in my own experiments (eg. Niv et al., 2005).

2.1.2 Characteristics of free-operant behavior

Free-operant experiments are typically conducted in an operant chamber. In the experiments we are interested in, this is a box containing manipulanda such as levers or chains, a food magazine to which reinforcers can be delivered, and different controllable stimuli such as lights and sounds that can be turned on or off. The experimental session is typically controlled by a computer that implements the experimental protocol by controlling the delivery of reinforcers and the status of the different stimuli. The computer also registers the occurrences of a predefined subset of actions (such as presses of the lever, or movements of a flap guarding the opening to the food magazine), if and when these are performed by the animal. Importantly, in the free-operant experiments I will describe below, the animal is required to perform an 'instrumental response', an action that is not common in its behavioral repertoire, in order to earn rewards according to some predefined schedule of reinforcement.

Ratio versus interval schedules

As mentioned previously, animals exhibit quantitatively different instrumental behavior depending on the type of reinforcement schedule used. The most direct demonstration of this is by 'yoking' an interval schedule to a ratio schedule. In this, one group of animals is trained on a (fixed or random) ratio schedule.

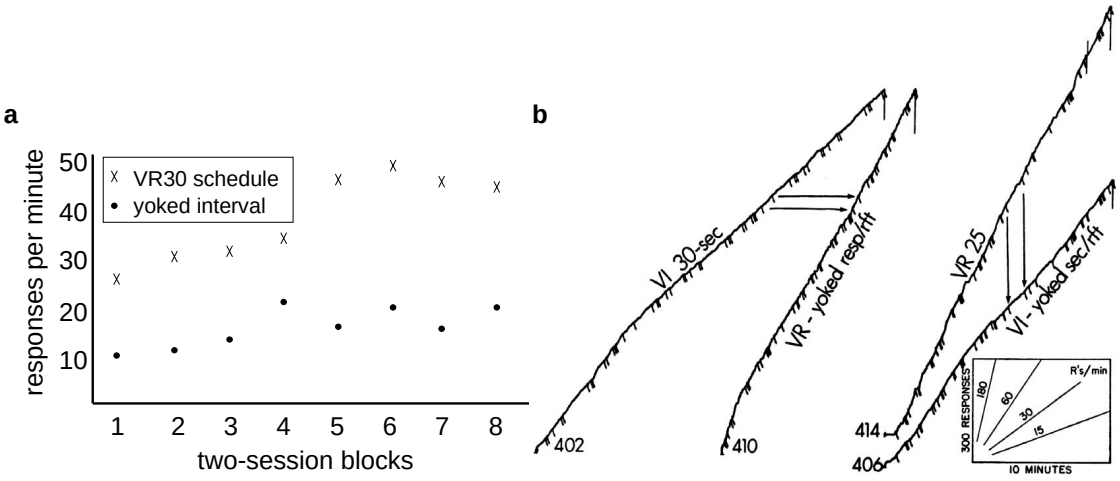

Figure 2.2: Responding on ratio and interval schedules. **a.** Leverpress response rates of rats trained on a VR30 schedule, over 20 sessions of training, and a corresponding group of rats that were trained on an interval schedule, with each rats' intervals yoked to the inter-reward intervals experienced by a ratio-schedule rat (adapted from Reed et al., 2000). **b.** Cumulative response records of two pairs of pigeons on yoked ratio and interval schedules (adapted from Catania et al., 1977). Responses over time are marked by a moving pen such that the slope of each trace represents the rate of responding (box inset in bottom right). Large pen displacements mark reward delivery. The number of responses per reward of pigeon 402, trained on a VI30 schedule, served as the variable ratio required per reinforcer from pigeon 410 (right-going arrows). The time intervals between rewards experienced by pigeon 414, trained on a VR25 schedule, served as the intervals in a variable interval schedule for pigeon 406 (down-going arrows). In all cases, responding on the interval schedule was at a considerably lower rate compared to that on the ratio schedule.

Another group is then trained on a variable interval schedule, with the intervals between rewards chosen such that they match those experienced *de-facto* by the ratio-schedule group. Thus the reward rates experienced by the two groups are similar[3]. Despite this similarity, the experimental results reveal markedly reduced response rates in the interval-schedule group, compared to those of the ratio-schedule group (see Figure 2.2; Zuriff, 1970; Catania et al., 1977; Dawson & Dickinson, 1990; Baum, 1993; Cole, 1999; Reed et al., 2000).

The "Matching Law" and the relationship between response rate and reinforcement rate

In random interval schedules, the relationship between response rate and reinforcement rate is typically a hyperbolic saturating one (Figure 2.3a). More famously, if two response options are concurrently available (say, right and left levers), each associated with a separate interval schedule, animals 'match' the ratio of responding on the two levers to the ratio of their experienced reward rates. For instance when one lever is reinforced on an RI30 schedule, while the other is reinforced on an RI15 schedule, rats will press the latter lever roughly twice as fast as they will press the first lever. Figure 2.3b demonstrates this result in response rates of pigeons pecking at two concurrently available keys associated with a variety of interval schedules.

[3]Only the reward-baiting intervals can be controlled by the experimenter, which means that the inter-reward intervals experienced by the interval-schedule group are not strictly matched to those experienced by the ratio-schedule group

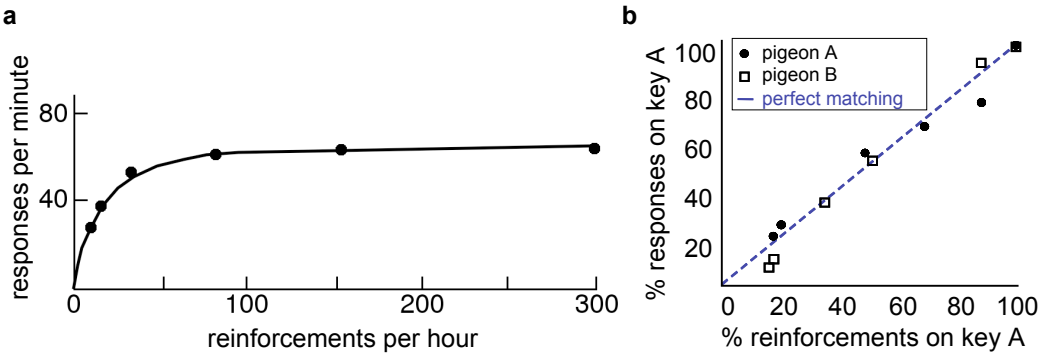

Figure 2.3: a. The relationship between response rate and reinforcement rate (the reciprocal of the interval between reinforcements) on a VI schedule, is hyperbolic (ie, of the form $y = B \cdot x/(x+x_0)$, solid line fit). This can be seen as an instantiation of Herrnstein's "Matching Law" for one instrumental response (adapted from Herrnstein, 1970). **b.** Classic "Matching Law" behavior: the proportion of pecks on key A is roughly equal to the proportion of rewards obtained on this key (adapted from Herrnstein, 1961). Note that rates in these cases are measured as overall number of responses in a session (McSweeney et al., 1983), uncorrected for factors such as eating time.

These molar relationships between response rate and reinforcement rate were discovered and studied in detail by Herrnstein, who formulated the "Matching Law" (Herrnstein, 1961, 1970, 1997):

$$\frac{r_1}{r_1+r_2} = \frac{R(r_1)}{R(r_1)+R(r_2)} \tag{2.1}$$

in which r_1 and r_2 designate response rates on each of two available options, and $R(r_i)$ their corresponding reward rates. An alternative formulation replaces response rates with time allocated to each of the responses:

$$\frac{t_1}{t_1+t_2} = \frac{R_1}{R_1+R_2} \longrightarrow \frac{t_1}{R_1} = \frac{t_2}{R_2}, \tag{2.2}$$

(where t_i is the time allocated to option i and R_i is the reward per unit time from option i). This shows that the "Matching Law" implies equating the returns (reward rate per unit time invested) from the two options.

When only one instrumental action is available, Herrnstein explained the hyperbolic relationship between the animal's instrumental response rate (r_{lever}) and the (experimenter determined) rate of reinforcement $R(r_{lever})$ as an instantiation of the "Matching Law" for one action (Baum, 1993). By assuming the existence of other, intrinsically rewarding actions (performed at rate r_{other}, with the intrinsic reward rate of $R(r_{other})$), and assuming a fixed total rate of behavior in the experiment ($r_{lever} + r_{other} = r_{total}$; or a fixed total time in the experiment), equation (2.1) gives the hyperbolic relation:

$$r_{lever} = \frac{r_{total} \cdot R(r_{lever})}{R(r_{lever})+R(r_{other})}. \tag{2.3}$$

Although postulated as a general law relating response rate and reinforcement rate (eg. Herrnstein, 1990), the matching relationship is actually far from being universally true (eg. Wearden & Burgess, 1982; Dallery

& Soto, 2004; Soto et al., 2006, 2005). For instance, whether response rates match reinforcement rates on concurrent interval schedules is sensitive (Pliskoff & Fetterman, 1981; Baum, 1982; Boelens & Kop, 1983) to the existence of a penalty for switching from one option to the other (a 'change-over delay'; COD; Herrnstein, 1961, 1970; Williams, 1994). In the absence of some penalty, whether implicit in the need to travel between two distant levers (as with foraging on distant food patches; Baum, 1982), or explicitly imposed by requiring a certain number of actions or a minimal amount of time to pass before the schedule on the switched-to option resumes (Shull & Pliskoff, 1967; Sugrue et al., 2004), animals simply alternate rapidly between the two options (Herrnstein, 1961, 1970).

The fine-scale temporal structure of responding on random schedules

Even the simplest of free-operant schedules typically involves not only an instrumental response, but also other implicitly required actions. Most commonly, while a certain instrumental response (say, leverpressing; LP) fulfills the schedule requirement, yet another response (for instance, poking the nose into the food magazine; nosepoking; NP) is necessary in order to gain access to the reinforcer. Figure 2.4 shows the molecular structure of responding of well-trained rats trained in such a scenario, to press a lever in order to earn sucrose solution (which was delivered into a food magazine) on a VI30 schedule of reinforcement. (Data are averaged over animals and sessions.) Figure 2.4a shows that after a reinforcer was earned (the delivery of which could presumably be heard), the rats rapidly executed a nosepoke action (median latency to the first nosepoke: 0.49s). Leverpress and nosepoke response rates as a function of the time since this consummatory NP are shown in Figure 2.4b. In the first several seconds, measured responding (NPs and LPs) was low, after which responding resumed, with the average rate of leverpressing increasing up to a constant response rate, and the rate of nosepoking following suit[4] (see section 2.4 below for a discussion of these 'spurious' nosepokes in the absence of food in the magazine). The pause in responding was of similar average length during different days of preliminary training on different schedules (FR1, VI2, VI15), and thus is likely caused by the rats' eating of the food, and should not be confused with the PRP characteristic of fixed interval or fixed ratio schedules, which takes different length depending on the schedule parameter. Analysis of the leverpress rates of individual rats (the details of which are beyond the scope of this chapter) showed that leverpressing after the consumption pause was in fact at a constant rate, and the apparent gradual increase in response rates in the figure is an artifact of averaging over rats and inter-reward-intervals. A similar overall pattern is also characteristic of responding on RR schedules.

A more detailed analysis of the individual responses within the relatively constant rates of responding shows that the rats' behavior is actually very variable. Figure 2.5 shows the distributions of inter-response latencies between consecutive actions. The inter-response times (IRTs) seem to suggest a random process with some underlying rate, and a (relative, as well as an absolute) 'refractory period' (or period of absolute or relative inaction) following the execution of an action. Indeed, the IRT distributions can be described well by Gamma

[4]The early rise of the rate of nosepoking, before that of leverpressing, was not a stable characteristic in different rats and different experiments, and should not be considered consequential.

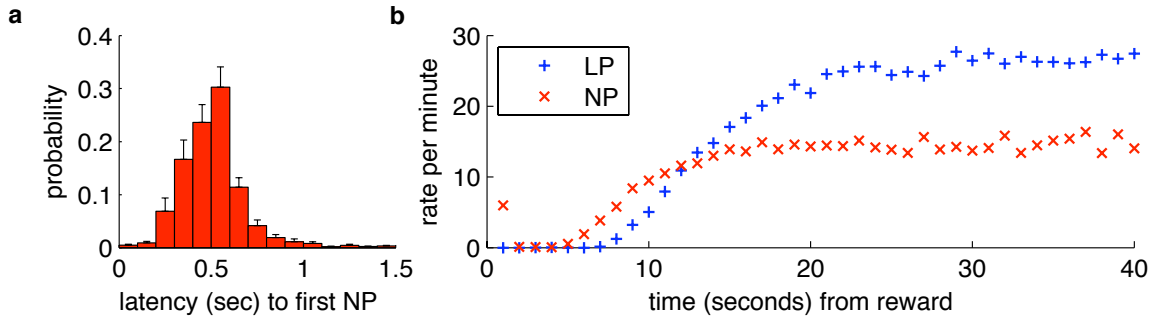

Figure 2.4: Responding of well-trained, food deprived rats on a VI30 schedule of reinforcement (see Chapter 6 for details of the experiment). Data are from the last five (of fifteen) days of training, averaged over 19 rats earning 30 reinforcers each per session. **a**. Probability histogram (and standard errors of the mean) of the time to the first (consummatory) nosepoke after a reinforcer is delivered. Overall median latency was 0.49s, mean 0.53s, standard deviation 0.39s. **b**. Subsequent rate of leverpresses (blue crosses) and nosepokes (red pluses), as a function of seconds from actual reward (ie, from the time of the first NP). Average leverpressing builds up to a relatively constant rate, following a rather long pause after gaining each reward, during which the food is consumed (see text). Here the response rate at timepoint t was computed by counting all responses emitted t seconds after a consummatory nosepoke providing the next reward was not yet baited (ie, counting responses only in the time between the first NP after a reward to the next to last LP before the next reward), and dividing by the number of inter-reward-intervals of this length (again measuring the interval length as that between the first NP to the next to last LP).

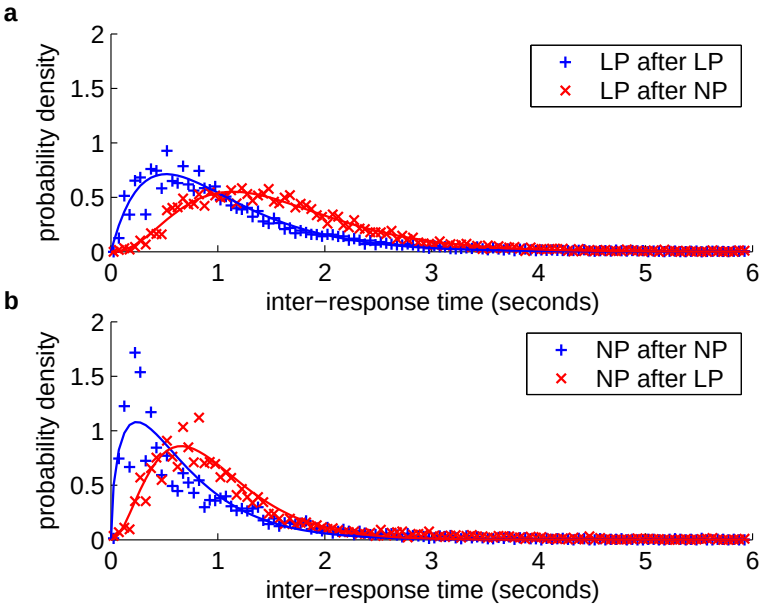

Figure 2.5: Inter response distributions for consecutive actions. The data analyzed are the same as those in Figure 2.4. Latencies were measured between every pair of consecutive actions at times in which reward was not available in the food magazine, nor baited (ie, excluding the interval to the rewarded LP, and that to the consummatory NP). **a.** Latencies to leverpresses, either after a leverpress (blue +-s) or a nosepoke (red x-s). **b.** Latencies to nosepokes, either after a nosepoke (blue +-s) or a leverpress (red x-s). Solid lines: best fit Gamma probability density function distributions. Note that the only actions recorded in the experiment were LPs and NPs, thus the potential occurrence of another unrecorded action between two consecutive recorded actions is not controlled for.

probability distribution functions (solid line fits; see also Killeen & Sitomer, 2003), which are commonly used to describe a Poisson (memoryless, constant rate) process, with a positive 'refractory period' after every event (see also Killeen & Fetterman, 1988).

The two-parameter (α, θ) Gamma distribution is defined over non-negative numbers $(x \geq 0)$

$$p(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\theta^{\alpha}\Gamma(\alpha)} \tag{2.4}$$

and can be interpreted as the waiting time until the α-th event in a Poisson process with a rate of $1/\theta$. Its maximum (and mode) is attained at $(\alpha - 1)\theta$ (for $\alpha > 1$), its mean is $\alpha\theta$ and its variance $\alpha\theta^2$. The IRT distributions are characterized by a low probability of very short IRTs and a peak at some intermediate latency, after which the probability of longer latencies drops fairly exponentially, implying that $\alpha > 1$.

The IRT distributions further reveal that the latency to perform an action is not independent of the previous action. The 'refractory period' and the latencies are on average longer when switching between two different actions (ie, the latency of an LP after an NP is longer on average than that of an LP after an LP (Figure 2.5a), and similarly for an NP after an LP compared to after an NP (Figure 2.5b)). One possible reason for this is that shifting from leverpressing to nosepoking, or vice versa, requires physical movement to a different location in the operant chamber, which, in itself, takes some minimal amount of time.

2.2 Methods: Modeling free-operant behavior

The above phenomena provide straightforward test cases for a model of free-operant behavior. Importantly, a model which treats the choice of response rate as the result of an optimization process, may shed light on the *reason* for these behavioral phenomena. In this section, I describe a reinforcement learning model in which an agent optimizes the choice of both *which action* to perform, and *with what latency* (or vigor) to perform it, taking into account both the immediate and the long term consequences of these choices. The model treats response vigor as being determined normatively, as the outcome of a battle between the cost of behaving more expeditiously and the benefit of achieving desirable outcomes more quickly. I will first describe the two characteristics of the model that set it apart from existing RL models of instrumental conditioning – these are the building blocks for modeling free-operant response rates rather than discrete action selection. The formal model will then be laid out, followed by two ways to derive the optimal behavioral strategy. The Methods section will conclude with a short discussion of several implementation practicalities.

2.2.1 Building blocks

We model the free-operant scenario common in experimental psychology, in which a rat is placed in an experimental chamber and can choose freely which actions to emit and when. Most actions have no pro-

grammed consequences; however, one action (leverpressing; LP) is sometimes rewarded with food, which falls into a food magazine. For simplicity we assume that food delivery makes a characteristic sound, signaling its availability for harvesting via a nose poke action (NP) into the magazine. In the following, we model this task as a continuous-time semi-Markov decision process (SMDP; Puterman, 1994), and treat within a reinforcement learning framework the problem of optimizing a policy (which action to perform, and at what latency, given the state) that maximizes the *net* rate of delivery of rewards minus costs.

There are two important building blocks on which the model relies. First, as already mentioned, both which responses to emit, as well as with what rate these are to be emitted, are subject to optimization. Thus we model response selection as the choice of a *pair*: an action $a \in \{\mathrm{LP}, \mathrm{NP}, \mathrm{Other}\}$ (where $a = \mathrm{Other}$ is taken to include the various miscellaneous behaviors such as grooming, rearing, and sniffing which animals typically perform during the experiment) and a latency $\tau \in (0, \infty]$ with which to perform it. The latency can also be seen as the instantaneous response rate, or the *vigor* with which the action is performed — to complete a leverpress within a shorter time, the animal must press more vigorously. A sequence of choices of $\{(a_i, \tau_i), i = 1, 2, ...\}$ pairs will constitute a choice of response rate for each of the actions a, which can be measured globally (as the reciprocal of the mean time between two choices of a; the traditional 'molar' measure) or independently per action (as the reciprocal of the mean latency to perform this action).

The second building block is the *cost* of performing the chosen actions. Specifically, if we do not assume that actions incur costs, in addition to possibly gaining rewards, the trivially optimal response rate will be to respond as fast as possible. The pattern of latency choices in figure 2.5 is characterized by low probability of very short latencies, suggesting that such latencies should be very costly, and the non-uniform distribution of different latencies suggests that at least some portion of the cost of an action should depend on its latency. Accordingly, we assume two types of costs for a chosen (a, τ) pair: a cost per unit action taken, and a cost which is proportional to the vigor of execution, ie, inversely proportional to the latency τ (Staddon, 2001). The first *unit cost* is simply a constant C_u for each action performed[5], and it is levied regardless of the latency with which the action is performed, such that even actions performed with a very long latency are not 'free'. The second cost (which we term the *vigor cost*), is $\frac{C_v}{\tau}$, where C_v is the vigor-cost constant. This cost severely penalizes very short latencies, but is negligible for long latencies[6]. Furthermore, to account for the costs of travel between different parts of the operant chamber, we assume that the unit cost constant C_u and the vigor cost constant C_v can depend on the identity of both the currently chosen and the previously performed action, and in both cases we assume slightly higher cost constants for transitioning between different actions. Together, these costs mean that pressing a lever three times in a second is more costly than pressing it three times in ten seconds, and that pressing a lever after a leverpress is less costly than doing so after a nosepoke. That some of the costs are vigor related means that response rate selection is not trivial — responding too quickly will be very costly, while responding too slowly means rewards will be obtained only after a considerable delay.

[5] This cost can be negative, thereby providing an 'internal' per-unit *reward* for an action. In all results reported here, the unit cost for $a = \mathrm{Other}$ was the only cost assumed to be negative. Note that such a per-unit reward is, in effect, a concurrent *ratio* schedule (FR1) for this action.

[6] While we chose this particular formulation for the vigor cost due to its simplicity, a variety of other monotonically decreasing functions in positive τs could be used instead, possibly yielding different qualitative results. As will be described in Section 2.3, the formulation we chose yielded results that closely match known behavioral patterns on several levels.

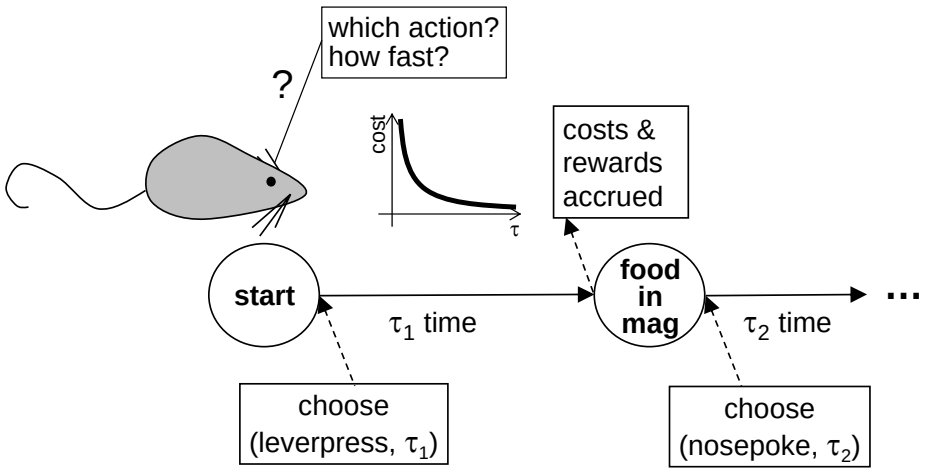

Figure 2.6: Model dynamics: The simulated rat begins at the start state, and selects both an action a to perform and a latency τ with which to perform it. For instance, here the rat selects the (action,latency) pair of (LP,τ_1). As a simplification we assume that the leverpress action is then executed (to the exclusion of all other actions) throughout the chosen latency τ_1. At the end of this interval the action is completed, at which time all costs and available rewards are accrued, and any state changes in the environment occur. In the ensuing state the rat again selects an (a, τ) pair, and the process repeats.

2.2.2 The model

The dynamics of the model are illustrated in Figure 2.6. We formally model the free-operant task as a continuous-time semi-Markov decision process with an observable state space. Based on the current state $s \in \mathcal{S}$, the simulated rat chooses an (a, τ) pair. For simplicity we consider punctate actions, with the latency τ related to the vigor with which this action is performed[7]. Critically, we simplify the dynamics of the task by treating this choice as an exclusive commitment to devoting all of the next τ time to performance of the action a. After time τ has elapsed, the action is completed, the rat receives rewards (if nosepoking when a reward is available in the magazine), incurs unit- and vigor-costs associated with its choice (see above), and the transition to the next state occurs[8]. The rat then selects a new (a, τ) pair based on the new state, and the process repeats. To allow comparison with experimental results, we also assume that eating a reward pellet is itself time consuming (see Figure 2.4 and Foster et al., 1997). Thus, whenever a nose-poking action is chosen and food is available in the magazine, a variable 'eating time' with a mean t_{eat} of several seconds, must pass before next decision point (and the next state) is reached.

[7]Ideally, the choice of action would specify the identity of the action, how fast (or how vigorously) to transition to it, and *for how long* to perform it, with the latter two independent of each other. However, as an initial approximation we assume punctate actions with a fixed, negligible length. This considerably simplifies the dynamics, at the expense of a partially inaccurate description of the decision process. Specifically, in this scheme, performing two consecutive nosepokes is different from performing one doubly long nosepoke, as the former consists of performing two actions rapidly, and the latter one action slowly.

[8]Note that state transitions can occur in our model only at the completion of actions. This means that the state of the world can not change during the execution of an action, ensuring that the optimality of a chosen response will not change during its execution. The shortcomings of such a modeling choice will be discussed in Section 2.4.3.

We consider the simplified case of an observable state space comprised of all the information relevant to the task. Moreover, we include in the state space all the history information that is necessary in order to make the task Markovian (that is, one in which the probability of transitioning to a state is only dependent on the current state and the chosen action and latency pair, and not on the past history of states). Specifically, the state space includes the identity of the previous action taken (a_{prev}), an indicator as to whether a reward is or is not available in the food magazine (i_r), and, as necessary, the number of LPs since the previous reinforcement (n_{LP}, for FR schedules) or the elapsed time since the previous LP (t_{LP}, for RI schedules, and $\{t_{LP1}, t_{LP2}\}$ for concurrent interval schedules in which two levers are available). Using this state space, the (probabilistic) transitions between the states $\mathcal{T}^{a,\tau}_{S\to S'} = P(S'|S,a,\tau)$ as well as the probability of reward for each action at each state $P_r^{a,\tau}(S) = P(reward|S,a,\tau)$ are Markovian and completely defined by the reinforcement schedule. For instance, in a random-ratio 5 (RR5) schedule, every $a =$ LP action has a probability of 0.2 of inducing a transition from the state in which no food is available in the magazine ($i_r = 0$), to that in which food is available ($i_r = 1$). In a fixed-ratio 5 (FR5) schedule, in contrast, an $a =$ LP action in a state in which four leverpresses have already been performed since the last reward ($n_{LP} = 4$) always transitions to the ($i_r = 1, n_{LP} = 0, a_{prev} =$ LP) state, while leverpressing in other states increments n_{LP} without leading to food in the magazine. In all schedules an $a =$ NP action in the 'no-reward-available' state is never rewarded, and, conversely, is rewarded with certainty ($P_r = 1$) in the 'food-available-in-magazine' state.

To recap, the action selection problem faced by the rat can be characterized by a series of states $S \in \mathcal{S}$ in each of which the rat must choose an action and a latency (a,τ) which will entail a unit cost $C_u(a,a_{prev})$ as well as a vigor cost $C_v(a,a_{prev})/\tau$ and result in a possible transition to a new state $S' \in \mathcal{S}$, as well as a possible immediate reward with utility U_r (Figure 2.6). The animal's behavior in the experiment is fully described by the successive actions and latencies chosen at the different states the animal encounters $\{(a_i,\tau_i,S_i), i = 1,2,3,...\}$ (where i counts choices). Note that the model dynamics specify an SMDP in which the rat's choices determine the dwell times at the different states, and the rewards and subsequent states possibly depend on these dwell times[9]. The action space here is continuous (due to τ), as is potentially the state space (specifically, in interval schedules, where it includes continuous elapsed time, t_{LP}), and all rewards and costs are harvested at state transitions and considered as point events.

To complete the formal specification of the task, we must now describe the goal for the rat, in the sense of what we consider it to be optimizing. There are two possible optimization metrics (Mahadevan, 1998), both leading to similar qualitative results, but allowing slightly different theoretical interpretations. In the following, we first treat the problem of optimizing a policy (which action to take and with what latency, given the state) in order to maximize the *average rate of return* (rewards minus costs per unit time). This formulation is slightly different from the more common exponentially discounted form of RL that we described in Chapter 1, and is better suited for modeling many aspects of free-running, free-operant behavioral tasks which have a cyclic nature and are not externally divided into trials and inter-trial intervals (Daw, 2003; Daw & Touretzky, 2002; Daw et al., 2002). Further, ethological studies provide empirical support for average

[9]Because the rat's choices uniquely determine the dwell times, the process is still strictly an SMDP despite the dependence of state transitions on dwell times, as this can be expressed as a dependence on actions instead.

reward rate optimization (eg. Bautista et al., 2001; Staddon, 1992), and this type of RL formulation has also previously been related to hyperbolic discounting (Ainslie, 1975), which, in behavioral tests, typically bests exponential discounting (Kacelnik, 1997). This formulation allows an interesting understanding of the underlying choice tradeoff, and so we discuss it in greater length than we do the second, more traditional, formulation, in which an *exponentially discounted sum or future rewards* is maximized.

2.2.3 Average reward optimization

In this section, we formalize the problem for the rats in terms of optimizing the long run rate of return (or net reward rate). That is, the goal of the rat is to choose actions and latencies that maximize the accumulated rewards less the incurred costs, per unit time. In order to find the behavioral policy which achieves this goal, we use *average reward* RL techniques to determine the value of each (a,τ) pair at each state in the task (defined in terms of the future expected rewards when taking the given action with the given latency, relative to the expected net reward rate; Puterman, 1994; Bertsekas, 1995b; Mahadevan, 1996). Once these so-called differential Q-values (Watkins, 1989) are known, optimal performance can be achieved simply by choosing, in every state, the (a,τ) pair with the highest value. In this chapter we will not discuss how these values may be learned from online experience, or the detailed timecourse of learning, which will be discussed at length in Chapter 5. Suffice it to say that the values *can* be acquired incrementally from experience with the task, and by using only local information (Schwartz, 1993b; Mahadevan, 1996), and that such learning rules have previously been used as models of phasic and tonic dopamine and serotonin signaling (Daw & Touretzky, 2002; Daw et al., 2002; Daw, 2003). Here we concentrate instead on the *properties of the optimal solution*, that is, how we expect well trained animals to behave in the steady state. In this case we can use model-based techniques such as value iteration methods (see below; Bertsekas, 1995b; Bertsekas & Tsitsiklis, 1996) to find the optimal values, the characteristics of which are our main concern.

In the average reward formulation, we define the policy-dependent *differential (or relative) value* of a state, denoted $V^\pi(S)$ (where $\pi(a,\tau|S)$ is the behavioral policy), as the expected integral of future rewards minus costs encountered from this state and onward, compared to the (policy dependent) expected net reward rate $\bar{R}^\pi$ (Schwartz, 1993b; Puterman, 1994; Mahadevan, 1996)

$$V^\pi(S) = E\left[\int_0^\infty (Rewards(S_t) - Costs(S_t) - \bar{R}^\pi)dt \,\middle|\, S_0 = S\right] \tag{2.5}$$

where the expectation is over the possible action and latency choices, and the resulting state transitions. Actions and latencies are assumed to be chosen according to the current behavioral policy $\pi = \{P(a,\tau|S), S \in \mathcal{S}, a \in \mathcal{A}, \tau > 0\}$, and state transitions are according to the schedule dynamics $\mathcal{T}^{a,\tau}_{S\to S'}$. Because in our simplified model rewards are gained and costs are levied only at punctate state transitions, we can redefine the values as a sum over discrete state transitions

$$V^\pi(S) = E\left[\sum_{n=0}^{\infty} (U_r(S_n) - C_u(a_n, a_{n-1}) - \frac{C_v(a_n, a_{n-1})}{\tau_n} - \tau_n \cdot \bar{R}^\pi) \,\middle|\, S_0 = S\right] \tag{2.6}$$

where $\{(a_n, \tau_n)\}_{n=0}^{\infty}$ is the series of actions and latencies chosen according to the current behavioral policy π, and $U_r(S_n)$ is the utility of the reward given at state $S = S_n$. By dividing this equation into its first and subsequent components, and using the previously defined notation, this sum can be expressed recursively

$$V^{\pi}(S) = \int \pi(a,\tau|S) \left[P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} - \tau \cdot \bar{R}^{\pi} + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} \cdot V^{\pi}(S') \right] d\pi \quad (2.7)$$

where (a, τ) is the immediate behavioral choice (according to π), and the sum over successor states turns into an integral if the state space is continuous. Thus the policy-dependent differential value of a state is the expected reward minus cost due to the current action and latency, plus the value of the next state (averaged over the possible next states), compared to the immediately forfeited net reward rate (the integrated rate of rewards minus costs).

Equation (2.7) is the *consistency relationship* that the correct state values must satisfy, which can be used constructively (in the 'relative value iteration' algorithm first presented by White, 1963) in order to find these values for policy π (Bertsekas, 1995b; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; see Bertsekas (1998) for a proof of the validity and consistency of iterating on differential values). In order to find the *optimal* differential values of the different states, that is, the values $V^*(S)$ (and net reward rate $\bar{R}^*$) given the optimal action selection strategy, we can simultaneously solve (either directly or using value iteration) the set of equations defining these values (the "Bellman equation" of the model, with the *max* operator in order to assume implicitly an optimal policy at each state; Puterman, 1994; Gosavi, 2004b):

$$V^*(S) = \max_{a,\tau} \left\{ P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} - \tau \cdot \bar{R}^* + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^*(S') \right\} \quad (2.8)$$

in which there is one equation for every state $S \in \mathcal{S}$. Now, given the optimal state values, the optimal differential value of an (a, τ) pair taken at state S, denoted $Q^*(S, a, \tau)$, is given by the simple model-based one-step lookahead equation using these values

$$Q^*(S,a,\tau) = P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} - \tau \cdot \bar{R}^* + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^*(S'). \quad (2.9)$$

The theory of Dynamic Programming ensures that there is a unique solution for the optimal attainable net reward rate $\bar{R}^*$, as well as the optimal differential state values $V^*(S)$ and the differential state-action-latency values $Q^*(S, a, \tau)$ (Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Gosavi, 2004b). Note, however, that the fact that the V-values are *differential* means that they are defined uniquely only with respect to each other, that is, they are defined up to an additive constant, and the same is true for the Q-values[10] (Puterman, 1994;

[10] The set of equations $\mathbf{V}^* = \mathbf{r}^* - \mathbf{c}^* + T\mathbf{V}^* - \bar{R}^*$ (where T is the (deterministic) optimal-policy-dependent transition matrix between states, and $\mathbf{V}^*$, $\mathbf{r}^*$ and $\mathbf{c}^*$ are the policy dependent values, immediate rewards and immediate costs, respectively, per state) can be written as $(1-T)\mathbf{V}^* = \mathbf{U}(\bar{R}^*)$ where $\mathbf{U}$ is a vector dependent on $\bar{R}^*$. This is a system of $N = |\mathcal{S}|$ equations with $N+1$ variables ($\mathbf{V}^*$ and $\bar{R}^*$). As T is a transition matrix with rows summing to 1, $(1-T)$ is rank $N-1$ and not invertible. However, it is easy to show that $\bar{R}^*$ is uniquely defined as the value for which $\mathbf{U}(\bar{R}^*)$ is orthogonal to the first eigenvector of $(1-T)$, and the values $\mathbf{V}^*$ are then uniquely defined up to an additive constant.

Gosavi, 2004b). Furthermore, by virtue of the definition of the Bellman equation, it turns out that the scalar $\bar{R}^*$ which uniquely solves it, is the optimal net reward rate, ie, the expected sum of all the rewards obtained minus all the costs incurred, per unit time, when following the optimal policy (Bertsekas & Tsitsiklis, 1996).

The advantage of the recursive formulation is that it no longer includes an infinite sum, and the values can be found using iterative Dynamic Programming methods such as 'differential value iteration' (Bertsekas, 1995b; Bertsekas & Tsitsiklis, 1996) or through online sampling of the task dynamics and 'temporal-difference learning' (Schwartz, 1993b; Mahadevan, 1996; Sutton & Barto, 1998). Here we use the former, and report results using the true optimal differential values. Specifically, for ratio schedules the simple transition structure of the task allows the Bellman equation to be solved analytically to determine the V-values. For interval schedules, which have a much larger state space, a simple analytical solution is no longer possible and we use differential value iteration with time discretization. We compare these model results to the steady-state behavior of well trained animals, as the optimal values can be taken to correspond to values animals have learned online, throughout an extensive training period. In Chapter 5, we will develop the latter temporal difference learning method, to show that the same results can be reached via a more biologically feasible learning process.

Implementation practicalities

As differential state values are only defined up to an additive constant, finding the unique solution to the Bellman equation requires that we clamp one of the values to zero (Gosavi, 2004b). Importantly, this should be the value of a *recurrent state* S_0, a state that is reachable from any other state under all optimal stationary policies[11] (Bertsekas, 1995a). The state we chose for this purpose is that of just having earned a reward for a leverpress ($i_r = 1, a_{prev} = \text{LP}$), which is reachable at least from all policies in which we are interested (ie, those that leverpress with non-zero probability). The differential value iteration algorithm we used, adapted from Jalali and Ferguson (1990) and Bertsekas and Tsitsiklis (1996), was thus:

Algorithm 2.1

$$
\begin{aligned}
V^T(S) &= \max_{a,\tau}\left\{ P_r^{a,\tau}(S)\cdot U_r - C_u(a,a_{prev}) - \frac{C_v(a,a_{prev})}{\tau} - \tau\cdot\bar{R}^{T-1} + \sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'}^{a,\tau} V^{T-1}(S') \right\} \\
V^T(S) &= V^T(S) - V^T(S_0) \\
\bar{R}^T &= \max_{a,\tau}\left\{ \frac{P_r^{a,\tau}(S_0)\cdot U_r - C_u(a,a_{prev}) - \frac{C_v(a,a_{prev})}{\tau} + \sum_{S'\in\mathcal{S}} \mathcal{T}_{S_0\to S'}^{a,\tau} V^T(S')}{\tau} \right\}
\end{aligned}
$$

[11] Strictly speaking, even the single-lever free-operant SMDP is not unichain under all stationary policies, and furthermore, there is no state that is recurrent under all stationary policies. This is important as the unichain property ensures that there is one global $\bar{R}$ applicable to all states (Puterman, 1994), and the recurrence allows a proof of convergence using a related stochastic shortest path problem (Bertsekas & Tsitsiklis, 1996; Bertsekas, 1998). However, the MDP does satisfy the less stringent conditions for convergence of differential value iteration given in Bertsekas, 1995a, namely that for each pair of states there is a stationary policy under which they are communicating. In fact, under all potentially interesting policies (ie, ones in which the rat leverpresses when there are no rewards available and nosepokes when there are rewards available) the process is unichain, and convergence problems were not encountered.

where T enumerates iterations of the algorithm, and every iteration includes an update of all the state values and of the net reward rate (defined using S_0). The algorithm is stable and typically converges quickly on the SMDPs described here (tens of iterations).

Unfortunately, in concurrent interval schedules in which there are two levers, there is no obvious choice for a recurrent state S_0 because the state includes the time since the last leverpress on the other lever, which could take on any value at the time of earning a reward on one lever. As a result, it is difficult to find the optimal values using standard differential value iteration[12]. However, the existence of an optimal policy is still guaranteed, and furthermore, it is known that for sufficiently small values of the discount factor γ this is also the solution of the corresponding exponentially discounted model. As a result, for concurrent interval schedules we find the optimal policy using the exponential discounting model with a low discount factor (although we have not rigorously worked out the specific factor which ensures the same policy), which is described below.

2.2.4 Discounted reward optimization

The alternative formulation of the model, in which the sum of (exponentially) time-discounted future net rewards is maximized, is quite similar to that of average-reward optimization, and the results of this model are qualitatively similar to the average-reward one. In fact, at least in the discrete-time MDP case, the average-reward formulation is equivalent to the limit of a time-discounted model as discounting becomes negligible (in terms of our formulation below, as $\gamma \to 0$ and $e^{-\gamma\tau} \to 1$; Tsitsiklis & Van Roy, 2002). It also turns out that, for our present purposes, the average-reward framework is more revealing, because the net reward rate quantity holds interesting properties, and controls the vigor of responding in a direct way. For this reason, we will concentrate mainly on the average-reward formulation. However, for completeness, and as we use it for concurrent interval schedules, we describe here the exponential discounting formulation for our model.

In discounted RL, the definition of state values (which are absolute rather than differential) is more straightforward: in order to maximize the sum of exponentially discounted net future rewards, the value of a state (or the Q-value of a state-action-latency triplet) can be simply defined as this sum (Sutton & Barto, 1998)

$$V^{\pi}(S) = E\left[\int_0^{\infty} e^{-\gamma \cdot t}(Rewards(S_t) - Costs(S_t))dt \,\middle|\, S_0 = S\right] \tag{2.10}$$

$$= E\left[\sum_{n=0}^{\infty} e^{-\gamma \sum_{i=0}^{n} \tau_i}\left(U_r(S_n) - C_u(a_n, a_{n-1}) - \frac{C_v(a_n, a_{n-1})}{\tau_n}\right) \,\middle|\, S_0 = S\right] \tag{2.11}$$

[12] This is not a problem for animals choosing actions and latencies in the real world, as we assume they do not use iterative Dynamic Programming techniques at all, but rather use online learning methods to approximate the optimal solution as detailed in Chapter 5.

where $\gamma \geq 0$ is the time discount factor, with higher values implying steeper discounting of the value of future rewards (and costs). Again, $\{(a_n, \tau_n)\}_{n=0}^{\infty}$ is the series of actions and latencies chosen according to the current behavioral policy $\pi = \{P(a,\tau|S), S \in \mathcal{S}, a \in \mathcal{A}, \tau > 0\}$, and $U_r(S_n)$ is the utility of the reward given at state $S = S_n$. The Bellman equation of this discounted model, from which we can find the uniquely defined optimal state values $V^*(S)$ (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), is then

$$V^*(S) = \max_{a,\tau} \left\{ e^{-\gamma\tau} \left[P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} \cdot V^*(S') \right] \right\}, \tag{2.12}$$

and, given the optimal state values, the optimal $Q^*(S,a,\tau)$ values are once again given by the one-step lookahead equation

$$Q^*(S,a,\tau) = e^{-\gamma\tau} \left[P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^*(S') \right]. \tag{2.13}$$

The value iteration algorithm for the exponentially discounted model is thus:

Algorithm 2.2

$$V^T(S) \quad = \quad \max_{a,\tau} \left\{ e^{-\gamma\tau} \left[P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^{T-1}(S') \right] \right\}$$

where every iteration includes an update of all the state values. Like the Algorithm 2.1, this algorithm is stable and typically converges quickly, although convergence is slower for lower discount factors[13].

2.2.5 Action selection

As mentioned, using the Q-values derived from V^*, the animal can select actions optimally (that is, such as to obtain the maximal possible net reward rate $\bar{R}^*$ or the maximal sum of discounted (net) rewards) by comparing the values of the different (a,τ) pairs at the current state, and choosing the action and latency that have the highest value. Alternatively, to allow more variable behavior (and occasional suboptimal 'exploratory' actions which are crucial when values are learned online as in Chapter 5), response selection can be based on the so-called 'soft-max' rule (or Boltzmann distribution) in which the probability of choosing an (a,τ) pair is proportional to its Q-value. Here we adopt this option, with the consequence that the policy we explore in this and subsequent chapters is *not* the deterministic optimal policy, but a near-optimal one in which actions that are 'almost optimal' are chosen almost as frequently as actions that are strictly optimal. Specifically, the probability of choosing (a,τ) in state S is

$$P(a,\tau|S) = \frac{e^{\beta Q^*(S,a,\tau)}}{\sum_{a',\tau'} e^{\beta Q^*(S,a',\tau')}} \tag{2.14}$$

[13] See Tsitsiklis and Van Roy (2002) for an analysis showing that, at least for MDPs, convergence can be speeded using an appropriate additive constant, with which the transient (and not only asymptotic) behavior of discounted and average reward RL is equivalent.

where β is the inverse-temperature controlling the steepness of the soft-max function (a value of zero corresponds to high temperature and uniform selection of actions, while higher values correspond to low temperatures and a more maximizing strategy). Soft-max action selection is indifferent to adding a constant to all Q-values, so it can be used with both absolute and differential values.

Finally, the model has a small number of free parameters, notably, the cost constants C_u and C_v for the different action pairs, the reward utility U_r, the eating time t_{eat}, the discount factor γ and the soft-max inverse temperature β. In this thesis I make no attempt to fit these to the experimental data, as this is not necessary for the aim of replicating basic aspects of free-operant behavior qualitatively in order to understand the normative foundations of response vigor. The results described below are thus general, robust, characteristics of the model, independent of specific settings of its parameters. Unless otherwise stated, the model parameters for all simulations below were: $U_r = 60$; $t_{eat} \sim \text{Uniform}(3, 9sec)$; $\beta = 4$; $C_u(a, a_{prev})$ were only dependent on the currently chosen action as follows: $C_u(\text{LP}) = 0.15$, $C_u(\text{NP}) = 0.15$, $C_u(\text{Other}) = -0.15$; and $C_v(a, a_{prev})$ were only dependent on whether the current and previous action were the same one, or different actions as follows: $C_v(\text{same}) = 0.5$, $C_v(\text{different}) = 1.5$.

2.3 Results

This section details the qualitative results of the above described reinforcement learning model of response rates, and compares its asymptotic behavior to that of well-trained animals behaving at steady-state. Specifically, I will compare the characteristics of free-operant behavior as described in section 2.1.2 to the behavior of the model, to show that this simple normative framework captures many classic features of free-operant animal behavior that have eluded previous RL treatments. My goal is twofold: on the one hand I wish to establish the validity of this model as a model of free-operant behavior, as a basis for using it in later chapters to model the effects of motivation on habitual behavior (Chapter 3) and to relate the choice of response rates to possible underlying neural mechanisms (Chapter 4). On the other hand, I am interested not only in replicating the basic characteristics of free-operant behavior, but also in understanding *why* animals might behave this way. To this end, I will analyze the model with the aim of elucidating the *normative* basis for the choice of response vigor. As a first step, let us examine the determinants of decision-making in the model.

2.3.1 Cost/benefit tradeoffs and the opportunity cost of time

Response selection in the model consists of a pair of choices: selecting which action to perform, and at what latency to perform it. Optimal responding is attained by choosing the (a, τ) pair with the highest Q-value, which, in the average reward formulation, amounts to maximizing the function

$$Q(S, a, \tau) = P_r^{a,\tau}(S) \cdot U_r - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} - \tau \cdot \bar{R}^* + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^*(S') \tag{2.15}$$

with respect to a and to τ. This can be viewed as the optimization of a tradeoff between the costs and benefits of instrumental responding. The benefits are represented in the first and last terms of the above equation: the potential acquisition of rewards and the transition into more valuable states of the world. Subtracted from these are the costs of responding: the unit cost, the vigor cost, and a third term — the net rate of reward $\bar{R}^*$.

How can we understand the conceptual role of this last term? Recall that one assumption of the model is that choosing to perform an action with a latency τ means that the next τ seconds will be devoted *exclusively* to the performance of this action. This means that any other benefits that could have been accrued during this time from performing any other actions, are forfeited. Thus latency selection must take into account the cost of lost time — the net reward rate that *could have* been earned during this time, had it not been devoted to the chosen action. Under this interpretation, the net reward rate is measures the *opportunity cost* of time. An (a, τ) pair that stands to earn more net benefit than is potentially lost during time τ is worth performing, while responses that will earn less than is lost should not be chosen. In the optimal policy, the net gain minus loss will be on average zero, as optimal action selection gains on average exactly $\bar{R}^*$ per unit time.

Note, however, that the choice of action and the choice of latency are influenced differently by these various costs and benefits. The first and second terms in equation (2.15) (the immediate reward and the unit cost) are not dependent on τ, and so they only affect the choice of which action to perform. The vigor cost and the opportunity cost both affect the choice of latency. Finally, depending on the schedule of reinforcement, the expected value of the subsequent state can depend on either the chosen action, or the chosen latency, or both (see Section 2.3.3 below), and so can affect either or both choices. Note, also, that the exponentially discounted model computes the same optimal tradeoff (as the models are formally equivalent in the limit of no discounting), even though there the average reward rate is not explicitly represented.

In terms of control of free-operant response rates, a key insight from the model is that by globally determining the opportunity cost of time, the expected net reward rate should influence the rate of *all* chosen actions. That is, if the net rate of reward is high, every second in which an action is not performed (and, necessarily, a reward is not delivered) is costly, and therefore it is worth the subject's while to perform actions more speedily, despite the energetic costs of doing so. In Chapter 3, this result will provide one component of an account of the effects of motivation on free-operant behavior, and in Chapter 4 we will suggest that the tonic level of dopamine represents the net reward rate, thereby offering a normative account of the global influence of dopamine on response vigor. But now, with this interpretation of the model in mind, let us examine its behavioral results.

2.3.2 Fine-scale temporal behavior

Figure 2.7 depicts the behavior of our model on an RI30 schedule, roughly equivalent to that with which the rats in Figure 2.4 were trained[14]. The results show that the model rat chooses the correct actions at

[14]Due to technical limitations of the operant chambers, the data in Figure 2.4 were acquired in a VI (rather than RI) experiment, in which the interval distribution was not exponential. As a result, the hazard function for reward baiting was not flat, and the

the different states, nosepoking after a reward is received, and leverpressing thereafter. Furthermore, in accordance with the behavior displayed by the rats, the model's leverpress rate is constant over time, bar a pause for consumption. In fact, the model rat's behavior is (perhaps not surprisingly) more optimal than that of the experimental rats, as it shows little excessive nosepoking when a reward is not available (several possible reasons for the sub-optimal nosepokes shown by experimental rats will be discussed in section 2.4 below).

Figure 2.7c shows that the distribution of latencies chosen by the simulated rat for each of the different actions, is roughly Gamma shaped. More revealing are the latencies for these actions only in the state in which no food is available (as was measured for leverpresses and nosepokes for the experimental rats). These can be directly derived from the differential Q-values of each (a,τ) pair in this state, which are shown in panel (d). The optimal Q-values for each action are smooth as a function of τ, low for short latencies (because of the high vigor cost associated with rapid responding), relatively low for long latencies (due to the higher opportunity cost of slow behavior), and highest for some intermediate latency τ^* which optimally balances the tradeoff between these. The Q-values further show that at the state in which no food is available, leverpressing (blue) is more valuable than either 'Other' (green) or nosepoking (red). In dashed lines below, the soft-maxed Q-values are, in fact, the probability to choose each (a,τ) pair in this state $S = \{a_{prev} = \text{LP}, i_r = 0\}$.

Our modeling choices (specifically, the form we chose for the vigor cost) are supported by the fact that, in accord with the empirical data, the optimal Q-values result in latency distributions that are well fit by a Gamma distribution. By comparing the soft-max distribution of $Q(\cdot,\cdot,\tau)$ for a certain action and state, as a function of τ alone, to a Gamma distribution with an inverse rate parameter $\theta > 0$ and a shape parameter $\alpha > 0$

$$p(\tau) = \frac{e^{\beta Q(\cdot,\cdot,\tau)}}{\sum_\tau e^{\beta Q(\cdot,\cdot,\tau)}} = \frac{\tau^{\alpha-1} e^{-\tau/\theta}}{\theta^\alpha \Gamma(\alpha)} \tag{2.16}$$

we can derive the form of $Q(\cdot,\cdot,\tau)$ which would result in a perfect fit:

$$Q(\cdot,\cdot,\tau) = \frac{1}{\beta}[(\alpha-1)\ln\tau - \frac{\tau}{\theta} + const]. \tag{2.17}$$

Figure 2.8 shows that such a difference between a log and a linear function of τ approximates the optimal Q-values fairly well, at least in the range of latencies that have a reasonable probability of selection. To summarize, the optimal Q-values in the model give rise to behavior in which appropriate actions are performed at different states, and in which, as in the behavior exhibited by rats, inter-response times are approximately Gamma-distributed.

---

probability of baiting increased with time since the previous reward. Nevertheless, it seems from the rats' behavior that their responding was not affected by this specific aspect of the schedule. Two reasons to believe this are, first, the rate of leverpressing did not increase as a function of time from the previous reward, and second, behavior on two different VI30 schedules with different baiting hazard functions was indistinguishable. This data are thus very likely similar to that from RI experiments, however, we do not have access to RI data at this point, which would allow the direct comparison.

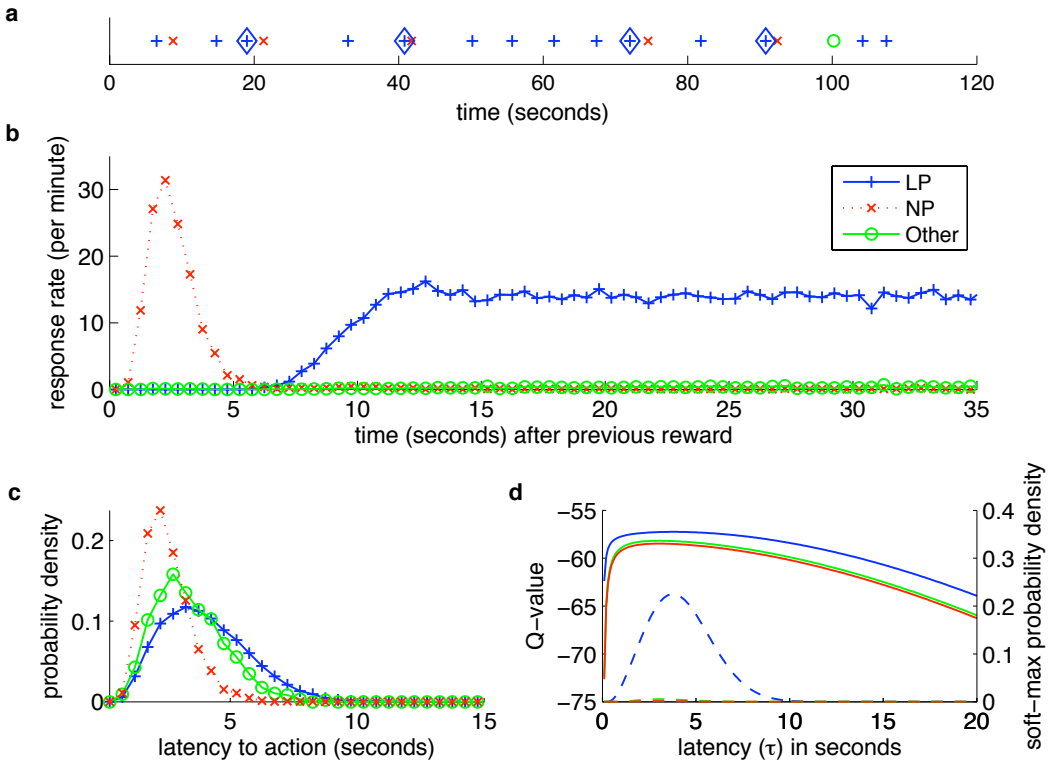

Figure 2.7: Data generated by the model captures the essence of the behavioral data (compare to Figure 2.4). Depicted are leverpress (blue; pluses), 'Other' (green; circles) and nosepoke (dotted red; crosses) responses on an RI30 schedule of reinforcement. **a.** Two minute excerpt from a session: a time-line of the rat's actions at the time of their completion (at the end of their chosen latencies). Diamonds surround leverpresses that resulted in food delivery. **b.** Response rates as a function of time since the previous rewarded leverpress. The simulated data show a rapid nosepoke response, followed by leverpressing at a constant rate (after some consumption time). **c.** As in the experimental data, inter-response times (response latencies) for each of the chosen actions are roughly Gamma distributed. **d.** The provenance of these distributions can be seen in the optimal Q-values for the (a,τ) pairs at $S = \{a_{prev} = \text{LP}, i_r = 0\}$ (solid lines; left axis). Note that the fact that the Q-values are negative is of no consequence, as these are *relative* values, ie, compared to the value of the just-rewarded state S_0 (which is 0), and only their values relative to each other are important. When filtered through a soft-max function, the probability of choosing the different latencies is roughly Gamma-shaped (dashed lines; right axis) for each of the actions. Data in (b,c) are averaged over 10,000 trials.

2.3.3 Ratio versus interval schedules

We can now compare the performance of the model in ratio schedules, as opposed to interval schedules. Recall that the only difference between random interval and random ratio schedules, is that whereas in the former the baiting probability is constant over time, in the latter, the rewarding probability per response is constant over time. That animals elect to respond at different rates in yoked ratio and interval schedules despite similarly distributed reinforcements from the environment, has been considered somewhat of a puzzle in the psychological literature, and although several explanations have been proposed, a consensus as to the underlying reason for this difference has not been reached (Dawson & Dickinson, 1990). Note that

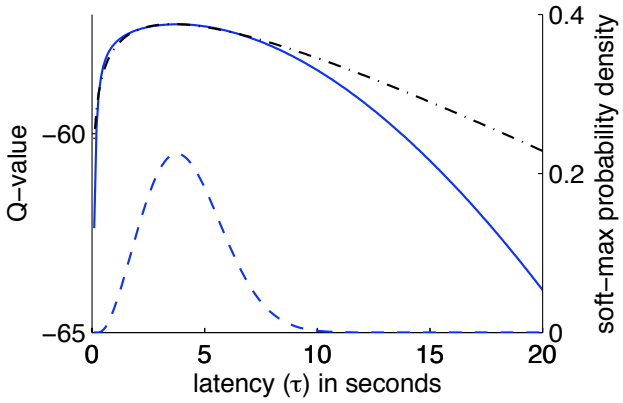

Figure 2.8: $Q^*(S,a,\tau)$ values (blue solid; left axis) and soft-maxed Q-values (blue dashed; right axis) for the lever-press response after a (nonrewarded) leverpress, as a function of response time, in an RI30 schedule. In dash-dot black are Q-values predicted from the Gamma fit to the IRT distribution, for comparison.

this phenomenon immediately rules out a simple function of the rate of reinforcement (for instance, that described by the "Matching Law") as the determinant of response rates in all schedules.

Simulating yoked interval and ratio schedules in our model shows that, as seen behaviorally, the model's response rates on ratio schedules are considerably higher than those on yoked interval schedules. Figure 2.9 illustrates this result for different RR schedules, and their yoked interval counterparts (in which the intervals between programmed reinforcements were matched to the intervals experienced on the ratio schedule). These differences in responding are fundamental characteristics of the schedules, and are not sensitive to the specific setting of the model parameters. Analysis of the model suggests that this phenomenon results from the fact that the optimal cost/benefit tradeoff that determines response vigor is qualitatively different for the two schedule types.

In the average reward model, and given the optimal net reward rate $\bar{R}^*$, we can analytically solve for the optimal latency to leverpress in a certain state S by finding the zero of the derivative of $Q(S,\mathrm{LP},\tau)$ with respect to τ. In random- or fixed-ratio schedules, this is especially straightforward as the vigor cost and the opportunity cost terms are the only ones depending on the chosen latency (notably, the transition probabilities to the subsequent rewarded or unrewarded states are independent of the action latency), giving

$$\frac{\partial Q_{ratio}(S,\mathrm{LP},\tau)}{\partial\tau} = \frac{C_v(\mathrm{LP},a_{prev})}{\tau^2} - \bar{R}^*, \tag{2.18}$$

and thus

$$\tau^*_{ratio} = \sqrt{\frac{C_v(\mathrm{LP},a_{prev})}{\bar{R}^*}}. \tag{2.19}$$

Importantly, this is true not only for leverpressing, but also for all other actions — the optimal latency for each action depends only on the vigor cost of emitting it, and the net rate of rewards.

In contrast, in random interval schedules the probability of transitioning to a rewarded state is $1-e^{-\tau/t}$ (where t is the schedule interval), which is dependent on τ. This is because the longer the latency to leverpress, the higher is the probability that the reward has become baited in this time (after the leverpress the process resets as a result of the memoryless distribution of intervals). Denote the instantaneous rate of

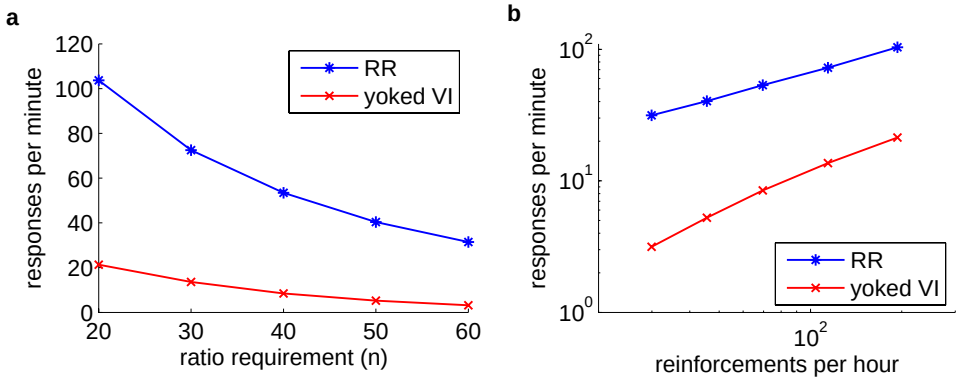

Figure 2.9: Comparison of performance on random ratio and random interval schedules of reinforcement. **a.** Steady state leverpress response rates on different RRn schedules (in which each leverpress was rewarded with probability $1/n$), and on corresponding yoked interval schedules, in each of which the baiting intervals corresponded to the times in which rewards were obtained on the ratio schedule. Value iteration for a yoked interval schedule was based on the mean inter-reward-interval obtained in the ratio schedule, and the behavioral data were generated with the exact sequence of yoked intervals. In all cases data were averaged over 6000 trials. Although the rewards were programmed at similar rates, interval schedule responding is markedly slower than ratio schedule responding (cf. Figure 2.2). Using the inverse latency to leverpress rather than the steady-state rate of leverpressing gives results that deviate only slightly from these, and only for long interval schedules, due to the effect of choices of $(a = \text{Other})$ on the measured leverpress rate. **b.** The same data plotted in log-log coordinates as a function of the obtained reward rate (the leftmost data point now corresponds to RR60 in which the reward rate was lowest, and the rightmost to RR20). The relationship between response rate and reward rate is relatively linear (on a linear as well as a log scale), in agreement with behavioral data (cf. the low reward rates in figures 10, 11 in Baum, 1993).

baiting $\lambda = 1/t$. For short enough leverpress latencies, λ approximates the derivative of the probability of transitioning to the rewarded state with respect to τ. Thus we can write

$$\frac{\partial Q_{interval}(S,\mathrm{LP},\tau)}{\partial \tau} \approx \frac{C_v(\mathrm{LP},a_{prev})}{\tau^2} + \lambda \cdot [V(S_r) - V(S_{nr})] - \bar{R}^* \tag{2.20}$$

where S_r is the rewarded leverpress state $\{t_{LP} = 0, a_{prev} = LP, i_r = 1\}$ and S_{nr} is the non-rewarded leverpress state $\{t_{LP} = 0, a_{prev} = LP, i_r = 0\}$. This gives the optimal latency to leverpress in interval schedules as

$$\tau^*_{interval} \approx \sqrt{\frac{C_v(\mathrm{LP},a_{prev})}{\bar{R}^* - \lambda \cdot [V(S_r) - V(S_{nr})]}}. \tag{2.21}$$

As $V(S_r) > V(S_{nr})$ and $\lambda > 0$, the denominator in $\tau^*_{interval}$ is strictly smaller than that in τ^*_{ratio}, explaining why the optimal latency is *always longer* (and consequently the response rate is lower) in interval schedules, as compared to ratio schedules. Intuitively, since longer IRTs increase the probability of reward per press in interval schedules but not in ratio schedules, the optimal leverpressing rate is lower in the former than in the latter. Note that in both schedules the optimal response rates show an inverse relationship to the net rate of rewards. The important implications of this result will be explored in detail in Chapters 3 and 4.

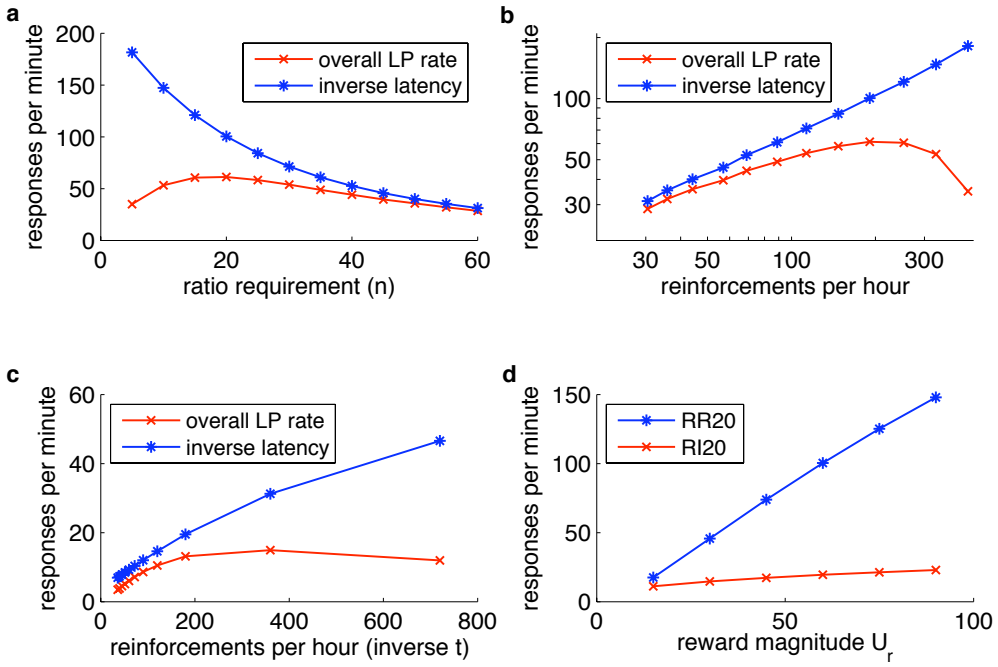

Figure 2.10: Raw leverpress response rates (red; measured as number of responses per minute in a 30 minutes session) and direct leverpress rates per minute (blue; measured as inverse of the mean chosen latency of leverpress actions) in different random ratio (RRn) schedules **(a,b)**, and random interval (RIt) schedules **(c)**, as a function of either the schedule parameter or the inverse interval between rewards (the reinforcement rate). Indirect measures show a complicated hyperbolic relationship with a decline in rates for high reinforcement rates, but a direct measure of leverpress rate, uncontaminated by issues such as time spent consuming the reward, shows a more straightforward monotonic power-law relationship (see text for details and analysis). **d.** The relationship between reinforcer magnitude U_r and leverpress rate on an RR20 (blue) or RI20 (red) schedule is linear.

2.3.4 Effects of reward magnitude and schedule parameter on vigor

In addition to their sensitivity to the type of reinforcement schedule, response rates on free-operant tasks are known to be influenced by the quality of the reward (its magnitude and/or concentration, Keesey & Kling, 1961; Bradshaw et al., 1978, 1981; Reed & Wright, 1988) and the schedule parameter or frequency of reinforcement (Baum, 1993). As discussed in Section 2.1.2, the hyperbolic saturating relationship between response rate and reinforcement rate has frequently been described in terms of the "Matching Law" for one instrumental action (equation 2.3). However, that response rates saturate (or even decrease) at high reinforcement rates is perhaps not a fundamental characteristic of response rate selection, but merely a result of the way response rates are typically measured: as the total number of responses in a session of a fixed length. These measurements do not account separately for the total time spent consuming the earned reinforcers, which increases as the reinforcement rate increases leaving less time for the instrumental response (Baum, 1993; Foster et al., 1997). As in our simulated experiments it is easy to obtain direct measures of the rate of responding on each action uncontaminated by competing responses (eg, by looking at inverse latencies), we can use the model to investigate the relationship between response rate and reinforcement rate or magnitude.

Figure 2.10(a,b) shows the relationship between leverpress rate and schedule parameter, for different RRn schedules. In red is the commonly used measure of raw number of responses per minute throughout the session. As the schedule requirement becomes larger (and the reinforcement rate lower), response rates diminish. However, for low schedule requirements (and high reinforcement rates) response rates are also low, as is seen experimentally (Baum, 1993). In blue is a cleaner measure of leverpress rate, namely, the average inverse latency to leverpress, which reveals a monotonic relationship between schedule parameter and response rate. Foster et al. (1997) show a similar result after correcting for consumption time: response rates are strictly diminishing for higher 'prices' of a reinforcer (as would be expected from a rational economic agent), at least in an open economy in which animals also receive food outside the experimental session.

Figure 2.10a suggests an inverse relationship between response rate and schedule parameter, which we can confirm analytically. In RRn or FRn schedules, the we can analytically compute the optimal net reward rate directly from the (trivially known) optimal policy — the net reward rate is the net gain from one cycle of reinforcement (the utility of one reward minus the costs of achieving it) divided by the time the cycle takes:

$$\bar{R}^* = \frac{U_r - \left[C_u(\mathrm{NP},\mathrm{LP}) + \frac{C_v(\mathrm{NP},\mathrm{LP})}{\tau^*_{\mathrm{NP},\mathrm{LP}}}\right] - (n-1)\left[C_u(\mathrm{LP},\mathrm{LP}) + \frac{C_v(\mathrm{LP},\mathrm{LP})}{\tau^*_{\mathrm{LP},\mathrm{LP}}}\right] - \left[C_u(\mathrm{LP},\mathrm{NP}) + \frac{C_v(\mathrm{LP},\mathrm{NP})}{\tau^*_{\mathrm{LP},\mathrm{NP}}}\right]}{\tau^*_{\mathrm{NP},\mathrm{LP}} + (n-1)\tau^*_{\mathrm{LP},\mathrm{LP}} + \tau^*_{\mathrm{LP},\mathrm{NP}} + t_{eat}}. \quad (2.22)$$

For illustration purposes, let us make the simplifying assumptions that all unit and vigor cost constants are equal (and thus all actions also have the same optimal latency) and that $t_{eat} = 0$, which gives[15]

$$\bar{R}^* = \frac{\frac{U_r}{n+1} - C_u - \frac{C_v}{\tau^*}}{\tau^*}. \quad (2.23)$$

Now, substituting the optimal latency $\tau^* = \sqrt{C_v/\bar{R}^*}$ (from equation 2.19), we get

$$\sqrt{\bar{R}^*} = \frac{1}{2\sqrt{C_v}}\left(\frac{U_r}{n+1} - C_u\right), \quad (2.24)$$

and so the optimal response rate $1/\tau^* = \sqrt{\bar{R}^*/C_v}$ is

$$\frac{1}{\tau^*} = \frac{1}{2C_v}\left(\frac{U_r - (n+1)C_u}{n+1}\right). \quad (2.25)$$

Thus, in ratio schedules the uncontaminated measure of leverpress rate is inversely related to n, as suggested by Figure 2.10a. It is also inversely related to the vigor cost constant C_v, and linearly related to the magnitude of reward U_r (as shown in Figure 2.10d). Because the optimal latency τ^* is proportional to n, the interval between rewards is proportional to n^2 (n leverpresses at latency $\propto n$ to achieve a reward), which means that the relationship between response rate and reinforcement rate in ratio schedules is actually not a hyperbolic one, but rather a power law. This is illustrated in Figure 2.10b where it is manifest as a linear relationship in

[15]This analysis is also possible (and its results are qualitatively similar) without these simplifying assumptions, however it involves solving a quadratic equation for $\sqrt{\bar{R}}$ when $t_{eat} > 0$. Note, also, that the figures illustrate the mean latency as averaged over the soft-max policy, while the analysis is in terms of the optimal latency (the peak of the soft-max distribution, rather than its mean). Simulated results for optimal latencies are qualitatively similar, but provide a less straightforward comparison to experimental data.

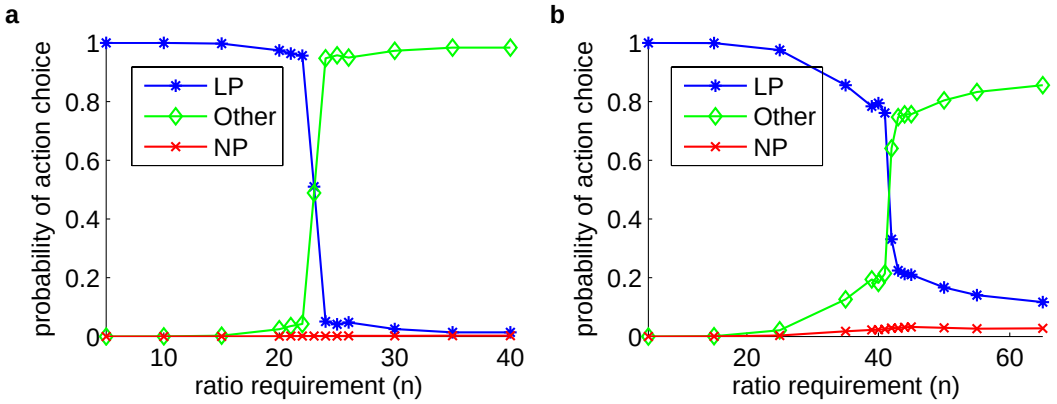

Figure 2.11: Probability of choosing the different available actions in an extinction test (using optimal Q-values and without further learning), for different ratio requirements. **a.** When $U_r = 20$ and $C_u(\text{Other}) = -0.6$, leverpressing abruptly stops at schedules higher than $n = 23$, illustrating 'ratio strain'. **b.** Ratio strain is dependent on the relative values of $C_u(\text{Other})$ and U_r. When $C_u(\text{Other}) = -0.3$, the breaking point is still abrupt, but occurs at a higher ratio requirement.

log coordinates (for comparison with Baum, 1993).

For interval schedules it is no longer straightforward to write a closed form solution for the optimal average reward rate. However, the 'molar' measure of overall response rate depicted in red in Figure 2.10c replicates the well-documented hyperbolic relationship between response rate and reinforcement rate on interval schedules (cf. Figure 2.3b), albeit with response rates diminishing at very high reinforcement rates as a result of consumption time, which has also been shown experimentally (Baum, 1993). As expected, the 'molecular' measure of inverse latency to leverpress (which is not contaminated by competition with other actions, or confounded with consumption time) shows a monotonic relationship, again suggesting a power law. Similar to the case of ratio schedules, the relationship between reward magnitude and response rate is linear for interval schedules (Figure 2.10d).

In sum, the model predicts that, when measured carefully, IRTs between consecutive leverpresses should show a power-law increase as the ratio or interval schedule requirement increases, and should decrease linearly as the magnitude of reward is increased. This is in accord with the few studies in which such measurements were taken (eg. Baum, 1993; Foster et al., 1997), but a full investigation of the different relationships for both schedule types has yet to be undertaken.

2.3.5 Breaking-point analysis in ratio schedules

Another characteristic of the difference between ratio and interval schedules is replicated by the model, namely, the susceptibility of ratio (but not interval) schedules to what is called 'ratio strain' (Baum, 1993). In this, responding is maintained by increasingly demanding ratio schedules, up to some critical point where the schedule behavior completely breaks down and responding stops (for instance, 'progressive ratio' sched-

ules increase the ratio requirement after each earned reward, and use this breaking-point as a measure of strength of responding). This phenomenon is not seen in interval schedules, even very long ones. Baum (1993) suggests an explanation in terms of optimality models that focus on the tradeoff between the scheduled reward for leverpressing and alternative intrinsic rewards for other actions. The idea is that schedule responding stops at the point in which the most rewarding action is no longer the instrumental action.

Indeed, in the model leverpressing competes with 'Other', which also leads to a (small) intrinsic reward. Both actions are essentially concurrently reinforced on ratio schedules, with 'Other' rewarding on FR1 with a small utility, namely, $-C_u(\text{Other})$. The model shows an abrupt breaking-point of leverpress responding, in favor of exclusive responding on 'Other', at the ratio requirement for which the net reward per press, $\left[\frac{U_r}{n+1} - C_u(\text{LP}) - \frac{C_v(\text{LP,LP})}{\tau^*_{\text{LP}}}\right]$, drops below the net reward for choosing 'Other', $\left[-C_u(\text{Other}) - \frac{C_v(\text{Other,Other})}{\tau^*_{\text{Other}}}\right]$, as illustrated in figure 2.11 for two different values of $C_u(\text{Other})$. Such a breaking down of instrumental responding does not occur in interval schedules, because in those leverpressing is always worthwhile: even in long interval schedules occasional leverpresses can harvest baited rewards, which means that infrequent leverpressing will have a high value and instrumental responding will not cease completely.

2.3.6 Concurrent interval schedules

We simulated concurrent interval (RI-RI) schedules by allowing two different leverpress actions to earn rewards on independent random interval schedules, and augmenting the state space so as to measure time from the last lever press on each lever. Recall that behavior on such schedules frequently conforms to the "Matching Law" (Herrnstein, 1970), but this is dependent on the existence of a penalty for switching from one lever to another (Herrnstein, 1970; Williams, 1994). Without this, and assuming low costs for each action, simple policies such as win-stay-lose-shift or alternating between the two levers, which are natural foraging strategies for many animal species, would be almost optimal (Sugrue et al., 2004). A penalty for frequent switching is sometimes explicitly introduced, but can often be implicit because switching from one response option to another requires time- and energy-costly travel due to the physical dimensions of the operant chamber. This latter type of change-over penalty can be naturally simulated in our model as a high cost for transitioning from one lever to the other.

In concurrent interval schedules there is no obvious recurrent state, so optimal Q-values were obtained using value iteration with time-discounted rather than average reward RL. Data were then simulated by generating actions from fixed Q-values in extinction conditions (ie, with no programmed rewards and no further learning). The resulting behavior approximated (in the soft-max case), or was exactly (when choosing according to the maximal Qs) the deterministic policy of choosing the richer option N times, followed by one choice of the leaner option, and back to the richer option, a policy that has previously been shown to be the optimal one for this situation (Houston & McNamara, 1981). In the model's behavior, within each cycle the N high-rewarding option leverpresses are executed at an increasing rate, as a result of the growing probability of reward on the other lever. Interestingly, animals responding on concurrent interval schedules actually

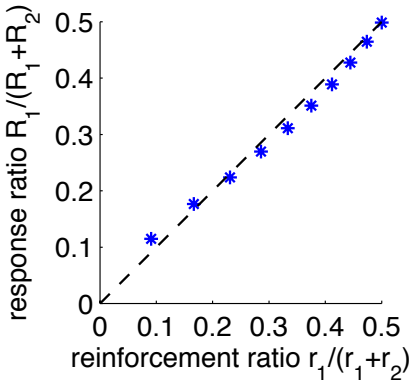

Figure 2.12: Under certain settings of the parameters (see text), the model's choices on a concurrent RI-RI schedule roughly conform to the "Matching Law", with the ratio of responses on each of the two levers following the ratio of their reinforcement rates. Blue stars: simulation result; dashed line: perfect matching. Here $C_v(\mathrm{LP1},\mathrm{LP2}) = C_v(\mathrm{LP2},\mathrm{LP1}) = 1$, $\beta = 0.05$ and $t_{eat} = 1$.

behave suboptimally as they don't seem to take the time since pressing the alternative lever into account (eg. Cleveland, 1999). See Daw and Touretzky (2001) for a discussion of possible causes for this deviation from optimality.

The "Matching Law" does not specify the order of choices of the two levers, but only the 'molar' ratio of response allocation (or time allocation) between them. Whether the deterministic policy or the soft-max results conform globally to the matching ratio or not, depends on the model parameters, and notably the vigor cost per lever-press and the change-over cost. The latter determines the time of the switch (what probability of reward on the lower-rewarding lever will justify the switch cost, with a higher cost delaying the switch), and the former determines the size of N (with a lower cost increasing N), that is, how fast responses on the high-rewarding lever will be performed, which in turn determines how many responses will be executed prior to the time of the switch. As has been also shown behaviorally (Staddon, 1992; Williams, 1994), manipulating the change-over cost can cause both undermatching or bias (in which the more frequently rewarding option is chosen less frequently than predicted by the ratio or reinforcements), and overmatching (in which the more rewarding option is chosen more frequently than predicted by the "Matching Law", and the policy is more 'maximizing' than matching). Figure 2.12 shows that with a proper setting of the change-over cost, indeed the response rates chosen by the model rat match the reinforcement rates of the two levers for a variety of interval schedules. Similar results are obtained when using either of the two commonly used behavioral measures, time spent on each lever (measured here as the cumulative latencies to actions on each of the levers), or number of actions on each lever. However, the implications of the parameter sensitivity of this result to the non-generality of matching behavior, is in line with previously noted empirical and theoretical challenges (Staddon, 1992; Williams, 1994, see also the Discussion below).

2.3.7 The effect of free rewards

One last and critical test for the model, is its behavior when rewards that are not contingent on a specific instrumental action are available, in addition to those earned instrumentally. This is a test case due to theoretical rather than experimental reasons (indeed, to the best of our knowledge, the experiment we have

in mind has yet to be done[16]). The theoretical conundrum is that by giving non-contingent rewards we can directly manipulate the net reward rate, however, if these rewards are really 'free' (and assuming they do not require a specific harvesting action, for instance, free brain stimulation rewards), this should, theoretically, have no effect on the animal's (optimal) policy. The special role of $\bar{R}$ in determining response vigor in our model (see Section 2.3.3) seems to imply that the model's behavior will be wrongly affected by free rewards.

To test this, we modeled free rewards which do not necessitate a consumption response as a constant low probability of reward ε per unit time. In effect, this adds $\tau \cdot \varepsilon$ to the Q-value of each action:

$$Q^*(S,a,\tau) = P_r^{a,\tau}(S) \cdot U_r + \tau \cdot \varepsilon - C_u(a, a_{prev}) - \frac{C_v(a, a_{prev})}{\tau} - \tau \cdot \bar{R}^* + \sum_{S' \in \mathcal{S}} \mathcal{T}_{S \to S'}^{a,\tau} V^*(S'). \tag{2.26}$$

Although this manipulation increased the optimal net reward rate $\bar{R}^*$ by ε, this had no effect on optimal action selection or response rates. This is because the increase in $\bar{R}^*$ affects the Q-values through the term $(-\tau \cdot \bar{R}^*)$, which is exactly counteracted by the new term $(\tau \cdot \varepsilon)$. The optimal ratio-schedule latency is thus

$$\tau_{ratio}^* = \sqrt{\frac{C_v}{\bar{R}^* - \varepsilon}} \tag{2.27}$$

giving the same response rate as in the case where there were no free rewards. Interval schedule responding is similarly unaffected.

2.4 Discussion: A new theory of free-operant responding

In this chapter, I have presented a computational model of optimal decision making in free-operant tasks, which incorporates the important, but often neglected, aspect of response vigor into a reinforcement learning framework. The model brings the computational machinery and neural grounding of RL models fully into contact with the vast reservoir of data from free-operant tasks. More importantly, it provides a normative basis for understanding the determinants of response rates, the most common dependent variable measured and tested in over half a century of behavioral psychology.

According to our model, response rates are determined as the optimal tradeoff between the costs of vigorous responding, and the benefits dependent on these instrumental responses. Individual action latencies are selected so as to exactly balance the (long-term) benefit minus the (immediate) cost of responding and the

[16] Dickinson and Balleine often report a 'contingency degradation' manipulation (eg. Balleine & Dickinson, 1998, 2000; Corbit et al., 2001; de Borchgrave et al., 2002; Yin et al., 2005). In this, every second in which a lever is pressed is rewarded with probability x, but for every second in which the lever is *not* pressed free rewards are given with the same probability. This is different from giving 'free' rewards, as in 'contingency degradation' rewards are explicitly contingent on not responding (which has the effect of reducing lever-pressing on the contingency-degraded lever), while in the case we discuss the free rewards are given *in addition* to those earned by lever pressing. Another closely related result from a scenario in which two different actions are rewarded with two different types of reward, is that adding non-contingent rewards of one type degrades performance on the action which earns a similar reward, in favor of the alternative response. In this case, specific satiety on each of the outcomes (which we did not model here) may be affecting response choices.

opportunity cost of remaining idle. The resulting 'molar' response rates are globally optimal in that they achieve the highest possible net reward rate, given the constraints of the reinforcement schedule.

The two sub-choices of which action to perform, and at what latency (or instantaneous rate) to perform it, are influenced by different aspects of the task (Niv, 2007). The choice action depends on the utility of the rewards potentially available for each of the actions, the probability that the action will indeed be rewarded, and the effort cost of performing the action. So pressing a lever that leads to food reward with 20% chance is preferable to an action that leads to the same outcome but with only 10% chance, and pressing an easily accessible lever is preferable to pulling a chain that necessitates much effort to jump and reach it. The optimal choice of how fast to perform the chosen action is determined by an altogether different cost/benefit tradeoff. On the one hand, faster performance is more costly. On the other hand, completing the chosen action slowly delays not only the availability of the possible reward for this action, but more importantly, it delays all future actions and rewards. The average reward RL framework is illuminating in that it makes explicit this tradeoff between the vigor cost and the expected net reward rate, which exactly quantifies the opportunity cost of wasted time.

Our model thus highlights the importance of the net rate of reward in determining optimal response rates, and shows that *higher reward rates should normatively be associated with faster responding, and lower rates with slower responding*. This key result will provide the basis for the following two chapters. In Chapter 3, we will ask: how do motivational states influence behavior? We will argue that motivational states determine the utility of different goal objects, for instance, a state of 'hunger' implies higher utility for food. As a result, hunger affects not only the worth of actions leading to food, but additionally, it increases the expected net reward rate in situations in which food rewards are typical. This, according to our model, should affect all response rates. Thus hunger should indeed enhance leverpressing for food rewards, but should also speed up other unrelated actions. This provides a normative basis for global 'drive'-like motivational effects, and specifically, for the effects of motivation on outcome-insensitive habitual behavior. In Chapter 4, we will discuss the neural basis of the control of response vigor. Extending existing ideas about the role of phasic dopamine signals in learning the values of actions, we will suggest that the important but often neglected *tonic level of dopamine* represents the expected net rate of reward. This idea explains why dopamine plays a critical role in determining the vigor of responding, and offers a normative account of a wealth of experiments showing that dopamine influences response rates. Furthermore, it provides a route by which dopamine could mediate the effects of motivation on vigor.

Note that although action selection in the model necessitates only locally available information, it is not at all myopic. First, the worth of each response option is computed in terms of state values that provide information about long term expected reward. Second, the net reward rate is a global measure carrying long-run policy-weighted information. Thus local action selection can be optimal in the long run, and is not prone to local maximizing at the expense of future benefits. This implementational locality (ie, the fact that long-run values can be learned and utilized using only local information; see Chapter 5), which nevertheless is guaranteed to achieve global optimality, is a fundamental characteristic of RL methods (Sutton & Barto, 1998).

2.4.1 Molar vs molecular explanations

Psychological theories of instrumental responding are frequently classified as being either at the 'molar' or the 'molecular' level. Consider, for instance, theories that aim to explain the difference between response rates in ratio versus interval schedules (Baum, 1989). One line of explanation appeals to the fact that interval schedules differentially reinforce long inter-response-times (IRTs), due to the higher probability that a reward has been baited in the time since the previous response, while ratio schedules reinforce all IRTs equally. Thus, through a 'response shaping' Skinnerian or Thorndikean 'Law of Effect'-type view, the former can be expected to generate responding with longer IRTs, that is, at a slower overall rate (Peele et al., 1984; Cole, 1999; Reed et al., 2000). In contrast to this 'molecular' line of reasoning, the other main line of argument is 'molar' in nature, and ascribes the difference in response rates to the difference in global feedback functions relating reward rate to response rate (McDowell & Wixted, 1986; Dawson & Dickinson, 1990; Baum, 1993). Ratio schedules provide a linear relationship between response rate and reward rate (the faster the responding, the higher the reward rate), while in interval schedules this relationship is nonlinear (saturating). In fact, in interval schedules, from some response rate and above, the reward rate is almost independent of the response rate, because the schedule sets an upper limit on the rate of rewards. Sensitivity to this difference in global feedback functions can thus be the reason for the different behavioral policies evident in the two types of schedules (Dawson & Dickinson, 1990; Baum, 1993).

As already suggested above, the model we present resides at both levels at once. For instance, our explanation of the discrepancy between response rates on yoked ratio and interval schedules can be considered a molecular one: the optimal choice of individual latencies is different in the two schedules. However, this is not the result of a local, molecular 'reinforcing' of some latencies as opposed to others. The primary determinant of why a particular latency is the optimal one is the net reward rate, which is a 'molar' measure. Furthermore, through the expected value of subsequent states, the optimal latency takes into account the constraints of the schedule, that is, its feedback function, which is a characteristic of 'molar' explanations of the ratio/interval paradox. Then again, the process by which the schedule-dependent responding is learned in the first place can be wholly local (see Chapter 5), again disguising the explanation as a 'molecular' one. The reason why the local and global cannot be teased apart in the model, is that truly optimal action selection should indeed combine long-term quantities in determining short-term actions.

However, one aspect to which our model points a directed spotlight, is the importance of using 'molecular' measurements of response vigor, and abandoning gross measurements such as overall responses in a session that confound reward rate (and thus time spent consuming) with the measurement of the vigor of the response in question. It seems that many a confusion in the literature might have arisen from the use of such contaminated measures. The concerted efforts to explain anomalies such as bitonic response rate curves can be abandoned, once precise measurements reveal that this is indeed an epiphenomenon (Baum, 1993). In this respect, response rates measured in extinction may be regarded as a cleaner measurement. Unfortunately, these are not steady state measurements, and the extinction-learning process may itself be a factor interacting with the measured rates.

2.4.2 Relationship to previous models

To the best of our knowledge, our model is the first to specifically target free-operant behavior using a normative reinforcement learning framework. Previous models aimed at similar data were for the most part descriptive rather than normative, while reinforcement learning models of conditioning dealt exclusively with action selection at predefined discrete timepoints, ignoring the question of response rate. In the following, I discuss the relationship between several previous models and the model we have presented here.

The "Matching Law"

The most venerable and influential account of free operant behavior is Herrnstein's "Matching Law" (Herrnstein, 1970). Originally a description of responding in concurrent interval schedules (Herrnstein, 1961), Herrnstein later suggested the matching relationship as a model of all free-operant behavior (eg. Herrnstein, 1990), and he and other researchers extended it to the case of only one instrumental action, to dealing with options that differ in reward magnitudes and not only reinforcement rates, to different delays in reinforcement, etc. (Baum, 1993; Williams, 1994). For instance, to extend the "Matching Law" to the case of one instrumental action, it was assumed that even in this case animals make choices between different actions, specifically, between the instrumental action and other intrinsically rewarding actions (Herrnstein, 1970), and that these alternative actions compete with the instrumental action for time allocation[17].

As a descriptive theory of responding, the "Matching Law" is fundamentally a molar theory of response rates rather than a theory of individual choice. In fact, the "Matching Law" does not correspond to any specific policy — many different policies specifying when to switch between options and how long to stay on each, can yield the same matching relationship. Melioration (Herrnstein & Prelec, 1991) is a dynamic model of individual choice that was suggested as the 'molecular' mechanism by which the "Matching Law" is implemented. By rearranging the matching relationship (equation 2.1) as

$$\frac{R(r_1)}{r_1} = \frac{R(r_2)}{r_2} \tag{2.28}$$

we can see that matching response rates to reinforcement rates equates the return (reinforcements per response) on the different options. Melioration keeps track of the returns of the two options, and chooses the option providing the highest return at each point in time. This naturally produces the matching relationship as the policy's fixed point. Surprisingly, however, melioration was never specified precisely at the 'molecular' level of how choice probabilities are updated as a result of differences in return, and how returns are tracked based on experience[18] (Herrnstein, 1991; Staddon, 1992). As a result, the predictions of the model

[17] The assumption of a fixed total amount of time for which different response options compete is reasonable. However, its extension to the formulation of the matching law in terms of response rates is questionable, as there it was assumed that the total rate of responding on all actions is fixed.

[18] It seems that Herrnstein 'discovered' rather than 'invented' the "Matching Law": he posed it as a description of behavior and not as a theory. Herrnstein (1991) notes that the reason melioration was not specified at the level of how subjects define and update

for individual response data were never explicitly stated and tested. Rather, both the "Matching Law" and melioration have been used mostly for 'molar' curve-fitting. So far, we have also presented a model of steady state responding, albeit one that makes precise predictions at the finer scale of sequences of choices between different response options. In Chapter 5, we will extend the model and specify it at the level of individual choices and updates. However, we are also at fault, as we also have yet to explore the critical fine-scale comparisons between the predictions of our model and empirical data.

At the level of normative theories, attempts to derive the matching relationship from higher principles have not seen much success (Williams, 1994). Equating returns has the flavor of an optimal allocation (or at least a locally optimal one), and indeed Staddon and Motheral (1978) showed that, providing that there is no change-over-delay (COD), this allocation maximizes reinforcement rate within a class of stochastic policies which probabilistically allocate responses to the two options. However, Houston and McNamara (1981) showed that the optimal policy in concurrent interval schedules is not a stochastic one. Rather, the optimal steady-state policy is a deterministic policy that stays on the richer arm for a fixed length of time and then switches momentarily to the leaner arm, retrieves any reinforcers that have become baited there, and immediately returns to the rich arm. Although this allocation of time or responses can approximate matching for some schedules and CODs, the matching relationship *per se* is not always optimal.

Herrnstein himself did not pose the "Matching Law" as normative. In fact, he specifically designed reinforcement schedules in which it is suboptimal to match the returns of the different payoff options, in order to pit the "Matching Law" against rational choice theories and prove the superiority of the former in describing empirical behavior (Herrnstein & Prelec, 1991; Herrnstein, 1990, 1991, 1997). In these schedules, the past allocation of responses to each of two options (eg. in a sliding window of 40 previous choices) determined the current reward that each option would deliver if chosen. Indeed, in many such cases subjects allocated their choices suboptimally (eg. Herrnstein, 1991). Our model, as a normative model that is guaranteed to find the globally optimal policy, cannot be expected to reproduce such behavior. However, note that in order to maintain the Markov property in such reinforcement schedules, the state representation must be augmented to include the relevant history of choices. It is reasonable to posit that animals and humans can not adequately represent a 40-long relevant history. Truncating the representation results in a partially-observable state space, in which case, like subjects' behavior, the model will be prone to local maxima. Egelman et al. (1998) showed that in the simplest case in which no history is represented, reinforcement learning will also settle on myopic matching of returns, rather than discover the globally optimal policy. Moreover, Herrnstein (1991) notes that, as the relevant behavior history is made shorter (ie, if rewards depended on the last 3-6 choices), subjects' response allocation approaches optimality. Thus, although in some cases behavior can be shown to be suboptimal, this does not in general invalidate normative models of choice.

That animals allocate responding to two differently rewarding options is understandable when considering concurrent interval schedules in particular. Because the schedules are independent, time spent on one sched-

response alternatives, is that "we do not know the learning algorithm that produces matching... We have examined the trial-by-trial choice patterns of our subjects and the only conclusion I can draw is that people appear to have many different ways to reach the matching point." (p.364)

ule counts as time elapsed on the other, so the longer the animal has been on one schedule, the more likely it is that a reward has been set up on the other one, and alternating between the schedules is worthwhile. In contrast, in ratio schedules each response has some probability of being rewarded, but the passage of time without responses does not change this probability. Some claims to the generality of the "Matching Law" were based on the observation that animals also match on concurrent VI-VR schedules, even when this leads to lower than optimal reinforcement rates (Heyman & Herrnstein, 1986), but here, as well, issues of state representation may be at stake. In concurrent ratio schedules, animals sometimes respond exclusively on the higher rewarding option (Herrnstein & Loveland, 1975), which is both the optimal policy and the one that satisfies the matching relationship (Herrnstein, 1970). However, in other experiments animals have been shown to allocate responses such as to match the ratio of the probabilities of reinforcement of the two options ("probability matching"; eg. Bitterman, 1965), which violates the matching relationship (Herrnstein, 1970). This suboptimal policy might be related to the use of soft-max rather than strict maximization: for certain settings of the inverse-temperature parameter β, choosing actions according to the Boltzmann (soft-max) distribution of their values (equation 2.14) will naturally lead to probability matching.

To summarize, the "Matching Law", while the most prominent model of free-operant instrumental behavior to date, cannot be considered a full model of individual action choice, or even a full description of a molar policy in concurrent schedules. Empirically it does not seem as if matching is as often the policy of choice as first believed (eg. Wearden & Burgess, 1982; Dallery & Soto, 2004; Soto et al., 2006, 2005), and although extensions to the basic model have been proposed (Baum, 1974), none is able to explain a wide variety of response rate data (Baum, 1993). Finally, in terms of normative models of behavior, the matching relationship is in many cases a suboptimal allocation of responding (Staddon, 1992; Williams, 1994), and in general can only be used as a description of behavior, not as a rationalization of it.

Killeen

Killeen and his colleagues have suggested a model of free-operant responding called 'Mathematical Principles of Reinforcement' (MPR; eg. Killeen, 1995; Killeen & Sitomer, 2003). This model is part of a framework of mathematical modeling of behavior, whose goal is also not a normative explanatory one. Rather, MPR aims to describe the basic laws of behavior from which complex behavioral patterns can be derived, much like the basic laws of mechanics and their relationship to explaining physical phenomena. Although mostly based on data from Pavlovian conditioning (eg. Killeen et al., 1978; Killeen & Sitomer, 2003), the framework does not explicitly distinguish between Pavlovian and instrumental behavior, and is intended to explain aspects of instrumental behavior as well (Killeen, 1998). Similar to the focus of this chapter, MPR deals mainly with steady-state behavior and not the dynamics of learning. Unfortunately, although MPR is a 'molecular' theory that makes detailed predictions regarding the effects of immediate prior history on action selection, it too, like melioration, has only been used to derive molar fits to average response curves.

MPR is based on three principles (Killeen, 1995, 1998; Killeen & Sitomer, 2003): (1) reinforcement causes unspecific arousal of behavior, (2) reinforcement causes outcome-specific motivation of reinforced responses, and (3) responding carries temporal and energetic costs. MPR further assumes the matching relationship as the fundamental determinant of action selection (as do many other theories, eg. Gibbon, 1992). The third of these principles we have also assumed; the second, in a simpler form (outcomes reinforce responses) is at the basis of reinforcement learning, and in its motivational sense can also be linked to our model (see Chapter 4); the first principle we have derived.

Killeen's first principle argues for a linear relationship between the overall rate of reinforcement and the level of *arousal*, which activates behavior in a general, diffuse way (Killeen et al., 1978). The idea is that each reinforcer activates behavior by some amount that depends on the reinforcer's magnitude and quality and the animal's deprivation level. This activation then decays exponentially with a fixed time constant. Killeen quantifies this exponential function by measuring the effect of single feedings on general activity of pigeons, and deriving the asymptotic arousal level that would result from slow exponential averaging over reinforcers, for different rates of reinforcement. Killeen (1998) furthermore claims that, even in learned situations, arousal is necessary for behavior. Evidence for this is that in avoidance experiments, even after many training sessions, response rates are low in the beginning of the session — presumably until arousal resulting from the aversive events accumulates. Many other behavioral phenomena, such as the reduction of responding in extinction, he similarly attributes to effects of arousal on responding, rather than to effects of learning. The level of arousal is related in MPR to both the probability and the rate of responding. However, due to the third principle, the relationship between arousal and instrumental response rate is not linear: responses demand some minimal execution time, and so one response blocks the emission of other responses and there is a ceiling on the overall response rate. From this, Killeen derives a hyperbolic relationship between response rate and reinforcement rate (Killeen, 1995, 1998), similar to that described by Herrnstein.

The arousal level postulated by Killeen is closely related to our notion of net reward rate and its general effect on response rates. However, some differences exist. At the algorithmic level[19], in Chapter 5 we propose that the net reward rate is computed online by exponential averaging of *reward prediction errors* rather than the rewards themselves (explaining for instance why conditioning can result in high levels of arousal at the beginning of a session, before any rewards have been received; see Chapter 4). At the implementational level, to translate arousal to response probability, Killeen assumes that at a specific level of arousal the probability of responding is constant. This implies an exponential distribution of IRTs (Killeen, 1998), which peaks at zero inter-response-time, at odds with our data. However, if we assume that execution times are random, and relate the model's IRTs to the interval between the end of one action and the beginning of another, this can be reconciled and Gamma-distributed latencies can be reconstructed. Despite the relationship between the models, the fundamental difference between our model and Killeen's is at the computational level: ours is a normative model that derives response patterns from optimality considerations, while Killeen describes behavior in terms of simpler behaviors and not in terms of its sources or reasons. However, that our results are consistent with the emphasis that Killeen puts on general arousal and its manifestations is encouraging.

[19] I am referring here to Marr's (1982) three levels of analysis: computational, algorithmic and implementational.

Staddon

Dragoi and Staddon (1999) also present a theory of instrumental behavior aimed at describing response rates. Their theory concentrates on the dynamics of learning and is based on multiple timescales: immediate reward expectation is compared to long-term reward expectation in order to determine response strength. The model consists of differential equations defining the dynamics of short-term and long-term associations and memory traces, which give rise to the short-term and long-term expectancies on which response strength is predicated. Whereas our model specifies individual responses, the dependent variable in Dragoi and Staddon's (1999) model is response rate, and individual responses are drawn at every tick of a clock with a probability defined by a soft-max over the strengths of potential responses. Thus although response strength dynamics are continuous, the model generates discretely timed behavior.

Through its multiple processes, their model replicates (but does little to explain) conditioning phenomena on several time-scales, including generalization, hyperbolic discounting, matching, positive and negative contrasts, the partial reinforcement extinction effect, and spontaneous recovery — many of which are beyond the scope of our model as they are not steady-state phenomena, or they relate to issues of state representation rather than value learning and the selection of response rates. However, the comparison between short and long term expectations is somewhat reminiscent of average reward reinforcement learning in which the values of actions are defined as the difference between their specific predicted benefits and the overall global benefit of the policy (the net rate of reward).

One notable exception to these descriptive models of free-operant responding is Staddon (2001), who suggests a normative theory that minimizes a cost function. In his model, like in ours, cost is proportional to response rate. However, the mathematical status of this model is uncertain, and its result — that response rate should be proportional (rather than inversely proportional) to schedule requirement — is in opposition to empirical data (and to our results). Staddon himself does not emphasize this model, as he does not believe that animals optimize anything in a literal sense, despite their often optimal behavior (Staddon, 2001).

Reinforcement learning

Finally, within normative models of animal decision making, reinforcement learning models of action selection are the most widely acclaimed (eg, Barto, 1995; Montague et al., 1995; Sutton & Barto, 1998; Dayan & Abbott, 2001; Daw et al., 2005). It is perhaps surprising, then, that despite the wealth of literature regarding response rates, these models have concentrated solely on action selection at discrete prespecified timepoints, saying nothing about the determinants of rate. Recently, McClure et al. (2003) made a first attempt at wedding the framework of RL models with response rates. In their model of a rat navigating a maze, actions are selected at prespecified timesteps, with repeated choices between running or keeping still resulting in an overall running rate. Our framework is fundamentally different from theirs, because by modeling the choice of latency, we directly optimize the rate of behavior. This allows the inclusion of vigor-related costs

that are absent in McClure et al's model (it is impossible to add vigor costs to a repetitive-discrete-choice model without violating the Markov assumption), as well as highlights the role of the net rate of reward in determining optimal response rates.

Another related framework of semi-Markov models of decision making, are RL models that chunk useful action schemas into temporally extended 'options' (Precup et al., 1998; Sutton et al., 1999). In these, the agent chooses between 'options' that involve sequences of discrete actions of varying lengths, and thus may take different times to complete. However, these models eschew an explicit choice of speed. We maintain that, regardless of whether an animal chooses a simple action or a complicated schema, the dimension of *how fast* the selected response should be performed is an important one meriting optimal choice. Although response rates can indeed be determined through ongoing selection between responding and resting, or through choices of different numbers of future sub-actions, we argue that our framework is a more natural and general way to treat ongoing, self-paced responding, and is more powerful in elucidating the tradeoffs underlying the selection of vigor.

Within the reinforcement learning framework, the use of the average reward RL rather than the more common discounted reward formulation, proved especially illuminating. Several lines of evidence suggest that indeed an average reward formulation may be better suited to modeling ongoing behavior and its neural substrates (Daw & Touretzky, 2000; Bautista et al., 2001; Daw et al., 2002). Here we have shown that teasing apart the contribution of the net rate of rewards to the value of an action, is especially meaningful in deciding on how much time to allocate to a specific action, as it provides the correct opportunity cost against which the benefit of the currently chosen action should be compared. In the discounted case, for which the simulated results are qualitatively similar, this aspect of decision making becomes obscure.

2.4.3 Limitations of the model and future directions

State transitions

Of course, as a crudely simplified model of decision making in free-operant situations, the model we have presented is severely limited. First, a critical simplifying assumption of the model is that once a decision is made regarding the next optimal action and the latency with which to perform it, the validity of this decision does not change while the action is executed. That is, we have assumed that the state of the world (eg, whether a reward is available in the food well or not) does not change while an animal is executing an action, and that the decision-maker fully determines *when* state transitions can happen. Though this may be true in free-operant schedules, our framework can not be used without modification to model tasks in which this assumption is invalid, such as instrumental avoidance conditioning (in which an aversive outcome occurs if a response is *not* performed). More generally, the model can not incorporate Pavlovian state changes (eg. stimuli appearing and disappearing, and rewards that are given regardless of the animal's actions), and is only a model of appetitive instrumental conditioning. In terms of the model states (the states caused and

predicated upon the animal's instrumental actions), Pavlovian outcomes, whether aversive or appetitive, are merely 'free' outcomes, unrelated to the chosen actions and distributed regardless of the state of the model. This ignores the fact that in a Pavlovian conditioning setting these seemingly free outcomes are actually signaled by predictive cues, which might usefully be incorporated into the state space to determine responding [20].

Moreover, the phenomenon of Pavlovian-instrumental transfer (PIT) in which the onset of a cue that has been associated previously with Pavlovian rewards enhances the rate of ongoing instrumental behavior, shows that Pavlovian cues can suboptimally influence instrumental response vigor in a state (stimulus) dependent way. This highlights a subtlety in the definition of the net reward rate in out model. According to the optimal solution, and consistent with common sense, 'free' rewards should *have no effect* on any ongoing instrumental behavior: any action and rate of responding that were optimal in the original task, are still optimal when additional free rewards are available. This implies that the effective net reward rate used to determine the optimal rate of instrumental responding should not be affected by Pavlovian rewards, that is, that the net rate of rewards controlling instrumental behavior should include only those rewards that are instrumentally earned. But inferring which rewards are earned instrumentally and which would have been delivered regardless of one's actions is not at all a trivial problem. Indeed, although animals show sensitivity to the contingencies between actions and rewards and reduce responding on a lever if rewards are offered at the same rate whether the lever is or is not pressed (as in 'contingency degradation', eg, Balleine & Dickinson, 1998; Corbit & Balleine, 2000; Corbit et al., 2002; Yin et al., 2005), responding in such cases is not completely eliminated, evidence for some confusion as to the causes of rewards. As a result of such overestimation of agency in obtaining Pavlovian rewards, the net instrumental reward rate would be overestimated, leading to instrumental response rates that are higher than is optimal. Indeed, in PIT, Pavlovian cues influence instrumental response rates despite the fact that they do nothing to change the tradeoff determining the optimal rate of responding (eg. Dickinson & Balleine, 1990; Colwill & Triola, 2002; Corbit & Balleine, 2003, 2005; Holland, 2004). Our model suggests that this may be the result of erroneous inclusion of Pavlovian rewards in the expected net rate of instrumental rewards.

Noise and partial observability

Other simplifications that we have made, which are perhaps easier to change, also limit our model's applicability to real-world situations. One is the assumption of an observable state space. The behavior of well-trained rats in conditioning experiments shows that they can readily recognize the sound of food being delivered into the food magazine. It is nevertheless naïve to assume that they know with certainty when food is available in the magazine, and when it is not. Ambiguity regarding state identity is one possible reason

[20] Note that our model is inadequate as a model of Pavlovian *responding* in a more fundamental way. As discussed in Chapter 1, Pavlovian responding is not necessarily normative — rather than a flexible, optimal, adaptation to a task, it seems as if Pavlovian responding is characterized by its inflexibility, and can be shown to be strictly suboptimal (for instance, in omission schedules). If indeed Pavlovian responses are a nonelective consequence of the predictive value of cues or states (Dayan et al., 2006), the applicability of normative models to explaining this behavior is limited.

for the excessive nosepokes shown in the experimental data (Figure 2.4b). Assuming a partially observable state space would render the model more complicated to solve, but there is considerable work in this area on which to base such a formulation.

Another source of noise and partial observability that can be more easily incorporated into the model, is timing noise. We have assumed perfect tracking of the passage of time. It is clear, however, that this is beyond the abilities of animals and humans alike. In fact, the subjective timing of intervals is known to be prone to scalar noise, that is, the standard deviation of the estimates of an interval's duration, is proportional to the mean length of the interval (Gibbon, 1977, 1992; Gallistel & Gibbon, 2000). In a more fundamental way, we have assumed that response selection is noiseless — in our model we used response timing as a proxy for response vigor, adopting the rather simplistic view that animals select a particular latency for their chosen action, and then set about doing it exactly that slowly and with no interruption. Realistically, animals cannot precisely time their own actions, just as they can not precisely time elapsed intervals. We can readily model the inclusion of temporal noise in the action execution process by drawing latencies from a distribution with a mean determined through the process of optimization (see also Chapter 5). We expect that this would not have a great impact on the general form of our results. The inclusion of timing noise would, however, allow our model to accommodate such findings as scalloped responding in fixed interval schedules (Gallistel & Gibbon, 2000). In our current noise-less model, the simulated rat could unrealistically simply time its behavior such that a single leverpress is emitted exactly when the interval has elapsed.

Accounting for excessive nosepoking

Although most studies of free-operant instrumental leverpressing report only leverpress rates, steady-state behavior in our own experiments (eg. Figure 2.4b) shows that, although futile, the rate of nosepoking in the absence of food is highly correlated to the rate of leverpressing throughout the session. This suboptimal behavior, which is not replicated by our model, can not be completely attributed to partial observability of the state space and confusion as to whether there is food in the magazine: the difference between the latencies to nosepoke when a reward is available, and when it is not, testify to the fact that the rats *can* discern between the two situations. One possibility is that in the first nosepoke the rats only partially consume the reward, and full consumption is distributed over several intermittent nosepokes (Peter Holland, personal communication). However, this would not explain the correlation between nosepoke and leverpress rates. Alternatively, the rats may be unsure about the identity of the instrumental action leading to reward. Because every reward was obtained by a leverpress followed by a nosepoke, it is conceivable that some animals in fact learn a combined action unit. That some animals perform combined LP+NP action units could explain the average response pattern that we see.

A third option is that, in the interval schedule from which these data were collected, the nosepoke actions are emitted in order to 'pass the time' and refrain from pressing too fast (perhaps even as an aid in tracking time; Peter Holland, personal communication). Moreover, in random interval schedules there is a higher

probability of reward baiting when a long time interval has passed since the previous leverpress, such as after nosepoking. In realistic learning conditions in which the transition structure of the SMDP must also be estimated from experience (see Chapter 1), this may result in incorrect attribution of the high probability of a transition to a rewarded state to the occurrence of a previous nosepoke. In fact, other aspects of stochastic learning can explain why excessive nosepokes arise in interval (but not ratio) schedules (see Chapter 5, section 5.4.3). Indeed, there is reason to believe that responding on ratio schedules is not characterized by such excessive nosepoking, although detailed data have yet to be examined. Of course, the fact that nosepoking is also the Pavlovian approach response in this scenario should not be ignored, and at least some of the excessive nosepoking may be attributed to appetitive approach behavior, which is, at present, beyond the scope of the model.

Average reward or discounting?

Finally, although the average reward RL formulation has proven so illuminating, it does not provide an accurate description of animal behavior, as animals are known to discount future rewards. For instance, in the model it would be more optimal to first earn several rewards by continuously leverpressing, and only then consume them all (thus avoiding some of the costly transitions from the lever to the food-magazine). In terms of net reward rate, this strategy would not suffer. However, in the experimental scenario, and in more realistic life situations, animals rarely adopt such a policy[21]. It is easy to justify the discounting of future rewards, which should lead to more immediate harvesting of available food at the slightly higher transition cost. For instance, the experimental session may end (and the rat be removed from the chamber) at any time, at which point uneaten rewards will be lost. It is therefore interesting to look at a model which combines a net reward rate measure to determine response vigor, with discounting of future rewards. This may not be trivial from the theoretical RL view, but seems behaviorally justified, and is left for future work.

2.4.4 Conclusions

To summarize, we have presented a reinforcement learning model in which animals optimize not only their choice of responses, but also their rate of responding, in order to harvest the maximal net benefits from their environment. This model, despite its simplicity, replicates many characteristics of free-operant behavior, and makes precise predictions about the relationship between response rates and the different experimental parameters. In addition to the normative starting point it offers for investigations of response vigor, our theory provides a relatively fine scalpel for dissecting the temporal details of behavior, such as the distributions of inter-response intervals at particular state transitions. There is thus great scope for revealing re-analyses of many existing data sets. These can help refine the model by providing more specific constraints, and highlighting where its simplifications cause it to deviate from animal behavior.

[21] To prevent the model from using such a strategy, we imposed the constraint that rewards could not be accumulated in the food magazine. That is, in the state in which $\{i_r = 1\}$, a nosepoke could harvest one and only one U_r.

This chapter is largely based on: Niv, Y., Joel, D. and Dayan, P.,
A normative perspective on motivation, **Trends in Cognitive Sciences**, 2006.

It is better to have an approximate answer to the right question than an exact answer to the wrong one –
John Tukey

Chapter 3

Motivation: The directing and the energizing

Abstract: Understanding the effects of motivation on instrumental action selection is fundamental to the study of decision making. Traditionally, two different functions have been ascribed to motivation: a 'directing' function of steering behavior toward the most valuable outcomes, and an 'energizing' function of activating ongoing behavior. However, despite decades of research, the specification of these functions has remained obscure, and the concept of motivation itself has proved slippery and labile. This has been partly remedied by a recent line of experimental investigations, which have convincingly shown that motivational states indeed 'direct' goal-directed behavior toward more valuable outcomes, by determining the utility of these outcomes. But how motivational states can influence outcome-insensitive habitual behavior is more mysterious. In this chapter, we view this question from a normative perspective, harnessing the model proposed in Chapter 2 to suggest an answer. We start by defining motivation as a *mapping* from outcomes to utilities. Invoking the reinforcement learning framework for maximizing accrued utilities, we then analyze how a motivational shift *should* affect optimal responding. Based on the computational view of habitual behavior as arising from a cache-based system, we suggest that habitual action selection can direct responding properly only in those motivational states which pertained during behavioral training. However, this does not imply insensitivity to novel motivational states. In these, we propose that outcome-independent, global effects of the motivational state can 'energize' habitual actions, as a well-founded approximation to the optimal solution in a trained situation.

3.1 Introduction

Motivation occupies center stage in the psychology and behavioral neuroscience of decision making and instrumental action selection. Yet, early attempts to conceptualize motivational effects through a single mechanism such as homeostasis, drive reduction or incentives, reached a dead end and were largely abandoned

(Berridge, 2004). For many years, the role of appetitive motivation in controlling behavior was not explicitly studied, but rather it was implicitly assumed that motivation (or, rather, deprivation) is necessary for behavior (hence the food or water deprivation of subjects common in most experimental psychology). The last two decades, however, have witnessed a resurgence in interest in motivational effects, and a renaissance in sophisticated analyses of motivation, primarily because different effects of motivation have been successfully used to tease apart different types of instrumental behaviors, namely goal directed and habitual control (see Chapter 1). Manipulations such as specific satiety or motivational shifts suggest that goal-directed and habitual actions are distinguished by the sensitivity of goal-directed, but not habitual responding to changes in the utility of their specific outcomes (Dickinson & Balleine, 2002). Although goal-directed and habitual behavior can be characterized by their differing motivational sensitivities, and the effects of motivational manipulations on goal-directed behavior are, by now, relatively clear, exactly how (and indeed, whether) motivation influences habitual responding has remained unresolved. This is particularly disturbing because habitual responding plays a very prominent part in both normal and abnormal behavior, hence without answering this question we can not claim a satisfactory understanding of either motivational control or habitual responding.

That our understanding of motivational control is lacking may be partly due to the fact that motivation itself is not a unitary construct (Berridge, 2004). In fact, Dickinson and Balleine (2002) trace back to Descartes two distinct influences of motivation on behavior: a 'directing' effect, determining the current goal(s) of behavior (eg. food or water), and an 'energizing' effect, which determines the force or vigor of behavior. The latter is closely linked to the Hullian concept of 'Generalized Drive' (Hull, 1943; Brown, 1961; Bolles, 1967), a motivational process that energizes all pre-potent actions. Whereas much is known about the directing aspects of motivation, the 'energizing' effects of 'generalized drive' have remained highly controversial.

In this chapter, we confront these challenges by using the model presented in Chapter 2 to underpin a computational analysis of the optimal effects of motivational states on action selection. We start by suggesting a simple, normative notion of motivation, with which we can define precisely outcome-specific 'directing' effects and outcome-general 'energizing' effects. This allows us to analyze the effects that a change in motivational state should optimally have on action selection. Combining these results with the computational constraints of the habitual system (Daw et al., 2005), we suggest that, without retraining, the outcome-specific effects of motivation can only influence goal-directed behavior. We further hypothesize that, as a well-found approximation to the optimal effects of motivation on behavior, the habitual system can be controlled by motivation through outcome-general 'energizing' effects. This provides the age-old notion of 'generalized drive' with a firm normative basis. As only preliminary experimental results on the latter hypothesis exist, Chapter 6 will describe how it can best be tested, presenting supporting results from two behavioral experiments.

3.1.1 Motivation: A mapping from outcomes to utilities

'Motivation' is a concept that is widely invoked in a variety of meanings and contexts. For example, whereas the Oxford English Dictionary defines motivation as "the *factors* giving purpose or direction to human or

animal behaviour", Salamone and Correa (2002) define motivation as "the set of *processes* through which organisms regulate the probability, proximity and availability of extenal and internal stimuli" (p.5; emphasis my own in both). To obviate vagueness in our analysis of the effects of motivation, we would like here to precisely define what we mean when we refer to 'motivation'.

Our conception of motivation is strongly influenced by the field of reinforcement learning (Sutton & Barto, 1998). As discussed in Chapter 2, in reinforcement learning outcomes such as food or water have numerical utilities measured in some common currency, and the imperative is to choose actions that maximize some long-term measure of total utility. Because outcomes may have different utilities in different motivational states, we define motivation simply as *the mapping between outcomes and their utilities*. We further refer to 'motivational states' (eg, 'hunger' or 'thirst') as *indices* of different such mappings (such as one in which foods are mapped to high utilities, and another in which liquids have high utilities). 'Motivational shifts' will therefore refer to shifts between different motivational states.

This definition of motivation is a pragmatic rather than a philosophical or psychological definition. We thus avoid, for now, important issues such as the grounding of these mappings in evolutionary fitness, and the emotional aspects of motivational states, as these are fairly orthogonal to a reinforcement learning treatment of motivation. We will show that even this fairly impoverished notion of motivation carries much explanatory power. Note, also, that this definition is means-neutral, in that organisms need not know the motivation-determined utilities, or have these utilities affect behavior in any way, for a motivational mapping to be meaningfully defined. Even if a dehydrated worm does not know the utility of different locations in terms of hydrating it, or how to get to those locations, by mere definition of being 'thirsty' some locations are now worth more to it than others. This sort of abstraction is useful, since we will consider circumstances in which different decision-making systems may not have access to, or knowledge of, the true mapping, and can only approximate it.

How can an animal modify its behavior so as to maximize the utility it gains from its environment given its motivational state? As an illustration, consider a rat presented with two levers in an operant chamber. The left lever is programmed such that pressing it five times is rewarded with food, while the right lever rewards with water for every five presses (a concurrent fixed-ratio (FR5-FR5) schedule; Figure 3.1a). When in different motivational states (satiety or hunger), and given the different utilities of the outcomes in each of these (Figure 3.1b), how can the rat decide whether to press the left or the right lever, and at what rate? In the following section, we first analyze how a change in motivational state *should* affect the cost-benefit tradeoffs determining action selection and response vigor. We then discuss the two strategies for action control (Daw et al., 2005), namely, the goal-directed forward model based strategy, and the habitual cache-based strategy, and how each *could* be controlled by motivational states.

We note also the (considerably simpler) problem of learning the motivational mapping itself. It may seem trivial that food is worth more when hungry, and water is worth more when thirsty. But this might, in fact, depend on a learning process, and may be a result of past learning from experience with these outcomes in different motivational states (Dickinson, 1997). For instance, imagine a situation in which, famished after a long day's hike, you enjoy a hearty dinner at a road-side restaurant. The next time you pass by the restaurant

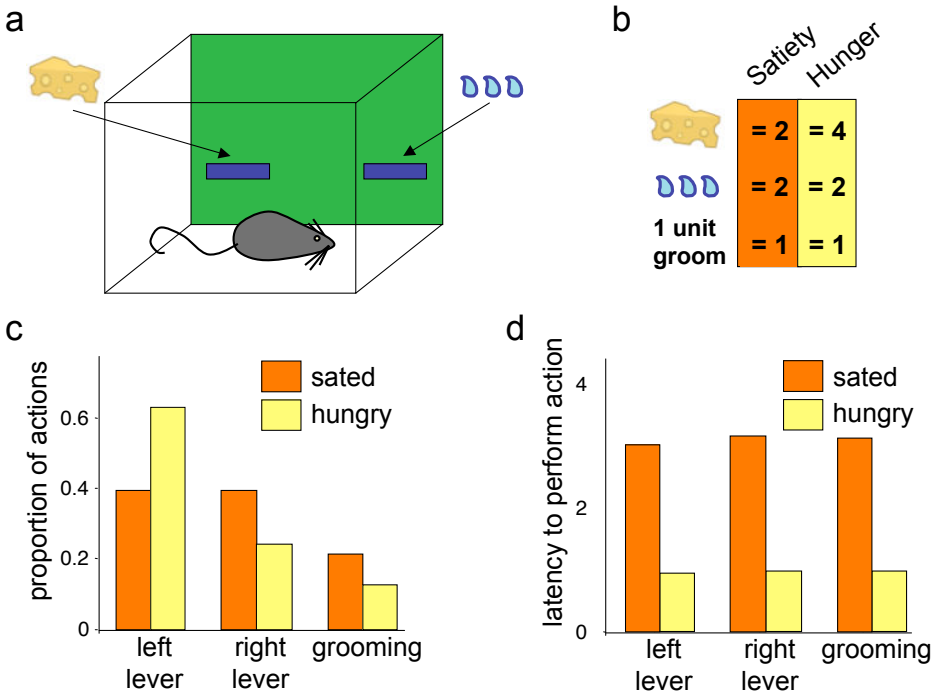

Figure 3.1: a. A simulated rat, trained in an operant chamber, can do one of three things: press the left lever (L) to obtain cheese on an FR5 schedule, press the right lever (R) to obtain water on an FR5 schedule, or groom to obtain an internal reward. Cheese and water outcomes are harvested at a food-well, by making a nosepoke (NP) action. **b.** We assume, for this illustration, that even when relatively sated, the cheese and water have slightly higher utilities than grooming. A shift to hunger, however, markedly enhances the utility of cheese, compared to the other utilities which are left unchanged. **c.** Not surprisingly, as a result of the shift from satiety to hunger, the rat chooses to press the left lever more often, at the expense of either grooming or pressing the right lever (which are still performed, albeit less often). **d.** A second consequence of the motivational shift is that all actions are now performed faster. This 'energizing' effect of the motivational shift is thus not specific to the action leading to the favored outcome, and can be regarded an outcome-general effect. Simulation parameters: $U_r = 20$ for either food or water when sated, $U_r = 40$ for food when hungry; $t_{eat} = 0$; $\beta = 0.2$; unit costs depended only on the chosen action $C_u(\mathrm{L}) = C_u(\mathrm{R}) = 0.15$, $C_u(\mathrm{NP}) = 0.05$, $C_u(\mathrm{Groom}) = -1$; and vigor costs were only dependent on whether the current and previous action were or were not similar: $C_v(\mathrm{same}) = 0.5$, $C_v(\mathrm{different}) = 1$.

you remember the meal you had there - but was the meal so good because of the quality of the restaurant, or merely because you were so hungry? This question is not only theoretical, as its answer should affect your decision to frequent the restaurant again. Similarly, the subjective utility of cheese that a rat has only experienced when hungry, might be ambiguous when the rat is suddenly shifted to a sated motivational state (and tested in extinction, ie, in the absence of rewards; Dickinson, 1997; Dickinson & Balleine, 2002).

Indeed, results from 'incentive learning' experiments show that for a motivational state to control responding through determining the utility of outcomes, these outcomes need to be *experienced* (consumed) in this motivational state (Dickinson & Balleine, 1994, 2002; Balleine, 2000; Colwill & Rescorla, 1986). For instance, in our concurrent ratio task, a shift from satiety to hunger might not affect the navigation choices of a rat that has never consumed cheese when hungry, presumably because the motivational mapping itself is unknown to it. However, even a chance to eat cheese in its home cage (not in the operant chamber) when

hungry, will suffice to allow it to learn the cheese's now higher utility, and alter its preference for the right and left levers (eg. Balleine, 1992; Lopez et al., 1992; Balleine et al., 1995).

3.2 Results: The effects of motivational shifts on vigor and action selection

Using this definition of motivation, we can directly model the effects of motivational shifts such as from satiety to hunger, by letting the motivational state of the animal determine the utilities U_r of the available outcomes within the framework presented in the previous chapter. Unfortunately, available data do not pin down the form of this mapping precisely, and we set it arbitrarily (for instance, a morsel of cheese might be worth two times as much to a hungry rat as to a sated rat). For simplicity, we also assume that motivation is constant during an experimental session, and ignore small changes in utility potentially resulting from progressive satiation over the course of a session.

Recall that in Chapter 2 we modeled decision-making in a free-operant scenario such as that in Figure 3.1a as a continuous-time semi-Markov decision process, in which the goal of the rat is to maximize the net rate of reward ($\bar{R}$). The task was characterized by a series of states S, in each of which the rat chooses an action a and a latency τ, entailing a unit cost $C_u(a, a_{prev})$ and a vigor cost $C_v(a, a_{prev})/\tau$, and causing a transition to a new state S' with probability $\mathcal{T}^{a,\tau}_{S\to S'} \geq 0$, as well as an immediate reward of utility U_r with probability $P^{a,\tau}_r(S) \geq 0$ (see Chapter 2, Figure 2.6). One way to behave optimally in this scenario is to compute long-run differential values for each action and latency at each state (denoted $Q(S,a,\tau)$ and corresponding to the expected long-term net gains contingent on performing (a,τ) at state S), and then choose the action and latency with the highest Q-value at every state. Because the value $Q(S,a,\tau)$ equals the sum of all future rewards minus costs, it can be defined recursively as the sum of the immediate rewards (from the environment) minus costs (for the chosen action, and for the opportunities forgone in the chosen latency), plus the (long-run) value of the next state, giving[1]:

$$Q(S,a,\tau) = P^{a,\tau}_r(S)\cdot U_r - C_u(a,a_{prev}) - \frac{C_v(a,a_{prev})}{\tau} - \tau\cdot\bar{R} + \sum_{S'\in\mathcal{S}} \mathcal{T}^{a,\tau}_{S\to S'} V(S') \tag{3.1}$$

where $V(S')$ is the (policy-dependent) long-run differential value of the successor state S', ie, $\langle Q(S',a,\tau)\rangle_\pi$ (see Chapter 2, section 2.2.3).

From equation (3.1) it is obvious that changing the utility of an outcome U_r will affect the values $Q(S,a,\tau)$ of actions that can lead to that outcome with probability $P^{a,\tau}_r$ greater than zero, and can indirectly affect the Q-values of all actions that lead to states from which this outcome can later be reached (through changes in $V(S')$). Due to these effects, after a shift from satiety to hunger the values of actions ultimately leading to food will be higher, which will result in a higher probability of choosing these actions, as in the classic 'directing' effect of motivation.

[1] We repress here the dependence of Q, V and $\bar{R}$ on the current (π) or the optimal (*) policy, as our point regarding the effect of a change in U_r is true for either case.

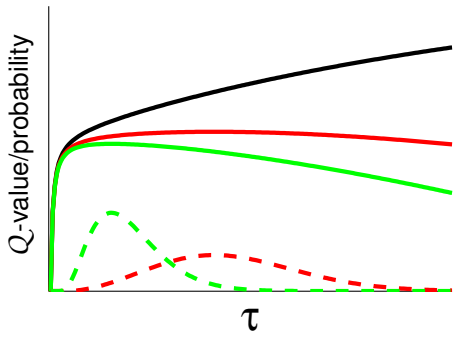

Figure 3.2: Q-values for an action a at different latencies τ. In black (top) are the 'unadjusted' Q-values, before subtracting $(\tau \cdot \bar{R})$. In red (middle, solid) and green (bottom, solid) are the values adjusted for two different net reward rates. The higher net reward rate (green) penalizes long-latency actions more, thereby causing *faster* responding, as shown by the corresponding soft-max action probability curves (dashed), which have an optimum at a shorter latency for higher $\bar{R}$.

Importantly, this is not the only way that a motivational shift will affect action and state values. According to the analysis of the effects of reward magnitude on behavior (Chapter 2, section 2.3.4), changing the utility of a reward also affects the obtainable net reward rate $\bar{R}$, with an outcome of higher (subjective) magnitude making the net reward rate higher. As the optimal latency to leverpress τ_{LP} is inversely proportional to $\bar{R}$ (for instance, recall from Chapter 2, section 2.3.3, that in a ratio schedule $\tau_{LP} = \sqrt{\frac{C_v(LP)}{\bar{R}}}$), the higher net reward rate when hungry means that the rat should choose to leverpress with a shorter latency, ie, at a faster rate. Note, though, that because the optimal τ for actions irrelevant to obtaining food has a similar dependence on $\bar{R}$, they too should become more vigorous as a result of hunger. The key to understanding this is the role that $\bar{R}$ plays in the tradeoff determining optimal response vigor. In effect, the expected net reward per unit time quantifies the opportunity cost for doing nothing (and receiving no reward) for that time; its increase thus produces general pressure to work faster. Intuitively, hunger encourages the rat to complete irrelevant actions faster, in order to be able to resume leverpressing quickly. The provenance of this effect is presented graphically in Figure 3.2. The higher $\bar{R}$ increases the cost of sloth, since every τ time without reward forgoes an expected $(\tau \cdot \bar{R})$ net reward. Higher net reward rates thus penalize slow actions more than they do rapid ones, tilting action selection toward faster behavior, for *all* pre-potent actions.

To summarize, when the rat is made hungry, the higher utility of food should result both in higher values for actions eventually leading to food, and in a higher net reward rate. The former will lead to outcome-specific 'directing' effects by which the state of hunger affects the propensity to perform actions that lead to food, while the latter will lead to an outcome-general 'energizing' effect on the vigor of all actions. These two effects are illustrated in Figure 3.1(c,d), which shows results from simulating a motivational shift in the concurrent fixed ratio task. Panel (c) shows that leverpressing for cheese is enhanced by hunger, at the expense of leverpressing for water, or grooming. Panel (d) shows that, as a result of hunger, all actions are performed at a faster rate regardless of the identity of their outcome. Note that because the 'directing' effect of motivation affects action selection and the 'energizing' effect affects vigor selection, these two effects of motivation can be seen as somewhat orthogonal.

We have so far established that the 'directing' and 'energizing' effects of motivation can be normatively grounded when considering behavior that aims to maximize net reward rate. An all-knowledgable controller determining behavior optimally should show both effects. However, it is not in all cases that we can expect

(even optimal) behavior to accurately reflect the motivational mapping. For instance, when circumstances change suddenly, some experience with the task might be necessary in order to learn new action values and adjust behavior to the new situation. Indeed, in order to delineate the (neural) interactions between action controllers and motivational influences, experiments testing the effects of motivational shifts are typically interested in uncovering the *immediate* effects of such a manipulation, prior to new experience-based learning (it is for this reason that the effects of motivational shifts are tested in extinction). It behooves us to ask, then, what could the immediate effects of motivational shifts be, given the different computational strategies which we think the brain uses to compute action values?

As discussed in Chapter 1, there is extensive evidence (see Dickinson & Balleine, 2002, for a review) that decision-making in mice, rats and primates can be based on (at least) two neurally distinct (Balleine, 2005) controllers, which employ different computational strategies (Dickinson, 1985; Daw et al., 2005). Goal-directed action selection, driven by forward-model tree-search (or so-called 'response-outcome' associations, Dickinson & Balleine, 1994, 2002), and habitual action selection, based on cached state-action Q-values (Watkins, 1989; or 'stimulus-response' associations, Dickinson, 1985). Below, we discuss how each action selection scheme can be influenced by motivation, given its computational constraints. We show that the division between outcome-specific 'directing' and outcome-general 'energizing' effects of motivation aligns computationally and psychologically with the division between goal-directed and habitual control.

3.2.1 Goal-directed behavior: A 'brute force' solution

Almost by definition, the goal-directed system is thought to employ what is called a *forward model*, working out the ultimate outcomes consequent on a sequence of actions by searching through the tree of state-actions-consequences, and choosing actions based on the outcomes' current utilities (Figure 3.3a; Daw et al., 2005). By using specific satiety and conditioned taste-aversion procedures to selectively alter the utility of a single outcome, it has been shown that action selection in this system is sensitive to manipulations that alter outcome utilities (Balleine & Dickinson, 1998, 2000; Corbit et al., 2001; Balleine et al., 2003; Killcross & Coutureau, 2003; Coutureau & Killcross, 2003; Holland, 2004; Yin et al., 2005, 2005). Further, studies employing shifts in motivational state have shown that these too affect goal-directed behavior through the determination of outcome utilities (see Chapter 1, section 1.1.3).

Figure 3.3a illustrates such a controller. Depicted is a (partial) forward model of the task (a state-action-outcome tree). Using this controller to decide which lever to press at the initial state $S_{(0,0)}$ (that is, in the state in which neither the left not the right lever have been pressed) is straightforward: the rat can mentally search through the tree (essentially simulating possible action choices and the consequent evolution of the states of the environment) and find the path with the highest overall utility. Because the rat's motivational state determines the mapping between outcomes encountered while traversing the forward model and their subjective utilities, action choice will automatically incorporate motivational influences and the rat will choose the left lever when hungry, the right lever when thirsty, and either when sated. Goal-directed control is therefore

motivationally straightforward, with outcome utilities directing actions to the most valued outcomes appropriately. However, this form of search in a forward model constitutes a 'brute-force' solution to the action selection problem, involving high costs of computation and working memory for the forward model search and evaluation mechanism. In order to compute accurate action values, a search into the distant future may be necessary, involving an increasingly branching tree of options. Even in the simple concurrent FR2-FR2 case we have depicted, and although we have omitted the choices of grooming and nosepoking in most of the states, the forward model is relatively complex. Most worrying, then, is that in the complex situations with which we are faced daily, the search costs may be sufficiently prohibitive such as to render a full forward-model search intractable (Samuels, 1959; Baum & Smith, 1997; Daw et al., 2005). In fact, the reason we have here ignored the choice of response vigor, and concentrated only on discrete action selection, is that a continuous action space (that incorporates, for instance, the choice of response latency) much complicates the evaluation of the forward model (see also the Discussion below).

3.2.2 Habitual behavior: A motivation-insensitive shortcut?

Action selection in the habitual system circumvents the need for representing and evaluating a large tree of possible future action sequences, and is instead based on previously learned and stored *(cached)* values of different actions at different states (Figure 3.3b). These so-called $Q(S,a)$-values summarize previous experience in a scalar measure of the estimated long-term return contingent on choosing a specific action a when at a specific state S. 'Model free' reinforcement learning methods provide a neurally feasible framework for learning such values from experience with the task dynamics (ie, from trial-and-error behavior, see Chapters 1 (section 1.2.2) and 5). These rely on the self-consistency of consecutive values, that is, the fact that the expected future reward from the current time and onward must equal the current reward plus the expected future reward from the next timestep and onward (as seen in equation 3.1 and in the various Bellman equations, eg, equations 1.7, 2.7, 2.9 and 2.12). For example, the Q-values in Figure 3.3b are the optimal values (that is, those corresponding to the optimal policy) that would be learned using standard undiscounted temporal difference reinforcement learning.

Importantly, in a caching system, values are defined in terms of the expected cumulative (or average) future utilities consequent on performing an action at a specific state. Adding together the utilities of different possible future outcomes (food, drink, mates, etc.) in some common currency, cached values are thus ignorant of the specific identities of the expected outcomes. At decision points, actions are chosen by comparing their relative cached values, rather than their consequent outcomes (Daw et al., 2005). If the values indeed correspond to the long-run expected payoffs from each action in each state, optimal action selection is guaranteed. Though less powerful than methods involving forward models, this sort of controller offers substantial computational savings, which underlies its popularity in reinforcement learning. Further, as reviewed in Chapter 1, these methods have been tightly linked to phasic dopaminergic signals in the brain (Montague et al., 1996; Schultz et al., 1997).

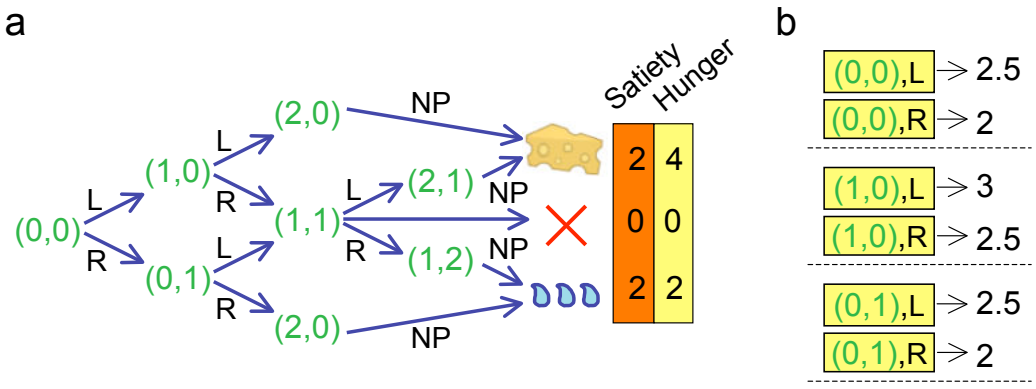

Figure 3.3: Two ways to estimate action values for action control. For illustration purposes we concentrate here only on discrete actions, ignoring vigor costs and the choice of response vigor. **a.** A forward model of a simplified version of the task depicted in Figure 3.1. In this episodic concurrent FR2-FR2 task, an NP action terminates a trial and resets the lever counters. States (green) represent the previous number of presses on the L and R levers in the current trial. Although we have not depicted grooming and nosepoking at many of the choice points, the tree of options is fairly complex. Importantly, the current motivational state of the rat defines the relevant mapping between outcomes and utilities (shown in the boxes). Thus, by searching through the forward model the rat will find choice L optimal at $S = (0,0)$ when hungry (yellow), and will choose L and R equally often when sated (orange). **b.** In contrast, a caching system does not represent a model of task dynamics, but rather stores (caches) learned values (in common-currency units) for every action a at every state S. Here we show a part of the cache, that was learned in the state of hunger. For simplicity, the cost per unit action was set to $C_u = 0.5$ for each of $\{\mathrm{L},\mathrm{R},\mathrm{NP}\}$, giving, eg, $Q(S_{(0,0)},L) = 4 - 2C_u(\mathrm{L}) - C_u(\mathrm{NP}) = 2.5$. Action selection in this system simply involves choosing the action with the greatest cached value at the current state. Since the values are divorced from the identities of the outcomes produced by different actions, changes in the outcome-utility mapping cannot be translated to the appropriate changes in values. However, the motivational state can be stored as part of the state representation, such that separate cached values can be learned for each motivational mapping.

As, by contrast with goal-directed actions, habitual behavior is operationally defined by its insensitivity to the specific identities of its consequent outcomes (Adams, 1982; Yin et al., 2004), how, then, can habits be influenced by a change in the motivational mapping of outcomes to utilities? One straightforward way by which habitual behavior could be influenced by a new motivational state is through learning of new values, based on performing the different actions at the different states, and experiencing the new utilities. Indeed, the motivational state effective at the time of learning can be used to index the values learned in different motivational states, and keep them separate. Thus habitual behavior can indirectly learn a motivation-dependent behavioral policy which properly directs actions to maximize outcome utility in different (but known) motivational states.

But what about the immediate effects of a motivational shift, prior to retraining? The experimental status of this question is still debated — some studies argue for complete insensitivity to outcome revaluations (Adams, 1980, 1982; Robbins & Everitt, 1996; Killcross & Coutureau, 2003; Coutureau & Killcross, 2003; Yin et al., 2004), while others claim that habitual behavior is directly and immediately sensitive to revaluation manipulations (Dickinson et al., 1995). We suggest that this confusion stems from the frequent treatment of outcome revaluation by a motivational shift as equivalent to outcome revaluation by specific satiety or conditioned taste-aversion. Indeed, unlike goal-directed control, a habitual system cannot direct

action selection based on new outcome utilities without the learning of new values described above, explaining the lack of sensitivity to the latter outcome revaluation procedures, which specifically target and change the utility of a single outcome. Nevertheless, in the case of motivational shifts, we claim that even without new learning, habitual behavior can be partially adapted using two different well-founded approximations to the desired effects of the new outcome-utility mapping. One involves a form of generalization gradient, based on an internal representation of the motivational state; the other involves an 'energizing' effect on ongoing behavior. Below we describe these approximations.

Approximation 1: Generalization decrement

We argued above that the cached values can be indexed by the current motivational state. In this case, a change in motivational state from training to test will also induce a change in states S, and thus potentially lead to a change in the estimated action values. Given the evidence for generalization decrement following a change in stimuli between training and test (Domjan, 2003; ie, a *reduction in responding* when tested with stimuli different from those with which the behavior was trained), one may expect that a shift to a novel motivational state may also, in general, lead to decreased conditioned responding (Brown, 1961; Bolles, 1967; Dickinson & Balleine, 2002).

The psychological learning literature talks about responding to a new stimulus as comprising a summation of decrements with respect to each of the previously trained stimuli (eg. Pearce, 1987, 1994). From a normative perspective we can view the predictions learned for different states as forming a 'value landscape', in which, as a result of generalization to a novel situation, expected value (and hence instrumental responding) may either decrease or increase compared to the training situation. Because in conditioning experiments animals are typically trained to expect a special reward in only one situation (the training chamber), generalization to a novel situation will normatively involve predicting fewer rewards. However, it is conceivable that in certain scenarios generalization to a new experimental situation would involve increased expectation of reward, and thus potentially a form of *generalization increment*. For instance, after training in a high deprivational state as well as in a low deprivational state, generalization to a novel motivational state of intermediate deprivation may involve an increment in conditioned behavior compared to performance in the low-deprivation training (and a decrement compared to the high-deprivation training). Note, though, that in terms of the effects of a motivational shift on habitual behavior, if, as is usually the case, the behavior was trained only in one specific motivational state (eg. 8 hours food deprivation), by virtue of the 'value landscape' only having one peak at the trained motivational state, we could expect a motivational shift to result in generalization decrement, ie, a decrease in habitual responding, even if the new motivational state is one in which the outcomes have a *higher* utility, for instance, 16 hours food deprivation.

Approximation 2: 'Generalized drive'

The second form of generalization stems from the fact that outcomes tend to have higher utilities in more deprived states, making the expected net reward rate higher. As discussed above, this should exert an invig-

orating effect on all pre-potent actions, regardless of their outcome. This means that in motivational states such as hunger or thirst, in which the net reward rate is high (since the utilities of food or fluid outcomes are high), all pre-potent actions should be performed quickly, and in states such as satiety, with lower net reward rates, all actions should be performed more slowly. Therefore, provided only that it has an idea as to whether the net rate of reward in a new motivational state will be higher or lower, the habitual system can respond approximately appropriately, by modulating the rate of performance of all actions regardless of their consequences, even before any new learning. This result gives the old (and controversial) psychological notion of 'generalized drive' a new, normative, interpretation as an optimal solution to an action-selection problem. By incorporating sensitivity to *expected* net reward rates in determining rates of responding, the habitual system can at least approximate the optimal choices of response rates immediately, even if not the optimal choice of actions. Of course, given additional training, this approximation will be refined and action selection will become precisely correct once the new values are learned.

Dickinson et al. (1995) showed habitual responding to be directly sensitive to a motivational shift from hunger to satiety, even without the need for an 'incentive learning' stage (which is necessary for goal-directed behavior to come under the control of a new motivational state). With the above suggested approximations in mind, we can now interpret these results. First, that the effect did not necessitate incentive learning is perhaps not surprising — presumably the influence of motivation on habitual behavior is not mediated by a change in the utility of a specific outcome, thus learning about such a specific utility has no effect. Second, either of the two approximations we have suggested, generalization decrement as a result of the novel motivational state or a 'generalized drive' effect as a result of a lower expected net reward rate, could potentially explain the experimental result. Chapter 6 will detail how the use of motivational up-shifts and side-shifts, as well as training with several different outcomes, can tease these possibilities apart, and can make a conclusive case for or against our 'generalized drive' hypothesis. Results from two such experiments described there indeed support a role for both 'generalized drive' and generalization decrement in habitual responding, and show no evidence for outcome-specific effects of either motivational side-shifts or up-shifts.

3.3 Discussion: Two sides of motivational influence

In summary, we have used a normative analysis of the effects of motivational states (or, in our view, mappings from outcomes to their utilities) on behavior, to show that motivation should have two orthogonal effects: an outcome-specific 'directing' effect on action selection, and an outcome-agnostic 'energizing' effect on the choice of vigor. Analysis of the computational characteristics of habitual and goal-directed behavior suggests that the 'directing' effects of motivation can only influence goal-directed behavior, while 'energizing' effects can also be seen habitual responding. This distinction calls for the operational definition of habitual behavior to be slightly refined — habits are not, in general, insensitive to outcome revaluation, but only do not show outcome-specific sensitivity to revaluation manipulations.

Of course, the motivational influences that we derived using the model from Chapter 2 define *the* optimal way a policy should adjust in response to a motivational shift, independent of the specific implementation

by which the policy is computed (ie, a forward model or cached values). Thus, theoretically, outcome-independent 'generalized drive' effects that do not necessitate incentive learning should also be seen in goal-directed behavior. These effects have not been reported previously, but, because such effects might be overwhelmed by 'directing' effects, teasing them apart will require a careful analysis of inter-response latencies and not overall response counts.

As with the computational limitations on habitual control, however, the forward model mechanism is also severely limited, especially when computing a forward model of a continuous-action policy (as is the case when deciding on both action selection and action vigor) and it is likely that it does not compute the full optimal policy. As a result, goal-directed behavior might be *insensitive* to the outcome-general 'energizing' effects of motivation. Alternatively, it is conceivable that dopaminergic influences on the goal-directed system (specifically, through the mesolimbic projection to the dorsomedial striatum) are used to determine goal-directed response vigor based on a *cached* measure of the net rate of reward, similar to the influence of tonic dopamine on response vigor in the habitual system (as will be discussed extensively in the next chapter). In as far as the current literature can shed light on these two possibilities, studies in which goal-directed behavior was subjected to a motivational shift without an 'incentive learning' stage (ie, in which there are presumably no 'directing' effects), do not seem to show any immediate energizing effects (Balleine, 1992; Lopez et al., 1992; Balleine et al., 1995; Dickinson et al., 1995)[2]. However, whereas action selection in the goal-directed system has been shown to be relatively dopamine independent, the effects of dopamine on the vigor of goal-directed responding have yet to be carefully investigated.

Furthermore, we have previously suggested (Daw et al., 2005) that goal-directed forward-model planning and value-caching habitual control represent two extreme ends of a spectrum of approaches to the substantial computational complexities of control in sequential decision-making tasks. It is possible that control incorporating a mixture of forward models and cached values may be evident in some domains (for instance, one that uses a forward model to expand only a few levels of the future tree of possibilities, and then falls back on cached values at the leaves of this tree), and that therefore the simple pattern of motivational sensitivity we have described may not be universal. However, we might speculate that the distinction between, and nature of, 'directing' and 'energizing' are not merely a happy coincidence useful in dissecting different methods of control, but rather direct products of the computational requirements of control in rich environments.

As for a generalization from our results, which dealt only with instrumental behavior, to the domain of Pavlovian control, an open question remains as to whether there is a similar dissociation in Pavlovian behavior, both in terms of two response controllers and in their susceptibility to motivational influences. Unlike Dickinson and Balleine (2002), who emphasize the difference between instrumental and Pavlovian motiva-

[2]Daw et al. (2005, 2006) suggest a different theoretical interpretation of the effects of incentive learning. According to this, a shift to a novel motivational state (or any other outcome revaluation manipulation that is not accompanied by a chance to experience the outcome's new utility) increases the uncertainty that the goal-directed system places in its value estimates. They argue for a normative arbitration between goal-directed and habitual control in which the uncertainty each system has in its value estimations is compared, and the more certain system assumes control (see also Chapter 1, section 1.2.3). Thus, as a result of the increased uncertainty of the goal-directed system due to outcome revaluation, behavioral control reverts to the habitual system, which does not show outcome-specific susceptibility to motivational influences. Unfortunately, this explanation also offers no explanation as to why 'energizing' effects are not seen following a motivational shift without incentive learning.

tional process, our analysis does not in any way rely on such a distinction, and, to the contrary, we expect that the Pavlovian control structure parallels that of instrumental behavior. Sporadic studies point to this direction (Holland & Straub, 1979; Blundell et al., 2003), however, the effects of motivational shifts and outcome revaluation have not yet been examined in Pavlovian conditioning as rigorously as in instrumental conditioning.

3.3.1 Relationship to previous models

Models and theories of the role of motivation in driving behavior were made popular in the early years of experimental psychology by Hull (1943), but had received considerably less attention following the 'cognitive revolution' in psychology. Our theory relates back to the Hullian concept of drive ('big D'), and differentiates this from other aspects of response choice, in that it shows that the effects of drive are in many ways orthogonal to those of the specific utilities contingent on different actions. Thus, our hypothesis regarding the control of vigor is different from the suggestion of Dickinson and Balleine (2002), that *incentive motivation* to a specific outcome (Hull's 'K') energizes actions leading to it. However, reaction times in discrete trial tasks show that an effect of incentive motivation does indeed exist, in that responding for a more desirable outcome is in general faster than that for a lesser reward (eg. Watanabe et al., 2001; Takikawa et al., 2002; Lauwereyns et al., 2002b). How our model can relate to and account for this result will be discussed in Chapter 4.

Killeen's 'Mathematical Principles of Reinforcement' (MPR; see Chapter 2, section 2.4.2) also relates to both incentive motivation and general arousal (Killeen et al., 1978; Killeen, 1995; Killeen & Sitomer, 2003). The first principle of MPR is that incentives excite responding, according to an arousal level which is proportional to the rate of reinforcement (Killeen, 1995). Moreover, similar to our notion of motivation, in MPR an outcome's utility is dependent on the deprivational state of the animal. The main difference between MPR and our analysis here is that, in Killeen's framework, reinforcement rate is separately accounted for each outcome, such that in MPR, arousal is more related to incentive motivation than to 'generalized drive'. Our model thus stands alone as one attempting to provide a precise computational (and in our case, normative) explanation for the concept of 'generalized drive'.

3.3.2 Neural underpinnings of motivational control: clues from PIT

The division of motivational influences into outcome-dependent and outcome-independent effects has an interesting parallel in the phenomenon of Pavlovian-instrumental transfer (PIT; Estes, 1948; Lovibond, 1983, see Chapter 1, sections 1.1.4 and 1.4). In PIT, stimuli classically conditioned to predict the occurrence of affectively significant outcomes affect the vigor of instrumental responding. As with motivational influences, there are two sorts of PIT: specific, in which a stimulus only affects instrumental responding for a similar outcome, and general, in which a stimulus has a general influence on all instrumental actions regardless of

their outcome (Cardinal et al., 2002; Holland, 2004). This latter effect is reminiscent of the 'generalized drive' effect which we have tied to the net rate of reward. Neurally, general PIT has been shown to be dopamine dependent (Dickinson et al., 2000; Wyvell & Berridge, 2000), and relying on the nucleus accumbens core and central nucleus of the amygdala (Hall et al., 2001; Holland & Gallagher, 2003), while the nucleus accumbens shell and the basolateral amygdala have been implicated in mediating specific PIT (Corbit et al., 2001; Killcross & Blundell, 2002; Balleine et al., 2003; Holland & Gallagher, 2003).

Building on what is known about the neural substrates of the two forms of PIT, and on the rapidly accumulating literature regarding the substrates of goal-directed and habitual control (see Chapter 1), we can begin to speculate as to the neural basis of the two forms of motivational influence on behavior. In accord with computational models of dopamine function (Montague et al., 1996) and the role of dopamine in habitual learning and action selection (Faure et al., 2005; Salamone & Correa, 2002), we propose that the 'generalized drive' or 'energizing' effect of motivation on responding is dopamine-dependent (Weiner & Joel, 2002; Niv et al., 2005), possibly mediated by the nucleus accumbens and the central nucleus of the amygdala (Cardinal et al., 2002). The following chapter will expand on this hypothesis. In contrast, 'directing' motivational control, through determination of specific outcome values, may be dopamine-independent, and is possibly mediated by the posterior basolateral amygdala (Robbins & Everitt, 1996; Cardinal et al., 2002; Balleine, 2005) and the orbitofrontal cortex (Cardinal et al., 2002). Moreover, the suggested dopamine-dependence of 'energizing' effects, together with the demonstrated dopamine-dependence of general PIT, prompts the tantalizing suggestion that the bases for the two may be the same, providing a potentially strong link between motivation and classical (Pavlovian) conditioning in controlling instrumental behavior.

3.3.3 Conclusions

Motivation turns out to be a rich and complex topic, because it has multiple facets to which the various action-selection systems are differentially sensitive. Oddly, it has been easier to use motivation to dissociate these systems than it has been to use them to elucidate motivation. Our definition of motivational states in terms of mappings between outcomes and utilities provides a simple normative scaffold on which to understand both optimal and approximately optimal sensitivity to outcome utilities. These ideas regarding the ways motivation influences action selection, and specifically habitual control, are not only significant for the understanding of motivation, but also provide a possible normative foundation for the much debated concept of 'generalized drive'. The use of computational models grounds this concept in precise predictions about what the effects of 'generalized drive' should be, and how they should be measured in order to tease them apart from qualitatively different, orthogonal effects of other aspects of motivation.

This chapter is largely based on: Niv, Y., Daw, N.D., Joel, D. and Dayan, P., *Tonic dopamine: Opportunity costs and the control of response vigor*, **Psychopharmacology**, 2007.

Of course, it's all chemical to start with – "Dopamine", the movie

Chapter 4

Neural substrates: Dopamine and response vigor

Abstract: Dopamine neurotransmission has long been known to exert a powerful influence over the vigor, strength or rate of responding. However, there exists no clear understanding of the computational foundation for this effect because predominant accounts of dopamine's computational function focus on a role for phasic dopamine in controlling the discrete selection between different actions, and say nothing about response vigor. In Chapter 2, we presented a model of free operant response rates that treats exactly this aspect of response selection. In the average reward reinforcement learning model we constructed, optimal control chooses a best response latency so as to balance the costs of acting quickly against the benefits of getting reward earlier, and maximize the overall net reward rate accrued. In this framework, the long run net rate of reward plays a key role as an opportunity cost against which the costs and benefits of responding are compared. In the previous chapter, we discussed how this signal can mediate the effects of motivation on habitual response rates. Here, we review evidence suggesting that the net reward rate is reported by tonic levels of dopamine in the striatum. This hypothesis unites psychologically and computationally inspired ideas about the role of tonic dopamine in the striatum, and explains, from a normative point of view, why higher levels of dopamine might be associated with more vigorous responding.

4.1 Introduction

Dopamine is perhaps the most intensively studied neuromodulator, due to its critical involvement in normal behaviors, including learning and performance in appetitive conditioning tasks, and also in a variety of abnormal behaviors such as addiction, electrical self-stimulation and numerous neurological and psychiatric disorders. Influenced particularly by the dramatic effects of pharmacological manipulations of dopamine neurotransmission on response rates, psychological theories of dopamine function have long focused on a

putative role in modulating the vigor of behavior. These theories attribute the vigor effects to a variety of underlying psychological mechanisms, including incentive salience (Beninger, 1983; Berridge & Robinson, 1998; Ikemoto & Panksepp, 1999), Pavlovian-instrumental interactions (Dickinson et al., 2000; Murschall & Hauber, 2006), and effort-benefit tradeoffs (Salamone & Correa, 2002). However, despite their psychological foundations, these theories do not, in general, offer a computational or normative understanding for *why* dopaminergic manipulations might exert such influence over response vigor.

Pharmacological and lesion studies show that elevated levels of striatal dopamine (for instance, as a result of administering agonists such as amphetamine) are first and foremost associated with enhanced responding (Jackson et al., 1975; Carr & White, 1987; Robbins & Everitt, 1996; Ikemoto & Panksepp, 1999; Giertler et al., 2003). Conversely, striatal dopamine depletion or antagonism profoundly reduces response rates (Sokolowski & Salamone, 1998; Aberman & Salamone, 1999; Salamone et al., 1999, 2001; Correa et al., 2002; Mingote et al., 2005). Figure 4.1a is representative of a host of results from the lab of Salamone, which show that lower levels of dopamine in the nucleus accumbens result in lower response rates. This effect of dopamine depletion seems more pronounced in higher fixed-ratio schedules, those requiring more work per reward. As a result of this apparent dependence on the response requirement, Salamone and his colleagues have hypothesized that dopamine enables animals to overcome higher work demands.

As discussed in Chapter 1, a different influential line of empirical and theoretical work on the involvement of dopamine in appetitive conditioning tasks arose from electrophysiological recordings of midbrain dopamine neurons in awake, behaving monkeys. These suggested a theory based on the framework of reinforcement learning, according to which the phasic (bursting and pausing) spiking activity of dopamine cells reports to the striatum a 'prediction error' signal (Ljungberg et al., 1992; Schultz et al., 1993; Schultz, 1998; Waelti et al., 2001; Bayer & Glimcher, 2005), which can be used efficiently both for learning to predict rewards and for learning to choose actions so as to maximize obtained rewards (Sutton & Barto, 1990; Friston et al., 1994; Barto, 1995; Montague et al., 1996; Schultz et al., 1997).

However, this computational framework suffers from three deficiencies that prevent it from providing a comprehensive picture of the role of dopamine in conditioned responding: First, as discussed in Chapter 2, the underlying reinforcement learning problem was formulated for discrete choice tasks, and so the theory is silent on the issue of the *strength* or *vigor* of responding. Bar the interesting exception of McClure et al. (2003), which we discuss later, reinforcement learning models of the dopamine system say nothing about the most obvious behavioral effect of pharmacological manipulations of dopamine, namely, their profound impact on response vigor. Second, this framework generally assumes that dopamine influences behavior only indirectly, by controlling learning (eg, Wickens, 1990; Wickens & Kötter, 1995). Although some behavioral effects of low-dose dopaminergic drug manipulations indeed emerge gradually, as if by learning (Wise, 2004), more immediate effects are seen with higher drug doses (or medial forebrain bundle stimulation; Gallistel et al., 1974), and effects on response rates have been shown in the absence of learning (Ohyama et al., 2000, 2001). Finally, whereas the unit recording data and associated computational theories are only concerned with the phasic release of dopamine, the tonic level of dopamine constitutes a potentially

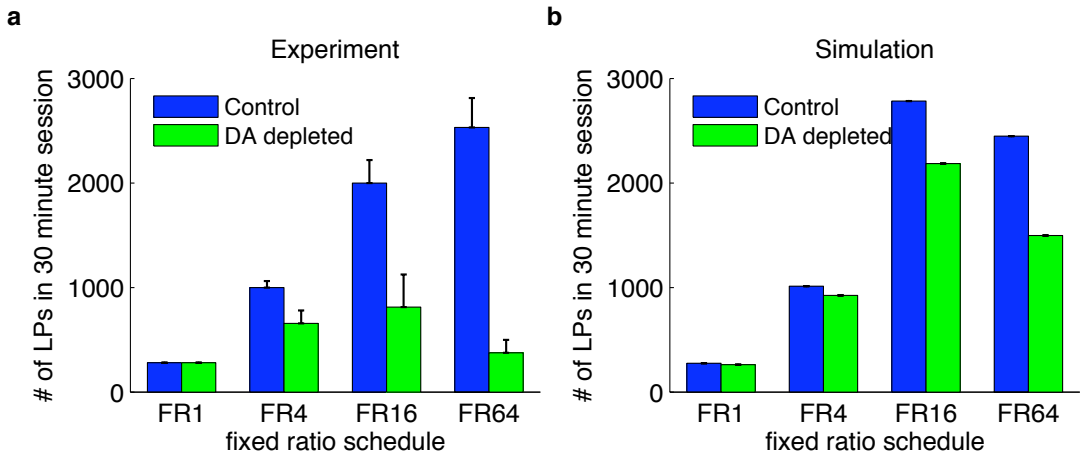

Figure 4.1: Effects of dopamine depletion on fixed ratio responding: **a.** Total number of leverpresses per session, averaged over five 30 minute sessions (error bars: s.e.m.), by rats pressing for food on different fixed ratio (FR) schedules. Rats with nucleus accumbens 6-hydroxydopamine lesions (green) press significantly fewer times than control rats (blue), with the difference larger for higher ratio requirements. Adapted from Aberman and Salamone (1999). **b.** Simulation of dopamine depletion: overall leverpress count averaged over five 30 minute sessions (error bars are negligible), for different fixed-ratio requirements. In blue is the control condition, in green is simulated dopamine depletion (attained by decreasing the net reward rate $\bar{R}$ by 60%). That the effects of the depletion appear more pronounced in higher schedules actually results from an interaction with time spent eating (see text). Simulation parameters were as in Chapter 2, bar the utility of the reward which was set at $U_r = 150$ to warrant responding on FR64.

distinct and carefully controlled channel of neurotransmission (Grace, 1991; Goto & Grace, 2005; Floresco et al., 2003; Bergstrom & Garris, 2003), for which a key role in enabling (Schultz, 1998) or energizing (Weiner & Joel, 2002) behavior has been suggested. Indeed, dopamine alterations affect a wide range of behaviors, many of which do not seem to be accompanied by phasic activity in dopamine cells. Furthermore, dopamine agonists can reverse many behavioral effects of dopamine loss (such as in L-dopa treatment of Parkinsonian symptoms), although they probably do not fully restore phasic dopaminergic transmission (Le Moal & Simon, 1991; Schultz, 1998). More directly, dopamine agonists, or artificial increases in dopamine level (eg, using amphetamine) have been shown to invigorate a wide range of behaviors (Lyon & Robbins, 1975; Evenden & Robbins, 1983; Taylor & Robbins, 1984, 1986; Ljungberg & Enquist, 1987).

Here, we suggest that the extended reinforcement learning model of response vigor in self-paced behavior that we have presented in Chapter 2, together with its links to the 'energizing' effects of motivation which we suggested in Chapter 3, can address these three lacunæ. By extending the formal normative account from the discrete action selection domain to the continuous domain of vigor or rate selection, our model highlighted the prominent role of the expected net rate of reward (which we designate $\bar{R}$) in determination of optimal response vigor. Recall that the net rate of reward exerts significant influence over overall response propensities, by acting as an *opportunity cost,* which quantifies the cost of sloth. That is, if the net reward rate is high, every second in which a reward is *not* delivered is costly, and therefore it is worth subjects' while performing actions more speedily even if the energetic costs of doing so are greater. In this chapter we argue on computational, psychopharmacological and neural grounds that this net reward rate may be

reported by tonic levels of dopamine, and show how it can account, without mediation through learning, for a wealth of reported effects of dopamine manipulations on response vigor in a variety of tasks. To complete the picture of dopaminergic control of normative response selection, we then consider how tonic and phasic dopamine signaling may interact.

4.2 Methods: Modeling dopamine in free operant response choice

The model we presented in Chapter 2 constitutes an extension of the classic reinforcement learning framework, so as to accommodate the choice not only of which action to perform, but also of response vigor, and thus model optimal response rates in free operant tasks. Specifically, our abstraction of the subject's optimization task directly paralleled the abstraction of discrete-choice tasks in standard reinforcement learning. This reliance on this well-studied framework was advantageous not only because reinforcement learning has a sound mathematical basis in the engineering theory of Dynamic Programming (Bertsekas & Tsitsiklis, 1996), but also because of its very close relationship with psychological accounts of behavioral learning (Sutton & Barto, 1981).

Our model is easily accommodated within the framework of temporal difference learning (see also Chapter 5), and naturally preserves the link between phasic dopaminergic signals and online learning of optimal values (which was discussed extensively in Chapter 1). Recall from Chapter 2 that the main differences between our model and classical reinforcement learning models of action selection are two: the choice of a response latency τ which accompanies the choice of an action a (which transforms the problem from a Markov Decision Process into a continuous time semi-Markov Decision Process), and the formulation of the problem in terms of optimizing the long run average rate of net utility per unit time $\bar{R}$ (rather than a discounted sum of future rewards). As a result, in our model the differential value of an (action,latency) pair at state S is defined as

$$Q(S,a,\tau) = R(S,a,\tau) - C(S,a,\tau) + V(S') - \tau \cdot \bar{R} \tag{4.1}$$

where we have here used shorthand notation with $R(\cdot)$ and $C(\cdot)$ representing immediate rewards and costs, respectively, and $V(S')$ is the differential value of the successor state to which the environment transitions after τ time passes and the chosen action a is completed.

This extension preserves fully the reliance of the learning process on a phasic prediction error signal, such as that suggested to be reported by dopamine (see Chapter 1, section 1.3.1). However, the classic prediction error signal (Chapter 1, section 1.2.1)

$$\delta = R(S,a,\tau) - C(S,a,\tau) + V(S') - V(S), \tag{4.2}$$

now must include an extra term due to the average reward formulation, namely, the subtraction of (or com-

parison to) the forfeited net rate of reward, ie, the opportunity cost of time:

$$\delta = R(S,a,\tau) - C(S,a,\tau) + V(S') - V(S) - \tau \cdot \bar{R} \tag{4.3}$$

(see also Chapter 5, section 5.1.1). Note that while after each state transition the immediate rewards and costs, as well as the value of the current and subsequent states, are expected to change, the new term $\tau \cdot \bar{R}$ does not depend on the current state: the net reward rate $\bar{R}$ is the same for all states[1], and τ is a property of the chosen response. Here we suggest that the baseline (or *tonic*) level of dopamine represents the constant part of this term, the net rate of reward $\bar{R}$.

Treatments of phasic dopaminergic firing as conveying a prediction error signal (such as δ in equation 4.2 above) have, in fact, always assumed a relatively constant baseline rate of dopaminergic firing around which the *signed* prediction error signal can be conveyed (eg. Montague et al., 1996). That is, a firing rate below the baseline rate (ie, a pause in firing) has been taken to signal a negative prediction error, while a burst of firing at a higher than baseline has been suggested to reflect a positive prediction error (Bayer & Glimcher, 2005). Because these models have most frequently been concerned with the electrophysiologically recorded firing patterns of dopamine neurons, the baseline *firing rate* has been taken as the reference point below which signals can constitute a negative prediction error. However, in the target areas to which the signal is conveyed, it may be more appropriate to consider the basal concentration of the neurotransmitter itself (especially as dopamine release is nonlinearly related to neuronal firing rate, eg, Gonon, 1988; Montague et al., 2004, and because the basal dopamine concentration is not directly related to the baseline firing rate, see below). This has been termed the *tonic level of dopamine*.

What we are proposing here is that this basal concentration of dopamine in target areas actually represents the net rate of reward. From the viewpoint of reinforcement learning, a net reward rate baseline would make the extension to learning via the prediction error in equation (4.3) straightforward (given proper temporal integration). From the viewpoint of the role of tonic dopamine, we are suggesting that it plays a new computational role: the tonic level of dopamine is not just a 'zero' point below which phasic signals are considered negative, it also carries a precise meaning as the expected or measured net reward rate. This implies that different tonic levels of dopamine should normatively correspond to different rewarding situations (and different behavioral policies), as will be discussed below.

Specifically, to model the effects of dopamine manipulations (such as dopaminergic lesions, or the pharmacological administration of dopamine antagonists or agonists, which predominantly affect tonic levels of dopamine), we assume a change in the net rate of reward $\bar{R}$, independent of any change in the task contingencies. The effects of this manipulation on the optimal behavioral policy are not completely straightforward: the consistency equations of the model (an equation such as 4.1 for each action, latency and state), from which we derive the differential values $Q(S,a,\tau)$ in order to find the optimal policy, have a unique solution which specifies both $\bar{R}$ and the differential Q-values such that they are mutually consistent. That is, the set of equations defining the relationships between the different Q-values can only be satisfied for a specific

[1] The net rate of reward is the same for all states because of the unichain assumption – see Chapter 2.

value of $\bar{R}$, which represents the true net reward rate that the optimal policy would achieve. Once $\bar{R}$ is extrinsically perturbed from this value by $\Delta\bar{R}$, the set of equations is no longer solvable. Because the net reward rate influences the Q-values only through the term $(-\tau \cdot \bar{R})$, we modeled the (immediate) effect of an extrinsic perturbation of $\bar{R}$ by simply adjusting the previously computed Q-values to the new net reward rate according to $Q^{new}(S,a,\tau) = Q^{old}(S,a,\tau) - \tau \cdot \Delta\bar{R}$. Actions were then chosen as usual, using a soft-max probability distribution computed from these new Q-values.

4.3 Results: The net rate of reward and tonic dopamine

The policy-dependent net reward per unit time plays a critical role in the selection of vigor in our model. This is because the choice of a latency τ with which to perform an action commits this time to this action exclusively. The opportunity cost of this commitment is $(\tau \cdot \bar{R})$, since this much net reward *could* have been earned by following the 'default' policy (the one that has lead to $\bar{R}$ up to now) rather than the current action in these τ seconds. Thus, when choosing actions and latencies, the cost-benefit tradeoff must include the opportunity cost (together with the other vigor and unit costs of responding) on the side of the costs. If, for some action and latency, the overall cost is smaller than the expected reward, then performing this action at this latency is worthwhile. In this way, the net reward rate effectively introduces competition between different latencies of responding (Dragoi & Staddon, 1999).

In Chapter 2 (section 2.3.3) we analyzed the optimal policy in standard operant reinforcement schedules and showed that the optimal latency of *all* actions is *inversely proportional to the net reward rate*. This means that when the net reward rate is higher, optimal responding will be faster, and conversely, when the reward rate is lower, responding will be slower. We therefore posit, on computational grounds, a tonic, slowly changing, net reward signal that should exert a generalized form of control over response vigor. In Chapter 3 we showed that motivational states are a central determinant of the net rate of reward, which links this form of control to the 'energizing' (or 'generalized drive') effects that motivational states exert on all pre-potent behavior. Linking tonic dopamine to the net reward rate thus explains why tonic dopamine is related to the energizing of behavior (Weiner & Joel, 2002). According to our hypothesis, a higher tonic level of dopamine represents a situation akin to higher drive, in which behavior is more vigorous, and lower tonic levels of dopamine cause a general slowing of behavior.

Using this hypothesis we can now explicitly model the cost-benefit tradeoff experiments pioneered by Salamone and his colleagues (Salamone & Correa, 2002). In the free-operant variant of these, it has been shown that 6-hydroxydopamine lesions in the accumbens have minimal effects on responding on low fixed ratio (FR) schedules, while severely reducing responding on high FR schedules (Figure 4.1a; Aberman & Salamone, 1999; Salamone et al., 2001; Mingote et al., 2005). Figure 4.1b shows results from our model, with dopamine depletion simulated by reducing the net reward rate $\bar{R}$, while leaving all other aspects of the model intact (see Methods above). A similar pattern of results is seen, with dopamine depleted rats pressing less than control rats. This arises since the optimal latencies for leverpressing are longer once tonic dopamine

reports a lower expected net reward rate, thus fewer presses are performed throughout the 30 minute session.

Note, however, that although the effects of depletion seem less pronounced in lower ratio schedules, in the simulation this is due to a larger proportion of the session time spent eating in lower FR schedules (at a consumption speed that is unaffected by dopamine depletions; Sokolowski & Salamone, 1998; Aberman & Salamone, 1999; Salamone et al., 2001; Salamone & Correa, 2002), and not from a smaller effect of depletion on leverpress latencies. In the model, tonic dopamine depletion causes longer leverpress latencies in *all* schedules. However, in a schedule such as FR1, in which the rat performs several hundreds of leverpresses and is rewarded with several hundred pellets, the majority of the session time is spent consuming rewards rather than leverpressing for them. By comparison, in the FR32 condition, the rat presses over a thousand times and only obtains several tens of pellets, thus effects of the dopamine depletion treatment on leverpressing seem more prominent. The use of a global measure of number of responses in a session as a measure of response rate proves misleading as it confounds the choice of latency with the choice of alternative actions (such as eating). Thus, according to our model, dopamine not only allows animals to cope with higher work requirements, but rather is important for optimal choice of vigor at *any* work requirement.

4.3.1 Tonic and phasic dopamine

We have so far not specified a neural (or computational) mechanism by which tonic dopamine levels come to match the expected net reward rate. One simple computational truth is that if the phasic responses of dopamine neurons indeed report a prediction error for future reward, their integration over time should, by definition, equal the net reward rate obtained. Moreover, because the phasic signal corresponds to predicted (and not only obtained) reward, online integration of these signals will allow the net reward rate to be *predictive*. For instance, upon being unexpectedly put in an operant chamber where high utility food was previously experienced, the large prediction error would immediately elevate the tonic level of dopamine, inducing rapid behavior even before rewards are obtained. This is different from Killeen's suggestion that response vigor is related to an ongoing average of obtained rewards, which must accumulate anew at the beginning of each experimental session (Killeen et al., 1978; Killeen, 1998). Such a predictive aspect of the net reward rate is key to producing immediate effects of motivational shifts, even in extinction (ie, without experiencing any rewards). In the next chapter we use exactly such a mechanism to estimate net rates of reward in our online learning implementation.

Indeed, phasic dopamine signals are discernible outside the synaptic cleft (Phillips & Wightman, 2004; Roitman et al., 2004), so if tonic dopamine concentrations were solely determined by the slow accumulation of dopamine from phasic events filtered by reuptake (inducing, for instance, an exponential decay which would allow exponentially weighted averaging, see eg, Daw & Touretzky, 2002; Killeen & Sitomer, 2003), they could directly measure, throughout behavior, the long-term cumulative net reward signal we posit. However, we note here that, from both a neural and a computational perspective, a view of the tonic signal as just the running average of the phasic signals is probably incomplete. Computationally, it would be advantageous to

decouple the signals such that a learned mapping from, say, a specific context to the expected net reward rate could take effect without relying on phasic prediction errors (which are more responsive to precisely timed events than to diffuse changes in context). Neurally, there is physiological evidence that the two modes of dopamine transmission are indeed somewhat decoupled, with phasic signals resulting from bursting activity, and tonic dopamine levels determined mainly by the overall percentage of active (non-silent) dopaminergic neurons and by presynaptic glutamatergic inputs (Chéramy et al., 1990; Chesselet, 1990; Grace, 1991; Floresco et al., 2003; Lodge & Grace, 2006; although these two modes of activity also interact, Phillips et al., 2003; Lodge & Grace, 2005). Moreover, afferents of the ventral tegmental area (eg. the pedunculopontine nucleus and the ventral pallidum, respectively) appear to affect either bursting activity or population activity in dopamine neurons (Floresco et al., 2003; Goto & Grace, 2005; Lodge & Grace, 2006), providing a mechanism for independent modulation of phasic and tonic dopamine levels.

Note that since pharmacological manipulations of dopamine are likely to affect both tonic and phasic signaling, their effects on behavior can be subtle to tease apart. This can be illustrated by considering responding in Salamone and colleagues' cost-benefit T-maze task (Figure 4.2a; Cousins et al., 1996; Denk et al., 2005). In this task, a rat can obtain a large amount of reward (say, four food pellets) by choosing one arm of a T-maze, or a smaller amount (say, two food pellets) if it selects the other arm. However, the highly rewarding arm is partly blocked by a barrier which the rat must scale in order to reach the reward. Hungry rats typically choose the high-reward/high-effort arm in the majority of trials. In contrast, after nucleus accumbens dopamine depletion, they prefer the low-reward/low-effort arm (but only if there is some chance of earning rewards in that arm, see Figure 4.2b). We suggest that this reversal in discrete action propensities is due to *learned* effects on choice preferences, mediated by the phasic dopamine signal (cf. Walton et al., 2006). For example, if the phasic signal is blunted by the drug, this would reduce the efficacy of the four-pellet reward signal, making it, say, equivalent to the learning signal that would normally result from receiving only two food pellets. Of course the reward signal for the low-reward arm would also be blunted, say to the equivalent of one pellet. In this case, although the two-pellet difference in reward prior to the lesion was sufficient to justify the extra cost of scaling the barrier, the one-pellet difference after dopamine depletion might not, thereby altering the rats' choice toward the low-rewarding arm within a few trials of learning.

At the same time, we would expect reduced *tonic* dopamine to reduce the reported 'opportunity cost' of slower responding without affecting other aspects of the task. Thus, prior to new learning with blunted phasic signals, rats should still be willing to climb the barrier for four food pellets, however, they need not hurry to do so. As a result of this separation of tonic and phasic effects, we predict that transiently, for instance in the first post-depletion choice trial, dopamine depleted rats should maintain their preference for the high-effort/high-reward arm, albeit acting distinctly more slothfully. Even as the phasic-induced learning effects accumulate and produce a shift in discrete action choice, the tonic effects should persistently promote slower responding. Indeed, Denk et al. (2005) showed that treatment with the dopamine antagonist haloperidol significantly lengthened the latencies (from 1.6 seconds to 6 seconds on average) to reach the reward on trials in which the high-effort arm was chosen.

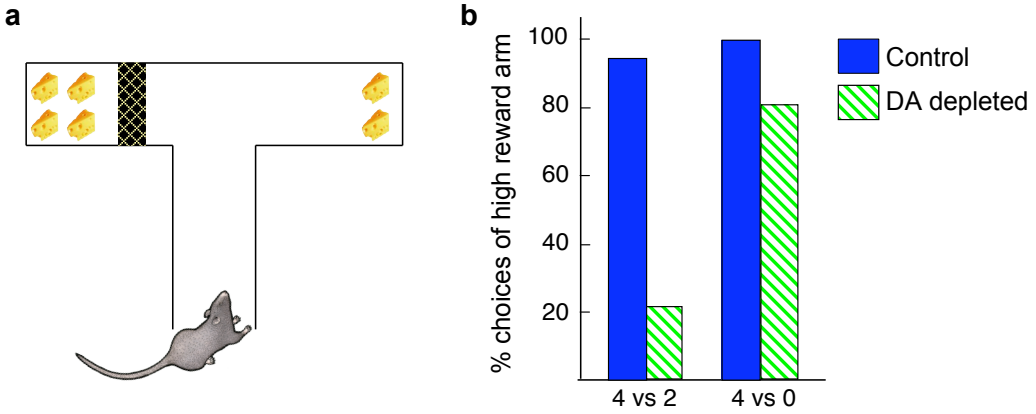

Figure 4.2: a. The T-maze cost/benefit task developed by Salamone and colleagues: At the start of a trial the rat is placed at the base of the maze, from which it can choose either of two arms. One arm leads to two food pellets, while the other is more rewarding (four pellets), but is also blocked by a scalable barrier. After the rat chooses an arm and consumes the reward, the trial ends and the rat is removed from the maze. Some inter-trial-interval elapses before the rat is placed in the start box again, and another trial begins. **b.** Rats typically choose the high rewarding arm, despite the effort that climbing the barrier necessitates (blue). The left bars show results from an experiment in which the high-effort/high-reward arm rewarded with four food pellets, and the low-effort/low-reward arm was baited with two pellets (4 vs 2), while the right bars are from an experiment in which the low-rewarding arm was empty (4 vs 0). After dopamine depletion, the green (striped) bars show that rats are no longer willing to climb the barrier, and prefer the low-effort/low-reward arm. However, this is only true if some reward can be harvested in the low reward arm. When the only reward available is that from the high-effort arm, dopamine-depleted rats continue choosing this option, demonstrating that their preference for the low-reward arm in the 4 vs 2 case was not caused by a simple treatment-induced inability to climb the barrier. Rather, the results suggest that the altered response pattern is the outcome of a dopamine-depletion-induced alteration of the cost-benefit tradeoff. Interestingly, the few choices of the non-rewarding low-effort arm by dopamine depleted rats completely disappeared by the second week of testing, as if by learning. Adapted from Cousins et al. (1996).

A final potential interaction between tonic and phasic aspects of dopamine is the finding that responding to cues predictive of higher reward is typically faster than responding to less valuable cues (eg. Hauber et al., 2000; Watanabe et al., 2001; Takikawa et al., 2002; Lauwereyns et al., 2002a; Schoenbaum et al., 2003). Although we associate vigor selection with tonic rather than phasic dopamine, electrophysiological recordings have shown a linear relationship between reaction times and phasic dopaminergic responding (Satoh et al., 2003; see also Roitman et al., 2004). The suggestion that phasic prediction errors affect tonic dopamine levels through integration (Phillips & Wightman, 2004; Roitman et al., 2004; Wise, 2004) can explain this correlation. Larger phasic prediction-error signals for stimuli previously associated with higher rewards (Fiorillo et al., 2003; Tobler et al., 2005) would thus result in faster responding to these cues. Interestingly, in contrast to the general effect of tonic dopamine on response times, which seems to be mediated by the nucleus accumbens (eg. Salamone & Correa, 2002; Salamone et al., 2007), reward-related reaction-time differences are not dependent on dopamine transmission in the nucleus accumbens (Brown & Bowman, 1995; Amalric & Koob, 1987; Hauber et al., 2000; Giertler et al., 2003, 2004), but are perhaps related to dopamine in the dorsal striatum (Amalric & Koob, 1987). Note that such effects of 'incentive motivation' for the outcome (Dickinson & Balleine, 2002; McClure et al., 2003; Berridge, 2004) might also involve temporal discounting (which we have not directly modeled here, see Chapter 2, section 2.2), by

which delayed rewards are viewed as less valuable than proximal ones. This is because the additional value of receiving a larger reward faster could offset the cost of a more vigorous response.

4.4 Discussion

The idea that dopamine controls response vigor is not new in the psychological literature. However, previous accounts have remained at the level of identifying a causal link between dopamine and response rates, unable to explain *why* this should be so. Here, extending the implications of the reinforcement learning model of free operant action selection that we proposed in Chapter 2 to the neural domain, we have suggested that tonic levels of dopamine in basal-ganglia structures encode the experienced (or expected, see below) net rate of rewards. This explains, from a normative stance, why dopamine plays a critical role in determining the vigor of responding, and provides a route by which dopamine could mediate the effects of motivation on response vigor. Our hypothesis dovetails neatly with computational theories suggesting that the phasic activity of dopamine cells reports appetitive prediction errors, and psychological theories about dopamine's role in energizing responses.

Response vigor and dopamine's role in controlling it have appeared in previous reinforcement learning models of behavior (Dayan & Balleine, 2002; McClure et al., 2003), but only as fairly ad-hoc bolt-ons. For instance, McClure et al. (2003) constructed a continuously varying response vigor (running speed in a maze) from a series of binary decisions of whether to respond or to do nothing. This allowed them to incorporate effects of (phasic) dopamine on response vigor, however, due to limitations on how vigor costs can be accounted for in a Markov Decision Process, their model but did not license the sort of analysis of the tradeoff between response effort and benefit on which we have focused here. In our semi-Markov framework, the optimization of vigor is treated as a reinforcement learning problem in itself, which allows a clear formulation of the cost-benefit tradeoff determining optimal vigor, and suggests a natural dopaminergic substrate for the opportunity cost of time. Our model puts in center stage the tradeoffs between costs and benefits that are explicit in the so-called cost-benefit T-maze tasks (Cousins et al., 1996), are implicit in free operant tasks, and are manifest in day-to-day decisions about the vigor of actions. To further account for immediate (unlearned) effects of motivational manipulations, the main assumption we make is that tonic levels of dopamine can be sensitive to *predicted* changes in the net rate of reward that are occasioned by changes in the motivational state, and that behavioral policies are in turn immediately affected. Embedding such sensitivity in a temporal-difference reinforcement learning system (see Chapter 5) can produce flexible adaptation of the vigor of habitual responding.

4.4.1 Predictions

Our theory makes several readily testable predictions. First, we predict that leverpressing latencies will be affected by dopamine depletions of the sort employed by Salamone and colleagues, even in schedules such

as FR1 that require less effort per reward. This effect, however, may not be straightforward to measure, as a 'molecular' measure of response latency is needed, rather than the 'molar' measure of number of responses in a session. Indeed, a more detailed reaction time analysis in Mingote et al. (2005) points in this direction. One simple way to eliminate the interaction with eating time, would be to test effects of dopamine depletion during extinction. This would also nicely separate immediate effects of changes in tonic dopamine levels from those of new learning due to a diminished phasic signal (see below), but albeit potentially at the expense of an interaction between dopamine depletion and extinction learning. Alternatively, high order conditioning schedules could be used to look at responding for conditioned stimuli, thereby eliminating the interference of rewards without inducing extinction.

As discussed in Chapter 3, we also predict similar effects of changes in motivational state. In particular, the higher the state of deprivation, the shorter the latency of all actions should be. Here, again, it is important to use a 'molecular' measure of response latency to distinguish the effects of satiety on response rates from its effects on eating time (which, in this case, do appear to be significant; Aberman & Salamone, 1999, but see Robbins & Everitt, 1996). Moreover, we predict that tonic levels of striatal dopamine will be higher in a deprived state than in a sated state (as also suggested by Weiner & Joel, 2002), given that the animal can indeed expect a higher net reward rate in the former (that is, given that relevant outcomes are available). Although difficult to measure directly, there is some supportive evidence for this (Wilson et al., 1995; Hernandez et al., In Press).

4.4.2 Immediate *vs.* learned effects

Previous reinforcement learning models have mostly concentrated on how phasic dopamine can affect behavioral preferences gradually and indirectly through a learning process. By contrast, we have modeled steady-state behavior in a well-learned task, and focused on explaining how a change in tonic dopamine, caused either pharmacologically or by a change in motivational state, can also affect behavior *immediately*, without requiring learning. Our hypothesis that ongoing measured levels of dopamine report the net reward rate, suggests a neural implementation for the suggestion in Chapter 3 that a change in expected net reward rate can immediately affect the rate of responding. The idea is that the neural action-selection system can take advantage of the fact that a higher net reward rate (arising, for instance, from a shift from satiety to hunger) will optimally be associated with more vigorous optimal responding. It can thus adjust response vigor directly on the basis of the tonic dopamine reported signal, even before learning new Q-values that correspond to the new situation. Importantly, such a mechanism provides some flexibility in rapidly adapting the overall rate of behavior to changes in circumstance that influence expected net reward rates.

Of course, the decision of how vigorously to respond is only one of the twin decisions underlying behavior in our framework. As discussed in Chapter 3, the decision as to which action to perform in a new motivational state is more difficult to adjust, as it requires re-estimating or relearning the values of different actions. In reinforcement learning models of the sort we have considered, relearning involves additional training

experience and utilizes the *phasic* dopamine signal. Thus, for instance, if a rat leverpressing for food is shifted from hunger to thirst, new experience (and learning mediated by phasic dopamine) will be necessary in order to direct responding to a different action to receive water. This complicated combination of direct motivational sensitivity (of vigor, through the tonic dopamine signal) and insensitivity (of choice, as a result of required learning), turns out to match well the characteristics of habitual behavior (Chapter 6; Dickinson, 1985; Dickinson & Balleine, 2002; Niv et al., 2006), to which this type of 'value-caching' reinforcement learning framework pertains. Habitual behavior, moreover, is indeed associated with dopamine and the striatum (eg. Yin et al., 2004; Faure et al., 2005).

We have not considered in this chapter the anatomically and psychologically distinct category of 'goal-directed' behaviors (Dickinson & Balleine, 1994) whose pattern of immediate and learned motivational sensitivity is rather different, and to which another class of 'forward-model' based reinforcement learning models is more appropriate (see Chapter 1; Daw et al., 2006). Although our current model addresses habit-based instrumental control, optimizing the vigor of responding is as much an issue for goal-directed instrumental control, and indeed for Pavlovian actions, and it is possible that a dopamine-reported net rate of reward plays a part in determining vigor for these as well.

We also have not yet treated the learning of a new task (which will be discussed in the next chapter), but concentrated only on the steady-state situation. Note, though, that it is at this stage, by which responding is nearly optimal with respect to the reinforcement schedule and task, that we may analyze the optimal relationship between reward rate and response vigor, and that these variables might stably be measured experimentally. In contrast, the learning stage is characterized by progressive (and likely complex) changes both in behavior and in the net reward rate obtained. As will be shown in Chapter 5, over the course of learning the animal must continually estimate the net reward rate from recently obtained rewards and costs. We predict that this will cause the tonic level of dopamine to change dynamically, in general, increasing as the subject learns the contingencies and optimizes responding, obtaining more rewards while reducing action costs. Higher net reward rates reported by higher levels of tonic dopamine, will then further enhance response rates in a feedback cycle. Realistically, however, even in a well learned task the net reward rate and the rates of responding may not be perfectly stable — for instance, during a session, both would decline progressively as satiety reduces the utility of obtained rewards.

4.4.3 Future directions: Reward *vs.* punishment; instrumental *vs.* Pavlovian

Among the most critical issues left for future work is the role of dopamine, and specifically, the role of tonic dopamine in aversively motivated conditioning. Daw et al. (2002) suggested that there is an opponency between serotonin and dopamine in controlling appetitive and aversive conditioning. This was based on data showing various forms of antagonism between these neuromodulators (eg. Fletcher & Korth, 1999), and on long-standing psychological ideas (Konorski, 1967; Solomon & Corbit, 1974; Dickinson & Balleine, 2002) that there exist two opponent motivational systems (see Chapter 1). Their model suggested that, since phasic

dopamine appears to reports appetitive prediction errors, phasic serotonin should report aversive prediction errors. This was construed in the context of an average-reward reinforcement learning model, rather like the one we have discussed here. Moreover, Daw et al. (2002) suggested that opponency also extended to the tonic signals, with tonic dopamine representing the rate of *punishment* (inspired by microdialysis data suggesting that dopamine concentrations rise during prolonged aversive stimulation), and tonic serotonin, conversely, reporting the reward rate.

Our suggestion to associate tonic dopamine with net rate of *reward* rather than punishment seems to reverse this prior mapping. However, the two views may not be as disparate as at first seems. For instance, in active avoidance tasks, responding is known to be under dopaminergic control. In this case, there is an analogous form of opportunity cost that forces fast avoidance, coming from the possibility of failing to escape a punishment. This link between net rate of punishment and vigor could potentially be realized by the same dopaminergic substrate as the appetitive energizing we have discussed (see also Ikemoto & Panksepp, 1999). In this sense, the opportunity cost relevant to determining response vigor should comprise both the net rate of (instrumental) rewards and the net rate of avoidable punishments. In fact, it may be more accurate to relate tonic dopamine to the net rate of *all motivationally significant instrumental outcomes*, in their absolute value (that is, summing up rewards and punishments rather than these canceling each other).

This, however, brings up another complication: although we have modeled instrumental behavior, ie, behavior that is performed in order to obtain rewards (or avoid punishments), the types of phasic dopaminergic prediction errors we have discussed here are also emitted in response to Pavlovian stimuli. Consequently, the tonic level of dopamine may also include the contributions of Pavlovian unconditional stimuli to the overall reward (and punishment) rate, which would prove problematic for the optimal control of vigor. As discussed in Chapter 2 (section 2.3.7), Pavlovian rewards that are given irrespective of any action performed by the animal, should *not* affect the optimal vigor of instrumental actions. Indeed, in the consistency equation defining Q-values, these 'free' rewards are separately accounted for and thus cancel out their counterpart in the net reward rate $\bar{R}$, resulting in the correct choice of vigor (that is unaffected by 'free' rewards). Thus the determinant of response vigor is still the total rate of instrumental outcomes, separated from the rate of Pavlovian events. However, this separation is hardly trivial to implement — not only does it necessitate two separate accounts of reward rate, but it assumes that rewards can be easily classified as to whether they were earned instrumentally or not, which is, in fact, is a tough inference problem on its own.

Assuming that such a separation does not occur, or is incomplete, instrumental response vigor proportional to a net reward rate that includes non-instrumental rewards, would be suboptimal. Indeed, as discussed in Chapter 2, Pavlovian-instrumental transfer (PIT) is one example of such suboptimal response choice. In this, the appearance of a previously-trained Pavlovian cue, which signals the upcoming availability of rewards regardless of any actions that the animal performs, has been shown to exert an invigorating effect on ongoing instrumental actions, such that a rat leverpressing for food will press faster once the Pavlovian cue is turned on (Estes, 1948; Lovibond, 1983). Of course, this change in vigor is suboptimal, because the cue does not signal any change in the instrumental reward schedule, meaning that the the optimal leverpress

rate should not change. PIT, at least in its general form (in which a cue effects instrumental responding even for different rewards; Holland, 2004) is, in fact, dopamine dependent (Dickinson et al., 2000; Wyvell & Berridge, 2000). In any event, whether PIT is indeed due to a cue-induced transient elevation of tonic dopamine (which would demonstrate the suboptimality of using tonic dopamine levels to determine response rate in a mixed instrumental-Pavlovian setting), how (and whether) such phasic Pavlovian and instrumental signals determine the tonic level of dopamine, and the hypothetical joint dopaminergic and serotonergic coding for appetitive and aversive signals present burning empirical questions.

A final pressing question has to do with the neuronal locus of the cost/benefit computations: in what brain area(s) are the different sources of information (immediate cost, potential immediate reward, expected future reward and opportunity cost) combined in order to decide which action to take and with what latency? It is tempting to think that a reinforcement learning mechanism contained in the basal-ganglia can perform this computation through simple temporal difference learning using dopaminergic signals. However, some factors in this computation, such as the immediate action costs, may be dopamine independent. Further, recent evidence (reviewed in Walton et al., 2006) implicates the anterior cingulate cortex (an area which receives dopaminergic projections from the midbrain, and innervates both dopaminergic neurons and neurons in the nucleus accumbens) in negotiating cost/benefit tradeoffs. As with dopamine depletion, lesions to the anterior cingulate cortex of rats performing the explicit cost/benefit T-maze (Figure 4.2a), alter the rats' preference away from the high-reward barrier-obstructed arm, unless the barrier is low enough or the reward high enough to warrant the extra effort (Walton et al., 2002, 2003; Rushworth et al., 2005). Furthermore, this effect emerges gradually, as by a learning process (Walton et al., 2006). Interestingly, despite the behavioral similarity of dopaminergic and cingulate involvement in this task, targeted dopamine lesions of the anterior cingulate cortex had no effect on preferences in the T-maze task, compared to unlesioned controls. This perhaps points to a top-down influence of the anterior cingulate cortex on dopaminergic neurons, rather that the other way around (Walton et al., 2006).

4.4.4 Conclusions

In conclusion, in this chapter we have leveraged the model presented in Chapter 2 to suggest a computational acount of striatal dopamine which incorporates both tonic and phasic dopamine signals into an action selection framework emphasizing both the identity of the chosen action, and the vigor of its execution. Our account emphasizes the non-binary nature of cost-benefit tradeoffs with which animal and humans are continuously faced, as the decision on action vigor (or latency) embodies a continuous valued decision regarding how much effort to exert given the available benefits. A slowly changing continuous valued tonic dopamine signal that represents the expected net rate of reward (and the opportunity cost of time) is exactly what is necessary to compute this tradeoff. Our framework thus makes another step in the direction of clarifying the role of striatal dopamine and its effects on behavior, even as it opens the door to a wealth experimental work that can quantify the precise interplay between cost and benefit, and between tonic and phasic dopamine.

In theory, there is no difference between theory and practice, but not in practice
– Jan L. A. van de Snepscheut

Chapter 5

Online learning of optimal response rates

Abstract: In Chapter 2, we presented a novel reinforcement learning model in which animals optimally chose both which action to perform and with what latency to perform it. The steady state results of this model were shown to correspond to the known characteristics of free-operant responding, and the implications of the model for the effects of motivation on response rates and the control of response rates by tonic levels of dopamine were subsequently discussed in Chapters 3 and 4, respectively. However, reinforcement learning models of animal learning are appealing not only because they provide a normative basis for decision-making, but also because they show that optimal action selection can be learned through online incremental experience with the environment and using only locally available information. As such, they lend themselves directly to a neural implementation, and indeed, neural substrates have been identified for key computational reinforcement learning constructs. Here, we complete the discussion of optimal learning of response rates by presenting such an online learning algorithm for our model. There are two major differences between learning in our model and previous online reinforcement learning algorithms: First, most prior applications have dealt with discounted reinforcement learning while we use average reward reinforcement learning. Second, unlike previous models that have focused on discrete action selection, the action space in our model is inherently continuous, as it includes a choice of response latency. We thus propose a new online learning algorithm that is specifically suitable for our needs. In this, building on the experimental characteristics of response latencies, we suggest a functional parameterization of the action space that drastically reduces the complexity of learning. Moreover, following from our analysis in Chapter 2 that showed that the net rate of reward is an important determinant of optimal response rates, and our hypothesis in Chapter 3 that the net reward rate mediates the effects of motivation on habitual behavior, we suggest a formulation of online action selection in which response rates are affected by the net reward rate. We show that our algorithm learns to respond with nearly optimal latencies, and discuss its implications for the differences between learning of interval and ratio schedules.

5.1 Introduction

In Chapter 2, we presented a reinforcement learning framework for modeling free operant behavior. In the model, we treated the choice of response rate as an optimization problem: given the characteristics of the instrumental task and the (hypothesized) costs of behavior, what is the optimal rate of responding? To answer this question, we sought to compute the long-run net benefit of choosing to perform each available action at each possible latency, for every state of the semi-Markov decision problem. Given such values, a policy that chooses (at each state) the action and latency with the highest value is optimal in terms of maximizing long-run benefits. We used the framework of average reward reinforcement learning (ARL) to formulate a recursive definition of the long-run differential value of a response that is based on the immediate expected costs and rewards for this response, plus the expected long-run differential value of the future state, minus the net reward rate forfeited during the time devoted to performing the response. Using Dynamic Programming techniques we solved the resulting system of consistency equations for the optimal differential state-action-latency-values, and derived from these values the optimal steady-state behavior.

However, as discussed in Chapter 1, standard Dynamic Programming techniques are 'model-based' because they require a model of the underlying dynamical system (Barto et al., 1989). That is, they require knowledge of $P_r^{a,\tau}(S)$, the probability of receiving an immediate reward of utility $U_r \geq 0$ when performing action $a \in \mathcal{A}$ with latency $\tau > 0$ in state $S \in \mathcal{S}$, and $\mathcal{T}_{S\to S'}^{a,\tau}$, the probability of transitioning to state $S' \in \mathcal{S}$ given the current state and the choice of action and latency, for every action, latency, and state[1]. In any real-world free-operant task, it is not realistic to assume that the animal has such *a-priori* knowledge of the dynamics of the environment. Rather, these quantities must be estimated from experience. RL techniques do exactly this. These methods use samples of the transitions and rewards observed during (possibly stochastic) interaction with the environment, in order to estimate a model of the environment and use it to search for the optimal behavioral policy (in 'model-based' RL), or to learn long-run state- or state-action values and derive the optimal policy directly (in the case of 'model-free' RL; Sutton & Barto, 1998). In this chapter, our goal is to provide an online 'model-free' algorithm for learning the optimal policy in a free-operant task, ie, for learning to choose actions and latencies such that maximize the net reward rate obtained, based only on information gained from experience with the task.

While online 'model-free' RL algorithms have been developed for a large variety of problems (Sutton & Barto, 1998), two characteristics of our problem preclude using these learning algorithms as is. One is the use of average reward RL, and the other is that response selection in our model includes the choice of latency, which means that the action space is continuous. In the following (Section 5.1.1), I will first describe in more detail why these two characteristics of our model require a special solution. I will then state the constraints that a free-operant online learning algorithm should adhere to (that is, the desiderata from such an algorithm). Because no previously suggested algorithm exactly matches these, in Section 5.2, I will present a new algorithm for online average-reward reinforcement learning of action and latency choice,

[1] These methods also require knowledge of the costs of different actions, however, we consider these a property of the animal rather than a property of the task. Thus, we assume in the following that the learning mechanism knows the action costs, although we note that these can also be sampled and learned from the actual experienced costs.

which borrows and combines ideas from previous algorithms, and, using some approximations, results in efficient online learning of optimal free-operant behavior. In Section 5.3, I will show that this algorithm indeed learns a behavioral policy that is sufficiently close to the optimal policy we derived in Chapter 2. I will also use the results of the model to discuss the differences between interval and ratio schedules, in terms of the learning dynamics and the steady-state solution. Finally, in section 5.4, I will briefly discuss previous related work, finishing with the limitations of our proposed algorithm, and directions for future work.

5.1.1 The problems

Let us first consider the simple case for which online learning algorithms have been well worked out: discounted RL with discrete action selection (ie, no selection of latencies) according to a fixed policy, and a small set of possible actions and states. Recapping from Chapter 1, the discounted RL consistency equation for state values in this case is:

$$V^{\pi}(S) = \sum_{a\in\mathcal{A}} \pi(a|S) \left[P_r^a(S)\cdot U_r + e^{-\gamma} \sum_{S'\in\mathcal{S}} \mathcal{T}^a_{S\to S'} \cdot V^{\pi}(S') \right] = \langle U_r \rangle_{P_r,\pi} + e^{-\gamma} \left\langle V^{\pi}(S') \right\rangle_{\mathcal{T},\pi} \tag{5.1}$$

where $e^{-\gamma}$ is the discount factor and the expectations are with respect to the dynamics of the environment and the behavioral policy (see also equation 1.8). One way to learn the long run state values $V^{\pi}(S)$ is by using the fact that this consistency equation is fulfilled only by the *true* values. If we estimate these values incorrectly, there will be an inconsistency between consecutive state value estimates (designated $\tilde{V}$)

$$\delta = \langle U_r \rangle_{P_r,\pi} + e^{-\gamma} \left\langle \tilde{V}^{\pi}(S') \right\rangle_{\mathcal{T},\pi} - \tilde{V}^{\pi}(S). \tag{5.2}$$

This *prediction error term* δ can be used to update the value estimates according to

$$\tilde{V}^{\pi}(S) \longleftarrow \tilde{V}^{\pi}(S) + \eta \cdot \delta \tag{5.3}$$

where $0 < \eta \leq 1$ is a learning rate or step-size. Such an *error-correcting learning process* will terminate only when the estimated values are self-consistent, and therefore are the true long-term state values.

The above learning process is 'model-based' because computing the expectations in equation (5.2) necessitates knowledge of the expected probabilities of immediate rewards and state transitions. However, it is easy to devise a related learning process which does not rely on such *a-priori* knowledge, and can be implemented online without any averaging. In order to use stochastic experience with the environment to learn the true values $V^{\pi}(S)$ by 'model-free' temporal difference (TD) methods, all that we need to do is to choose actions according to policy π at every timestep t, sample $\langle U_r \rangle$ and $\langle V^{\pi}(S') \rangle$ by observing the rewards and transitions that actually take place in the world as a result of these choices, and use these to compute stochastic samples of the inconsistency between the values of consecutive states (Barto et al., 1989; Bradtke & Duff, 1995)

$$\delta(t) = U_r(t) + e^{-\gamma} \tilde{V}^{\pi}(S_{t+1}) - \tilde{V}^{\pi}(S_t), \tag{5.4}$$

The value estimate of the state that led to the prediction error $\tilde{V}^{\pi}(S_t)$ is then updated according to equation (5.3), albeit using the new error term $\delta(t)$ from equation (5.4). This stochastic version of discounted 'value iteration' is guaranteed to converge on the true values, providing that all states are sampled infinitely often and that the learning rate decreases appropriately with training (Sutton & Barto, 1998).

As we saw in Chapter 1 (section 1.2.2), it is straightforward to derive similar 'stochastic value iteration' algorithms for 'on-policy' or 'off-policy' online learning of $Q(S,a)$-values (Watkins, 1989; Bradtke & Duff, 1995). Furthermore, by basing the policy π on these values, the policy can change as information about the state-action values is acquired, and, similar to 'policy iteration' methods, the process will converge not only on true policy-dependent values, but also on the optimal policy in terms of maximizing discounted future rewards. Asymptotic convergence of these algorithms is also guaranteed providing all states are visited infinitely often (Watkins, 1989; Watkins & Dayan, 1992).

An alternative online 'model-free' way to find an optimal policy based on a stochastic approximation to 'policy iteration' is by using an Actor/Critic architecture (see Chapter 1). In this, a policy $\pi(a|S)$ (ie, the probability of performing each action at every state) is explicitly represented by the Actor, and temporal difference errors computed by the Critic (equation 5.4) are used to update both state-value estimates in the Critic, and the policy in the Actor (Barto, 1995; Dayan & Abbott, 2001). Although widely used and of great interest due to its suggested neural counterparts, this method is less sound than 'value-iteration'-based methods, and can be proven to converge only if the policy is updated at least an order of magnitude slower than the values (Konda & Tsitsiklis, 2003).

Problem 1: Average reward reinforcement learning

How do these methods extend to the case of average reward reinforcement learning? Again, a good place to start is the consistency equation for state values (see Chapter 2, equation 2.7)

$$V^{\pi}(S) = \int \pi(a,\tau|S)\left[P_r^{a,\tau}(S)\cdot U_r - C_u(a,a_{prev}) - \frac{C_v(a,a_{prev})}{\tau} - \tau\cdot\bar{R}^{\pi} + \sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'}^{a,\tau}\cdot V^{\pi}(S')\right] d\pi \quad (5.5)$$

$$= \langle U_r\rangle_{P_r,\pi} - \langle C_u(a,a_{prev})\rangle_{\pi} - \left\langle \frac{C_v(a,a_{prev})}{\tau}\right\rangle_{\pi} - \langle\tau\cdot\bar{R}^{\pi}\rangle_{\pi} + \langle V^{\pi}(S')\rangle_{\mathcal{T},\pi} \quad (5.6)$$

which is different from equation (5.1) only in that we have subtracted action costs and the opportunity cost of time, rather than discounted the value of the future state. Here, similar to the discounted case, if actions and latencies are chosen according to $\pi(a,\tau|S)$, the reward probabilities and state transitions can be sampled from experience with the environment, and the inconsistency between subsequent state values (where consecutive states in continuous time and the actions chosen in each of these are now indexed by $n = 1,2,3...$)

$$\delta(n) = U_r(n) - C_u(a_n,a_{n-1}) - \frac{C_v(a_n,a_{n-1})}{\tau_n} - \tau_n\cdot\bar{R}^{\pi} + \tilde{V}^{\pi}(S_{n+1}) - \tilde{V}^{\pi}(S_n) \quad (5.7)$$

can be used to update the value estimates $\tilde{V}^{\pi}(S_n)$ according to equation (5.3). ARL temporal-difference Q-learning and Actor/Critic learning rules can be similarly derived (eg. Schwartz, 1993a, 1993b; Singh, 1994).

However, the extension of TD methods to ARL is not completely straightforward (Mahadevan, 1996) because, in addition to estimating the values of states (or state-action-latency triplets), the net rate of reward $\bar{R}^{\pi}$ must also be estimated. This is not trivial as consistent policy-dependent state or state-action values exist only in conjunction with the *true* policy-dependent net reward rate. On the one hand, if the policy is not constant but rather is improved as the value estimates are improved, the net reward rate estimate should change as the policy changes, tracking the true net reward rate. However, on the other hand, to allow stability of the estimated values, it must change slowly enough so that it is relatively constant with respect to the values. The method of estimating the net reward rate thus critically influences the convergence and stability of ARL online learning methods. For instance, different from discounted RL in which Q-values can be learned 'off-policy' (that is, the values Q^* corresponding to the optimal policy can be learned although actions are selected according to a suboptimal policy π), 'off-policy' learning will not converge in ARL as only $\bar{R}^{\pi}$ (and not $\bar{R}^*$) can be estimated from experience.

A second concern stems from the importance of stochastic sampling of actions and states for the stability of ARL. Recall that the soundness of ARL depends on the unichain property of the MDP (Puterman, 1994). That is, for there to be one net reward rate $\bar{R}$ common to all states, each state must be reachable from all others. Stochastic action selection can ensure this, however, the specific method of action selection can affect the so-called 'mixing-time' of the Markov chain of states, defined as the time constant of its exponential convergence to a stationary distribution (Baxter & Bartlett, 1999), which affects how quickly the measured net reward rate reaches a value that is close enough to the asymptotic net reward rate of the policy (Kearns & Singh, 1998)). This, in turn, determines the convergence and stability of an online learning algorithm. We will return to this issue in more detail in the Discussion, but we note it here because the issue of the specific mode of exploratory action selection will pervade this chapter.

Previous online learning algorithms for ARL suggest two alternatives for estimating the average reward[2] $\bar{R}$. In "$\mathcal{R}$-learning", the first online ARL algorithm, Schwartz (1993a, 1993b) suggested that because $V(S_t) = \langle U_r(S_t)\rangle + \langle V(S_{t+1})\rangle - \bar{R}$, the average reward estimate $\bar{R}$ can be stochastically updated toward the target $[U_r(S_t) + V(S_{t_+1}) - V(S_t)]$. In practice, this indirect method that bootstrappes state value estimates is rarely used (see Millan et al., 2002; Tadepalli & Ok, 1996, for two examples).

More popular is to estimate $\bar{R}$ directly from the obtained rewards, for instance, as a running average of the outcomes of the past T timesteps. Singh (1994), the first to suggest this method, argued that only those timesteps in which the greedy (optimal) action was chosen should be taken into account, in order to avoid penalizing the estimated average reward with the outcomes of trials in which a suboptimal action was deliberately chosen. Note, however, that it is only appropriate to estimate the reward rate solely from trials in which a greedy action was chosen, if action selection is according to the commonly used ε-greedy method (in which in a small ($\varepsilon << 1$) percent of the trials a random action is chosen instead of the optimal action; eg. Schwartz, 1993b; Singh, 1994), but not for action selection methods that do not choose the greedy action equally often at all states. Figure 5.1 illustrates this for soft-max action selection. In blue are the optimal

[2] In the discrete time MDP ARL setting, $\bar{R}$ has units of reward (rather than reward per timestep) and it is more appropriate to designate it the average reward (over time), rather than the reward rate.

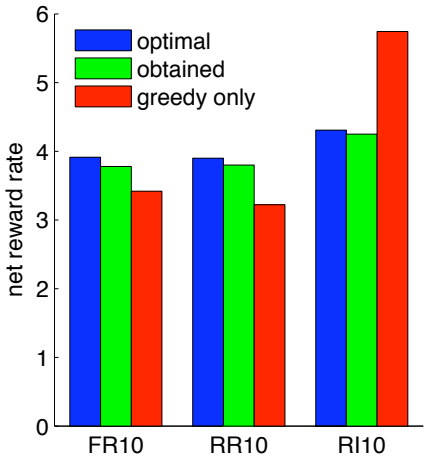

Figure 5.1: Estimating the net reward rate from optimal actions alone can result in biased estimates. Shown are the optimal $\bar{R}$ (blue), the net reward rate obtained in 5000 trials of soft-max action selection (green), and the estimated reward rate when only optimal actions are taken into account (red), for three different free-operant schedules. The obtained net reward rate is slightly lower than the optimally achievable one because behavior was generated using a softmax policy rather than the optimal deterministic (greedy) policy. Estimating the reward rate based only on optimal actions underestimates the reward rate in ratio schedules, and overestimates it in interval schedules. Simulation parameters were as in Chapter 2.

net reward rates (computed by solving the Dynamic Programming equations from Chapter 2) for fixed ratio, random ratio, and random interval free-operant schedules of reinforcement. In green is the net reward rate obtained with soft-max action selection based on the optimal steady-state Q-values, and in red is the reward rate estimate which takes into account only steps in which the optimal action was chosen. The latter estimates are significantly biased because, even in steady-state behavior, the amount of stochasticity differs between states S, and so the estimated net rate of rewards is not based on the true stationary distribution of the Markov chain. For instance, if the distribution of latencies to nosepoke is narrower than the distribution of latencies to leverpress, such a reward rate estimate will be too high because it will take into account a disproportionately large percentage of nosepoking actions (which are the only actions obtaining rewards).

Problem 2: Continuous actions

A second complicating factor is the space of possible actions in our model. Specifically, as a result of the choice of latency, the action space in our model is continuous. In general, dealing with continuous state and/or action spaces requires the use of function approximation. Unfortunately, RL implementations using function approximation are notorious for being unsound, and for providing ad-hoc problem-specific solutions at best. Moreover, the RL literature concentrates mainly on function approximation for large or continuous state spaces, rather than for continuous action spaces. Note, however, that selecting an action to perform typically requires some form of search within the available actions. A continuous action space thus poses a more difficult problem than a continuous state space (Smith, 2002b). Even the simplest schemes of action selection, such as choosing the action with the highest value, or selecting actions according to soft-max probabilities based on their values, become impossible with most general forms of function approximation (Santamaria et al., 1998; Smith, 2002b). Action selection from a continuous action space thus requires either a very special structure for the action-value function, or explicit storage and updating of a policy, such as in Actor/Critic methods.

A common solution to this problem is to preselect a discrete subset of actions and to find an optimal policy using these alone, with traditional discrete-action RL methods (eg. Barto et al., 1983; Santamaria et al.,

1998; for an extensive discussion see Smith, 2002a, 2002b)[3]. Indeed, to derive the steady state solution for the free-operant case using Dynamic Programming in Chapter 2, we discretized time and derived values for a discrete subset of latencies. However, by definition, discrete-action learning methods do not treat actions as residing on some continuum. In terms of online learning from stochastic experience, estimating values for a large number of discretized latencies while ignoring the distance relationships between them, seems an unsatisfactory (and inefficient) solution, especially given that the optimal state-action-latency values in our model are smooth as a function of latency (eg, see Figure 2.7). For instance, without assuming smoothness of values as a function of latency, learning that a leverpress with a latency of 3.4 seconds is highly valuable would not generalize at all to leverpressing with, say, a latency of 3.5 seconds.

A second approach is to approximate the continuous space with a special class of functions in which searching for the best action is feasible (eg. Baird & Klopf, 1993). However, these methods are only appropriate for generating either deterministic (greedy) policies, or ε-greedy policies. ε-greedy action selection is inappropriate as a model of animal free operant behavior, as it assumes a high probability of selection for only one specific (optimal) action-latency choice in each state, and a similar (low) probability for all other sub-optimal action choices (including latencies very similar to the optimal one, as well as very costly short-latency actions or extremely long latencies). This is at odds with our experimental observations regarding free-operant behavior (see, for instance, Figure 2.5), in which similar latencies are chosen similarly often, and close-to-optimal latencies are chosen more often than those that are very suboptimal (see also Desideratum II below).

The third approach, which our algorithm will follow, is to represent the policy itself (the probability of choosing each action and latency at a given state) as a parameterized function, and to learn the parameters of this function using online ARL methods. This approach can be seen as a natural extension to the Actor/Critic framework that explicitly represents a lookup-table policy (ie, the probability of performing each of a discrete set of actions at a given state, eg Barto et al., 1983). The first and influential learning algorithm for a parameterized continuous policy is due to Williams (1992). His "REINFORCE" algorithm, developed for immediate reward tasks, is essentially a prescription for building 'model-free', incremental, gradient-following learning rules for a parameterized stochastic policy $\mathcal{F}(\vec{\theta})$. The REINFORCE learning rule is $\Delta\theta_i = \eta_i(r_t - b_i)e_i$, where $\vec{\theta}$ are the policy parameters, η_i is the learning rate for parameter θ_i, r_t is the immediate experienced reward, b_i is some action-independent 'reinforcement baseline' (a good choice for which is the average reward), and e_i is the 'characteristic eligibility' of parameter θ_i equal to $\frac{\partial \log \mathcal{F}(\vec{\theta})}{\partial \theta_i}$. This method can be extended (as in Dayan and Abbott's (2001) 'Direct Actor' method) to general MDPs that involve delayed rewards, by substituting the immediate reward r_t with the long-term reward $r_t + V(S_{t+1})$. In Section 5.2, we will use this method to develop the update rules for our algorithm.

5.1.2 Constraints/Desiderata

There is much current interest in the extent to which animals *learn* optimally from their interaction with the environment, in addition to just showing optimal behavior *after* learning. The task of learning optimally

[3] In fact, due to the success of these applications, Smith (2002a) has difficulty finding problems in which use of a continuous action space is practically justifiable.

places much greater burdens on the implementational substrate, and so psychological and neurobiological considerations turn out to play a critical role in modeling. Ideally, we would be able to describe the whole class of learning algorithms consistent with the various sources of data; however, this is beyond the bounds of current theory. Instead, we list here the general constraints implied by the data. In the following section, we will provide a learning algorithm that is consistent with these.

I. The most basic constraint is that the algorithm should learn a policy over a discrete set of actions and a continuous set of latencies that is close to optimal in terms of the net reward rate obtained. We have already interpreted substantial experimental data under this rubric in Chapter 2.

II. Although optimal algorithms in non-strategic contexts are typically deterministic, animal behavior is not, and thus it is necessary to consider stochastic policies. Specifically for the domain of free operant responding, we are interested in those policies that distribute choice such that similar latencies are chosen similarly often[4] (although our algorithm will not obey the additional constraint of inherent timing noise, which may play a role in underpinning this behavioral characteristic). Furthermore, learning of strictly optimal deterministic policies is somewhat irrelevant in the case of animal learning This is because in stochastic environments strictly optimal solutions are possible only if the learning rate is gradually reduced to zero, which is an unrealistic assumption for learning in an ever-changing world.

III. Extensive data show that animals acquire good behavior in free operant tasks rather rapidly. This places severe constraints on any learning algorithm that is required to reach steady state performance within an amount of training that is on the order of that given to animals. Specifically for our case, due to the smoothness of action values as a function of latency in the optimal solution, we will consider generalization among different latencies key to efficient learning. A subtle liberation implied by this focus on moderate amounts of training, is that properties of the algorithm that are important only after very large amounts of training (such as provable convergence given a stationary environment), whilst theoretically desirable, are not a prime concern. Indeed, the widely influential Actor/Critic model (see Chapter 1), which turns out to provide the best infrastructure for our algorithm (see the continuous action space problem above) is the least sound of all RL methods, in terms of convergence.

IV. An obvious constraint is that the learning rule should be incremental and online with respect to samples from the environment.[5] General neurobiological considerations also suggest that it should only use information that is locally available, that is, its working memory demands should be minimal.

V. Finally, in the previous two chapters we reviewed extensive evidence that latencies depend immediately on the net rate of reward $\bar{R}$, acting through the medium of tonic levels of dopamine. The requirement for immediacy is a severe constraint, since the central property of caching algorithms of the kind that we are considering here is that they do not maintain online dependencies on any of the separate constituents of action values or policies. As an exception to this general rule, we will impose this dependency explicitly, essentially in the representation our algorithm employs.

[4]This could also ensure that the unichain property holds, given that the commonly used uniform probability distribution over actions is not defined over all positive latencies.

[5]We will ignore the role of post-learning consolidation and replay (Foster & Wilson, 2006).

5.2 Methods

In this section, I will develop a model-free online learning algorithm conforming to the above constraints, for the SMDP setting described in Chapter 2 (section 2.2.2). Our main goal is to learn a suitable policy for free operant behavior, ie, which actions to perform, and with what latencies, in order to reap rewards optimally. Dayan and Abbott (2001) describe two general approaches to learning a policy: in the 'Direct Actor' method, a parameterized functional form of the policy is assumed, and the parameters are learned, eg, by update rules that ascend the gradient of the average reward, as in "REINFORCE"; in the 'Indirect Actor' method, Q-values for the different actions are learned, which indirectly specify a policy. Because our experimental data readily suggests a functional form for the policy, I will first develop the algorithm following the 'Direct Actor' method. As I will discuss in Section 5.2.2, this turns out to be similar to using an 'Indirect Actor' with a specific functional form for the Q-values that is suggested by the Bellman equation for random ratio schedules.

5.2.1 An Actor/Critic policy-iteration algorithm for free operant behavior

Building on the apparent functional form of the experimentally measured inter-response intervals in a variable interval task (Figure 2.5), we will limit the form of the policy $\pi = P(a_i, \tau|S)$ where $i \in \{\text{LP}, \text{NP}, \text{Other}\}$, $\tau > 0$ and $S \in \mathcal{S}$, to a mixture of Gamma distributions (one distribution for each action a_i), such that

$$P(a_i, \tau|S) = \frac{m_i}{\sum_j m_j} \cdot \frac{\tau^{\alpha_i - 1} e^{-\tau/\theta_i}}{\theta_i^{\alpha_i} \Gamma(\alpha_i)}. \tag{5.8}$$

The policy is thus parameterized by $\Theta = \{\vec{\mathbf{m}}, \vec{\alpha}, \vec{\theta}\}$, where $m_i \geq 0$ are the mixing proportions (specifying the probability of choosing each of the actions a_i), and $\alpha_i > 0$ and $\theta_i > 0$ are the shape and scale parameters of the Gamma distribution specifying the choice of latencies for action a_i.

Let us first treat the problem of improving the policy based on the currently estimated state values (as in 'policy iteration'). Following Williams (1992), for some long-run measure of expected future reward r_{future} when starting from state S (eg, the immediate reward plus the value of the subsequent state), we can write the policy-dependent expected future reward in each state as

$$\langle r_{future}|S\rangle_\pi = \sum_{a_i} \int d\tau P(a_i, \tau|S) \cdot \langle r_{future}|S, a_i, \tau\rangle. \tag{5.9}$$

To update the policy parameters following the gradient of the expected long-run reward, we must take the derivative of the expected reward with respect to each of the parameters of the policy:

$$\frac{\partial \langle r_{future}|S\rangle_\pi}{\partial \Theta_i} = \sum_{a_i} \int d\tau \frac{\partial p(a_i, \tau|S)}{\partial \Theta_i} \cdot \langle r_{future}|S, a_i, \tau\rangle \tag{5.10}$$

(note that given a state, action and latency, the expected reward $\langle r_{future}|S, a_i, \tau\rangle$ is determined by the envi-

ronment and independent of changes in the policy parameters). This gradient can be written in the form

$$\frac{\partial \langle r_{future}|S\rangle_\pi}{\partial \Theta_i} = \sum_{a_i} \int d\tau p(a_i,\tau|S) \cdot \langle r_{future}|S,a_i,\tau\rangle \left[\frac{\partial \log p(a_i,\tau|S)}{\partial \Theta_i}\right] \tag{5.11}$$

from which the stochastic updates for Θ_i are now apparent: if, whenever action (a_i,τ) is performed in state S, we update Θ_i according to $\langle r_{future}|S,a_i,\tau\rangle[\,\cdot\,]$, then the updates will follow the gradient of the expected reward, properly averaged. $\left[\frac{\partial log p(a_i,\tau|S)}{\partial \Theta_i}\right]$ is the REINFORCE 'eligibility' of parameter Θ_i. As an unbiased proxy for $\langle r_{future}|S,a_i,\tau\rangle$, we can use the immediate rewards minus costs (shorthanded r_t), minus the opportunity cost $\tau_t \cdot \bar{R}^\pi$, plus the expected value of the subsequent state $V(S_{t+1})$, stochastically sampling the immediate reward and the transition to the successor state from the environment. From these we can subtract an arbitrary baseline term for all actions in this state (Dayan & Abbott, 2001), a convenient choice being $V(S_t)$. This results in the conventional TD error $\delta(t) = r_t - \tau_t \cdot \bar{R}^\pi + V(S_{t+1}) - V(S_t)$ as the proxy for $\langle r_{future}|S,a_i,\tau\rangle$. Following Kimura and Kobayashi (1998a, 1998b), we thus construct the following Actor/Critic algorithm:

ALGORITHM 5.1: A FREE OPERANT ONLINE ACTOR/CRITIC LEARNING ALGORITHM

1. **Initialize** policy parameters $\Theta_0 = \{\vec{\mathbf{m}}_\mathbf{0}, \vec{\alpha}_\mathbf{0}, \vec{\theta}_\mathbf{0}\}$, state values $V_0(S) = 0$ for all $S \in \mathcal{S}$, and net reward rate $\bar{R} = 0$; Get initial state S_0 from the environment and select initial action $a_0 = a_i$ and latency τ_0
2. **Repeat**:
   (a) Perform the chosen action at the chosen latency, observing the immediate net reward obtained $r_t = U_r - C_u(a_t, a_{t-1}) - \frac{C_v(a_t,a_{t-1})}{\tau_t}$, and the new state S_{t+1}
   (b) **CRITIC:** (relative value iteration)
      i. Compute the TD error: $\delta(t) = r_t - \bar{R}_t \int_0^{\tau_t} e^{-\eta_{\bar{R}} \cdot t} dt + V(S_{t+1}) - V(S_t)$, where $\eta_{\bar{R}}$ is the learning rate (decay rate) of the net reward rate estimate (see below)
      ii. Update the state value: $V(S_t) \leftarrow V(S_t) + \eta_v \delta(t)$, where η_v is the state value learning rate
      iii. Update the net reward rate estimate: $\bar{R}_{t+1} = e^{-\eta_{\bar{R}}\tau_t} \cdot \bar{R}_t + \eta_{\bar{R}} \cdot r_t$, where $\eta_{\bar{R}} << 1$ is the exponential averaging rate[a] for $\bar{R}$ (based on Daw & Touretzky, 2002)
   (c) **ACTOR:** (policy iteration by gradient ascent)
      i. Update the policy parameters:

$$\begin{aligned} m_j &\longleftarrow m_j + \eta_m \delta(t) \frac{1}{m_j}\left[\delta_{a_i a_j} - \frac{m_j}{\sum_k m_k}\right] \\ \alpha_i &\longleftarrow \alpha_i + \eta_\alpha \delta(t)\left[log\frac{\tau}{\theta_i} - \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)}\right] \\ \theta_i &\longleftarrow \theta_i + \eta_\theta \delta(t)\left[\frac{\tau - \alpha_i\theta_i}{\theta_i^2}\right] \end{aligned}$$

      where η_x is the learning rate for parameter x, $\delta(t)$ is the TD prediction error provided by the Critic, and $\delta_{a_i a_j}$ is 1 if the chosen action is a_i and zero otherwise (thus all the components of $\vec{m}$ are updated after every action, but α_i, β_i are updated only when a_i is chosen).
      ii. Select new action $a_t = a_i$ and latency $\tau_t = \tau$ with probability $P(a_i,\tau|S) = \frac{m_i}{\sum_j m_j} \cdot \frac{\tau^{\alpha_i - 1} e^{-\tau/\theta_i}}{\theta_i^{\alpha_i}\Gamma(\alpha_i)}$

[a] The discrete time update rule $\bar{R}_{new} \longleftarrow \bar{R}_{old} + \eta_{\bar{R}}(r_t - \bar{R}_{old})$ becomes, in continuous time, $\bar{R}_{t+1} = (1-\eta_{\bar{R}})^{\tau_t} \cdot \bar{R}_t + \eta_{\bar{R}} \cdot r_t = e^{\log(1-\eta_{\bar{R}})\cdot\tau_t} \cdot \bar{R}_t + \eta_{\bar{R}} \cdot r_t$, and for sufficiently small $\eta_{\bar{R}}$ we can substitute $-\eta_{\bar{R}} \approx \log(1-\eta_{\bar{R}})$.

Note that in this algorithm the action space is fully continuous in τ. The state space, however, is still discrete. In ratio schedules in which the number of states $|\mathcal{S}|$ is very small (on the order of twice the number of different actions (N_{actions}) for random ratio schedules, and on the order of $2 \cdot N_{\text{actions}} \cdot n$, where n is the schedule requirement, in fixed ratio schedules), this method will be most powerful as the model will be completely continuous in time, while only tuning $3 \cdot N_{\text{actions}} \cdot |\mathcal{S}|$ parameters in the Actor. In interval schedules a fully Markov state space must include the time from the last leverpress, making the state space continuous as well. In this case, function approximation of the state values (as in Sutton et al., 2000) would be in line[6].

5.2.2 Correspondence to 'Indirect Actor' methods

In 'Indirect Actor' methods a policy is not directly represented by the Actor, but rather it is derived indirectly from state-action Q-values whenever an action must be chosen (Dayan & Abbott, 2001). We must thus parametrize a functional form for the $Q(S,a_i,\tau)$ values for an action a_i in a certain state S as a function of τ, and devise an appropriate learning algorithm for updating the parameters of this function based on rewards and state transitions sampled from the environment. In accord with Desideratum II, we will choose a function that lends itself to action selection according to a soft-max distribution of the Q-values. Furthermore, it is reasonable to request (Bertsekas & Tsitsiklis, 1996) that this function be able to sufficiently approximate the optimal Q-values that were found using Dynamic Programming in Chapter 2. I will now use these two constraints to derive a functional form for $Q(\cdot,\cdot,\tau)$. I will then show its relationship to the Bellman equation, and the relationship of the resulting algorithm to the 'Direct Actor' method above.

Let us consider the choice of a latency τ, a selected action a. In figure 2.8, we showed that applying the soft-max operator to the optimal Q-values for a given action as a function of latency results in a Gamma-shaped function. Working back from the Gamma distribution through the soft-max operator, we showed that this means that the Q-value function is the difference between a logarithm and a linear function (see Section 2.3.2). To recap, by and ignoring the normalization of the Q-values in equation (2.16)

$$p(\tau|S,a) = \frac{e^{\beta Q(\cdot,\cdot,\tau)}}{\int e^{\beta Q(\cdot,\cdot,\tau)} d\tau} = \frac{\tau^{\alpha-1} e^{-\tau/\theta}}{\theta^{\alpha}\Gamma(\alpha)}, \tag{5.12}$$

(in which β is the soft-max inverse temperature and (α,θ) are the shape and scale parameters of the Gamma distribution), we can write

$$Q(\cdot,\cdot,\tau) = \frac{1}{\beta}\left[(\alpha-1)\log\tau - \frac{\tau}{\theta} - \log\Gamma(\alpha) - \alpha\log\theta\right]. \tag{5.13}$$

Thus, by setting $(A = \frac{\alpha-1}{\beta})$ and $(B = \frac{1}{\beta\theta})$, we can parametrize the Q-values as the difference between a logarithm and a linear function of τ

$$Q(\cdot,\cdot,\tau) = A\log\tau - B\tau - C \tag{5.14}$$

[6] We do not use function approximation for the state space because even in random interval schedules the effective number of relevant states is small as every leverpress resets the Poisson baiting clock to zero.

where $\Theta(S,a) = \{A,B\}$ are tunable parameters specific to each (state,action) pair, and C is a normalizing constant[7]. By construction, this functional form is amenable to the soft-max operation (which will result in Gamma(α,θ) distributed latencies). Note, however, that by ignoring the normalization of the Q-values, we have ignored the overall probability of choosing each action a_i, itself a parameter that must be learned. In Q-learning this probability is based on the relative future value of the different actions. But this is where the correspondence to the 'Direct Actor' method breaks, as there the policy maintains only an ordinal relationship between actions, and ultimately selects only the action that has the highest value (see section 5.4.3).

How does this parameterization relate to the Bellman equation defining the optimal Q-values? Recall that the Bellman equation (equation 2.9) was

$$Q^*(S,a,\tau) = P_r^{a,\tau}(S)\cdot U_r - C_u(a,a_{prev}) - \frac{C_v(a,a_{prev})}{\tau} - \tau\cdot\bar{R}^* + \sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'}^{a,\tau} V^*(S'). \tag{5.15}$$

In all the free-operant schedules we have discussed, the probability $P_r^{a,\tau}(S)$ is, in fact, independent of τ. This is because only one action is directly rewarded (with probability $P_r = 1$ regardless of its latency), and that is nosepoking when in the state in which food is available in the magazine. By definition, the unit cost term is also independent of response latency. This leaves three terms that can contribute to the functional form of $Q(\cdot,\cdot,\tau)$. Based on the vigor cost and the opportunity cost terms we can expect there to be one component which goes to $-\infty$ when τ approaches zero (such as $\log\tau$ or $1/\tau$) and one that is linear in τ, both in line with the suggested parameterization.

However, the state transition probabilities $\mathcal{T}_{S\to S'}^{a,\tau} = p(S'|S,a,\tau)$ can take the shape of any arbitrary probability distribution, and the state values $V^*(S')$ can also take any arbitrary form, depending on the structure of the task. The functional form suggested in 5.14 can thus represent optimal Q-values only for a restricted subset of tasks in which the relationship between $\sum_{S'\in\mathcal{S}} \mathcal{T}_{S\to S'}^{a,\tau} V^*(S')$ and τ is itself a sum of a logarithm and a linear function of τ. In particular, ratio schedules in which the expected value of the subsequent state is independent of response latency satisfy this constraint, and random interval schedules approximately conform to this constraint because the relationship between latency and subsequent state value is approximately linear.

Given this parameterization of the Q-values, and following Sutton and Barto (1990), we can now derive local learning rules of the form $\Theta_i \longleftarrow \Theta_i + \eta\delta(t)\frac{\partial Q(S,a,\tau)}{\partial\Theta_i}$ where $\delta(t)$ is the temporal difference prediction error as in equation (5.7). The partial derivatives of $Q(S,a,\tau)$ with respect to A,B are:

$$\frac{\partial Q(S,a,\tau)}{\partial A} = \log\tau - \frac{\Gamma'(\beta A+1)}{\Gamma(\beta A+1)} + \log(\beta B) \tag{5.16}$$

$$\frac{\partial Q(S,a,\tau)}{\partial B} = -\tau + \frac{\beta A+1}{\beta B} \tag{5.17}$$

[7] C is inconsequential because the Q-values are differential and so defined up to an additive constant, and the soft-max operator is indifferent to additive constants. We include it here only to demonstrate the relationship to the 'Direct Actor' method.

and so, reparametrizing with (α, θ) we get

$$\frac{\partial Q(S,a,\tau)}{\partial \alpha} = \frac{1}{\beta}\frac{\partial Q(S,a,\tau)}{\partial A} = \frac{1}{\beta}\left[\log\frac{\tau}{\theta} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\right] \tag{5.18}$$

$$\frac{\partial Q(S,a,\tau)}{\partial \theta} = -\frac{1}{\beta\theta^2}\frac{\partial Q(S,a,\tau)}{\partial B} = \frac{1}{\beta}\left[\frac{\tau}{\theta^2} + \frac{\alpha}{\theta}\right] \tag{5.19}$$

giving the same update rules for (α, θ) in the 'Indirect Actor' method as were derived for the 'Direct Actor' method. This is not surprising, because the 'Direct Actor' update rules were derived by taking the derivative of the log of the (Gamma-shaped) parameterized policy with respect to its parameters (Williams, 1992), while in the 'Indirect Actor' the parameterized form of the Q-values was derived by taking the log of a Gamma distribution. Our method of generalizing between different latencies using a Gamma-shaped function approximation of the policy within a 'Direct Actor' framework, is thus similar to an 'Indirect Actor' in which the Q-values are parameterized as the difference between a log and a linear function of the latency, in accord with intuitions derived from the free-operant Bellman equation.

5.2.3 Incorporating the net rate of reward into the algorithm

In the algorithm we have suggested, the net rate of reward $\bar{R}$ is used by the Critic to compute the prediction errors with which state values and policy parameters are learned, but online action selection by the Actor is independent of the net reward rate. However, in Chapter 4 we suggested that the expected net rate of reward is represented by tonic level of dopamine, and directly affects the choice of response latencies. How can such an influence be incorporated into our Actor/Critic algorithm?

The Actor's policy is parameterized such that the choice of which action to perform can be decoupled from the choice of response latency: the mixing proportions parameters m_i influence the former decision, while the latter is affected only by the Gamma distribution parameters (α_i, θ_i) of the chosen action. This means that, to influence response latency, $\bar{R}$ should influence either α_i or θ_i. In section 5.2.2 we discussed the relationship between (α_i, θ_i) and the Bellman optimality equation: according to the Bellman equation, changes in $\bar{R}$ should have an immediate effect on the term that is linear in τ, which is parameterized by θ_i. The net reward rate can thus influence the choice of response latencies through a direct link between θ_i and $\bar{R}$. What is the precise nature of this link? In the absence of normative guidelines, we will base this on the known relationship between the net reward rate and the optimal response latency in ratio schedules, as this analytic relationship also provides a reasonable approximation for the effects of changes in the net reward rate on behavior in interval schedules.

In Chapter 2 (section 2.1.2) we showed that the optimal response latency τ^* in ratio schedules is:

$$\tau^*_{ratio} = \sqrt{\frac{C_v(a, a_{prev})}{\bar{R}^*}}. \tag{5.20}$$

This means that a change in the net rate of reward, say, $\bar{R} \longrightarrow a \cdot \bar{R}$, will cause a $\frac{1}{\sqrt{a}}$-fold change in the optimal latency. To relate this to the Actor's policy, note that the maximum of a Gamma probability distribution is attained at $\tau^* = (\alpha - 1)\theta$. Thus, we can define $\theta = \hat{\theta}/\sqrt{\bar{R}}$, and learn the parameter $\hat{\theta}$ while still selecting actions based on $\Theta = (\vec{m}, \vec{\alpha}, \vec{\theta})$. Using the chain rule and $\frac{\partial\theta}{\partial\hat{\theta}} = \frac{1}{\sqrt{\bar{R}}}$ we can derive the new learning rule for $\hat{\theta}$:

$$\hat{\theta}_i \longleftarrow \hat{\theta}_i + \eta_\theta \delta(t) \left[\frac{\tau\sqrt{\bar{R}} - \hat{\theta}_i \alpha_i}{\hat{\theta}_i^2} \right], \tag{5.21}$$

and because the update for α depends on the value of θ, learning of α is now also influenced by $\bar{R}$

$$\alpha_i \longleftarrow \alpha_i + \eta_\alpha \delta(t) \left[log \frac{\tau\sqrt{\bar{R}}}{\hat{\theta}_i} - \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)} \right]. \tag{5.22}$$

5.2.4 Constraining the policy parameters

The learning algorithm we proposed uses stochastic samples of the environment dynamics to learn three parameters for each (state,action) pair: a mixing proportion m, and Gamma distribution parameters $\hat{\theta}$ and α. The values that these may take are restricted: the mixing proportions must be non-negative, the scale and shape parameters (θ, α) of the Gamma distribution must be positive, and α must be greater than 1 for the Gamma distribution to be first increasing and then decreasing. These constraints can easily be imposed by remapping the policy parameters

$$m = e^{\tilde{m}} \tag{5.23}$$
$$\alpha = e^{\tilde{\alpha}} + 1 \tag{5.24}$$
$$\hat{\theta} = e^{\tilde{\theta}} \tag{5.25}$$

and learning $\tilde{\Theta} = \{\vec{\tilde{m}}, \vec{\tilde{\alpha}}, \vec{\tilde{\theta}}\}$ rather than the original parameters. The Actor's update rules must thus be multiplied by the derivative $\frac{d\Theta}{d\tilde{\Theta}}$, giving, as Step 2(c)i in the algorithm, the policy update rules:

$$\tilde{m}_j \longleftarrow \tilde{m}_j + \eta_m \delta(t) \left[\delta_{a_i a_j} - \frac{e^{\tilde{m}_j}}{\sum_k e^{\tilde{m}_k}} \right] \tag{5.26}$$
$$\tilde{\alpha}_i \longleftarrow \tilde{\alpha}_i + \eta_\alpha \delta(t) \left[log \frac{\tau\sqrt{\bar{R}}}{e^{\tilde{\theta}_i}} - \frac{\Gamma'(e^{\tilde{\alpha}_i} + 1)}{\Gamma(e^{\tilde{\alpha}_i} + 1)} \right] e^{\tilde{\alpha}_i} \tag{5.27}$$
$$\tilde{\theta}_i \longleftarrow \tilde{\theta}_i + \eta_\theta \delta(t) \left[\frac{\tau\sqrt{\bar{R}}}{e^{\tilde{\theta}_i}} - e^{\tilde{\alpha}_i} - 1 \right] \tag{5.28}$$

and action selection proceeds according to $\Theta = \{e^{\tilde{m}}, e^{\tilde{\alpha}} + 1, \frac{e^{\tilde{\theta}}}{\sqrt{\bar{R}}}\}$.

In practice, to stabilize the algorithm we used such a remapping to constrain the Gamma distribution parameters even further, such that $\alpha > 1.05$ and $\theta > 0.05$. The value of $\bar{R}$ was also constrained to be at minimum 0.1, which proved useful in the beginning of training where negative reward rates can cause the animal to stop responding (by choosing very long response latencies). The learning rates were set to $\eta_{\bar{R}} = 0.0005$, $\eta_v = 0.05$, $\eta_m = \eta_\alpha = \eta_\theta = 0.01$. Importantly, the Critic learned faster than the Actor (Konda & Tsitsiklis, 2003), and learning of the net reward rate was slow compared to that of the other parameters;(Gosavi, 2004b). Finally, to prevent numerical overflow due to large values, the update rules for $\tilde{m}$ and $\tilde{\alpha}, \tilde{\beta}$ were multiplied at every step by $e^{-\tilde{m}}$ and $e^{-|\tilde{\alpha}|-|\tilde{\beta}|}$, respectively.

5.3 Results

In this section, I will first analyze the learning dynamics of our proposed algorithm, comparing the learning curves for interval and ratio schedules. To establish the validity of our algorithm as a free operant reinforcement learning algorithm, I will then establish that it indeed learns policies that approximate the known optimal policies (as derived in Chapter 2). Last, I will show that the effect of changes in the net rate of reward on response selection parallels that seen in the optimal solution (as in Chapter 4).

5.3.1 Learning dynamics

Figure 5.2 summarizes the results of twenty runs of our algorithm on a random ratio schedule of reinforcement, and twenty on a random interval schedule. Each run consisted of $3,000$ training trials (ie, training until $3,000$ rewards were obtained), comprising approximately $30,000$ actions per run[8]. Panel (a) shows the average learning curve for the random ratio runs, in terms of the net reward rate at each point in time, throughout learning (a learning curve from a single run is superimposed for illustration purposes). Panel (b) depicts the average learning curve for the random interval runs. The curves show that the algorithm fairly rapidly learns a policy that achieves a net reward rate close to the maximal possible. Although performance improves slightly thereafter, the learning curves nearly asymptote after approximately $7,000$ actions (~ 600 trials) for ratio schedules, and within fewer than $3,000$ actions (~ 200 trials) for interval schedules. For comparison, standard online Q-value learning with discretized action latencies necessitated approximately $10,000$ trials to asymptote for random ratio schedules (not shown).

Two differences between ratio and interval learning are apparent: learning of interval schedules asymptotes faster than that of ratio schedules, and the variance of the net reward rate at asymptote is substantially larger in interval schedules compared to ratio schedules. The single run learning curves illustrate the reason for this second observation: the net reward rate in interval schedules fluctuates widely in the steady-state. These fluctuations possibly result from the feedback structure of this schedule: initially, as the model learns the task, faster response rates results in rewards being harvested faster, and thus the net reward rate increases.

[8] Results from longer runs of $10,000$ trials did not give qualitatively different results.

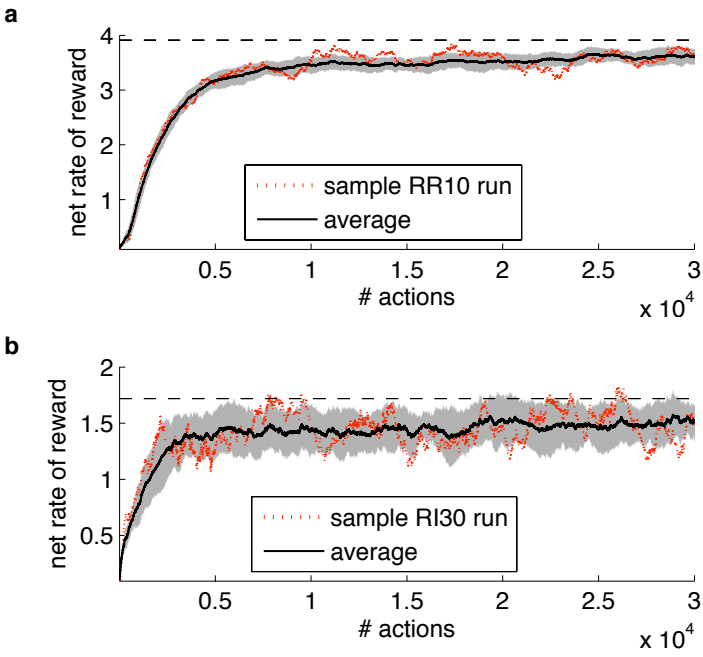

Figure 5.2: Learning curves for random ratio and random interval schedules. **a**. Net rate of reward as a function of time (actions), averaged over the twenty runs of the RR10 schedule (black solid; gray shading: standard deviation), with the curve from one sample run overlaid in dotted red. The net rate of reward is an indication of the overall quality of the policy, and a measure of learning. Dashed black horizontal line: optimal net reward rate achievable with the optimal deterministic policy. Note that this optimum is not a realistic target for an algorithm learning stochastic policies. **b**. Net rate of reward as a function of actions averaged over the twenty RI30 runs (color conventions as in a). In all runs task parameters were as in Chapter 2, and the learning algorithm parameters were as detailed in Section 5.2.4. Initial values were: $V(S) = 0$, $M(S,a) = 1$, $\alpha(S,a) = e^1$, $\theta(S,a) = e^{-1.5}$.

However, once reinforcements are obtained fairly soon after they are baited, faster responding has little effect on the rate of reinforcement, and, due to response costs, only causes a decline in the net reward rate. This, in turn, slows down responding (recall that the optimal latency to respond is inversely proportional to the net rate of reward[9]), which causes an increase in the net reward rate, again increasing response rates and so forth[10]. In ratio schedules, such a complicated feedback cycle does not exist because the relationship between response rate and reinforcement rate is linear.

The feedback relationship in both schedules can also explain the different rates of convergence of the learning process. The optimal policy for ratio schedules is well defined within a narrow range of parameter values. Thus specific parameters need to be learned in order to obtain a high net rate of reward, but once this range is reached all runs achieve similar net reward rates. Conversely, in interval schedules, many policies are close to the optimal: given the dominance of the schedule-limited rate of reinforcements in determining the net reward rate (each reinforcer has a utility $U_r = 60$ while a typical action cost is $(C_u + C_v/10) = 0.2$), different rates of leverpressing result in similar net reward rates. As long as the rat leverpresses at some minimal rate and nosepokes to retrieve its rewards, the net reward rate will be quite close to optimal. As a result, the learning algorithm achieves high net rates of reward rather quickly, although the policy may continue changing thereafter (eg, fluctuating between higher and lower response rates, as discussed above).

[9] These fluctuations are also seen in variants of our model in which action selection does not depend directly on the average rate of reward (ie, θ_i are learned rather than $\hat{\theta}_i$), suggesting that they result from an interaction between the learning process and the feedback function of the schedule, and not merely from the postulated dependence of the Actor on $\bar{R}$.

[10] Despite the rather oscillatory-sounding underlying dynamics, a Fourier transform of the individual learning curves (not shown) did not reveal a characteristic frequency of oscillations of the net reward rate.

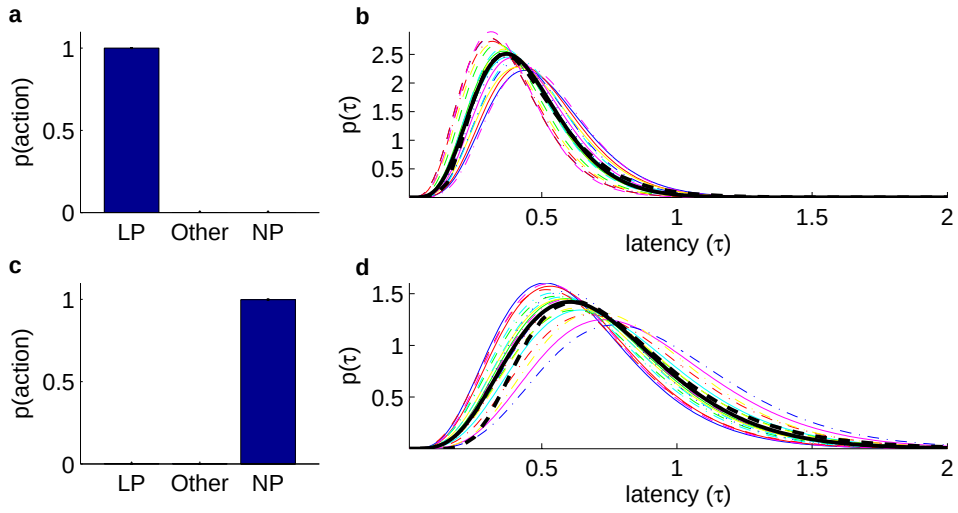

Figure 5.3: Policies learned in each of 20 runs (3000 trials each) of our online Actor/Critic learning algorithm on an RR10 schedule of reinforcement. **Top**: The Actor's policy for the state $(a_{prev} = \text{LP}, i_r = 0)$, that is, the state in which a previous leverpress was not rewarded. **Bottom**: The policy for the state $(a_{prev} = \text{LP}, i_r = 1)$, ie, after a rewarded leverpress. **a**. The mean probability (over all runs; black bars: standard deviation) of selecting each action shows that in the absence of rewards the leverpress action will be chosen. **b**. Color curves show the distribution of leverpress latencies in this state in each of the runs. The Gamma distributions are very similar and sharply peaked at a short (< 0.5s) latency. In solid black is the mean policy (a Gamma$(\langle\alpha\rangle, \langle\theta\rangle)$ distribution), for comparison with a soft-max policy derived from the optimal Q-values (as found by relative value iteration in Chapter 2), in dashed black. Note that the strictly optimal policy is deterministic; the soft-max temperature used to derive a comparable stochastic policy was arbitrarily chosen as that which equates the height of the peaks of the learned and derived policies. **c**. The mean probability of selecting each action shows that when a reward is available the rat will choose to nosepoke. **d**. In color are the nosepoke latency distributions in this state, in each of the runs. In solid black is the mean policy, for comparison with the optimal policy in dashed black.

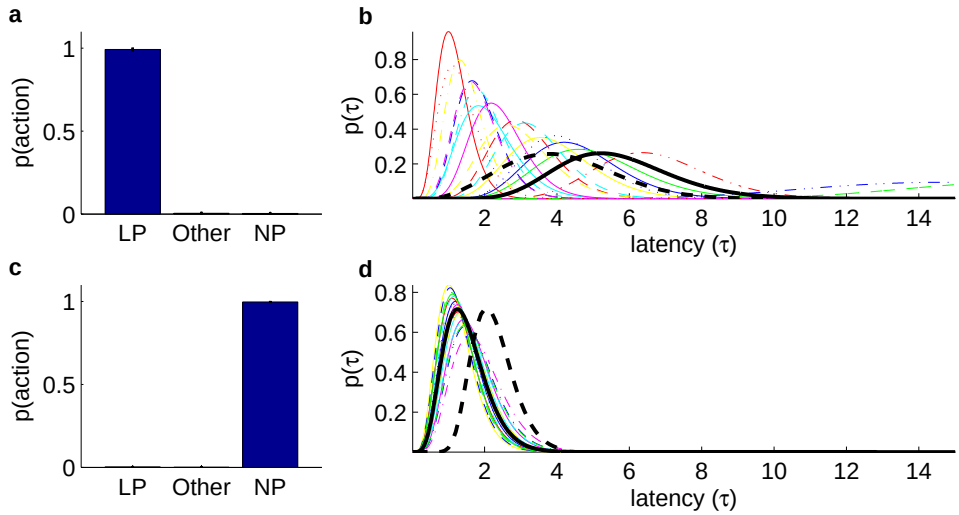

Figure 5.4: Policies learned in each of 20 runs (3000 trials each) of our online learning Actor/Critic algorithm on an RI30 schedule of reinforcement. **Top**: The Actor's policy for the state $(t_r = 1, a_{prev} = \text{LP}, i_r = 0)$, ie, just after an unrewarded leverpress. **Bottom**: The policy for the state $(t_r = 1, a_{prev} = \text{LP}, i_r = i)$, ie, right after a rewarded leverpress. See the caption of Figure 5.3 for details, though note the different time axes in the two figures.

5.3.2 Learned policies vs. optimal policies

Figure 5.3 details the learned policies at the end of each of the above random ratio schedule runs. The top panels depict the policy learned for the state in which no reward is available: the action chosen in this state is leverpressing (Figure 5.3a) and the color plots show that the leverpress latencies are short, their distribution sharply peaked (Figure 5.3b). The bottom panels depict the policy learned for the state in which a reward is available in the food magazine. Here, nosepoking is the preferred action, again with a relatively short latency. In solid black is the mean policy averaged over all twenty runs, which compares favorably to a policy derived from a soft-max over the optimal Q-values, nearly superimposed in dashed black.

Figure 5.4 shows the policies learned in the random interval schedule runs. In contrast to the relatively homogeneous results for the ratio schedule, in interval schedules, as might be expected given the fluctuating nature of the steady state of the learning algorithm (see Figure 5.2b), different runs ended with quite different policies: the distributions of leverpress latencies in the 'no reward' state are quite broad (note the different scaling of the x-axis compared to Figure 5.3), and their means vary over a wide range (Figure 5.4b). Furthermore, on average the learned policies are substantially slower than the optimal policy (compare the black solid curve to the black dashed one). Nosepoke latencies, however, are on average faster than those in the optimal policy (Figure 5.4d).

The discrepancies between the learned and the optimal policies for interval schedules do not necessarily invalidate our algorithm. As shown in Figure 5.5, the optimal $Q(\cdot,\cdot,\tau)$-value function for interval schedules is quite flat around its maximum, with the optimal latency only slightly better (in terms of expected value) than a wide range of alternative latencies. This is particularly evident for leverpressing in the absence of rewards. Indeed, as mentioned, the structure of interval schedules is such that the rate of responding has only minor influence on the rate of reinforcements, and due to the low costs of responding, a variety of leverpress rates will lead to very similar net rates of reward. In terms of our learning algorithm, this means that interval schedules provide a very shallow gradient around the maximum, with the result being a large variance of the learned policies, albeit with a small (perhaps negligible) price in terms of net reward rate obtained.

5.3.3 Action selection and changes in the expected rate of rewards

In an Actor/Critic model in which the policy is explicitly represented, the net rate of reward does not have an intrinsic role in determining response rates. Nevertheless, in order to tie action selection in the online learning model to the theoretical results on the effect of the net rate of rewards on response rates (Chapter 2), and to our hypotheses regarding the effects of motivation on habitual behavior (Chapter 3) and the relationship between tonic levels of dopamine and response vigor (Chapter 4), we have suggested that the Actor's policy be influenced directly by the net reward rate $\bar{R}$. In our formulation, a change in $\bar{R}$ has an immediate effect on each of the policy's Gamma distributions, through θ_i. We can thus expect the effect of a (dopaminergic) manipulation, in which $\bar{R} \rightarrow a\bar{R}$, to affect the mean latencies of all actions by a factor of

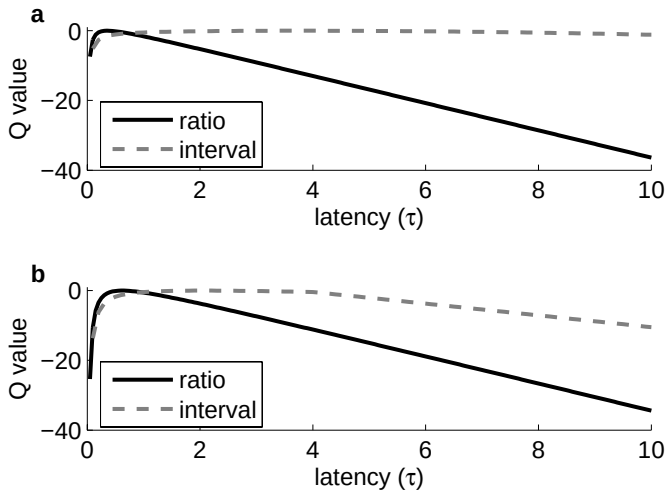

Figure 5.5: Optimal Q-values for the above random ratio and random interval schedules, as derived by relative value iteration (see Chapter 2). Note that these were derived using a model of the one-step transitions and the immediate rewards in the task (ie, using a model-based algorithm), different from our online learning algorithm which does not rely on such knowledge and is completely model-free. **a**. Q-values as a function of latency τ for leverpressing when no reward is available. In solid black are the values for an RR10 schedule, and in dashed gray are the values for an RI30 schedule. The relative flatness of the function for interval schedules around its maximum is apparent in comparison to ratio schedule values. **b**. Q-values as a function of latency τ for nosepoking when a reinforcement is available in the food magazine. In this case the function for interval schedules is not as flat as that for leverpressing, but still it is markedly flatter than that for ratio schedules, resulting in a shallower gradient for our stochastic gradient-climbing algorithm. The optimal stochastic policies in Figures 5.3 and 5.4 were derived from a soft-max operation over these Q-values, with the soft-max temperature arbitrarily chosen to facilitate comparison between policies (see above).

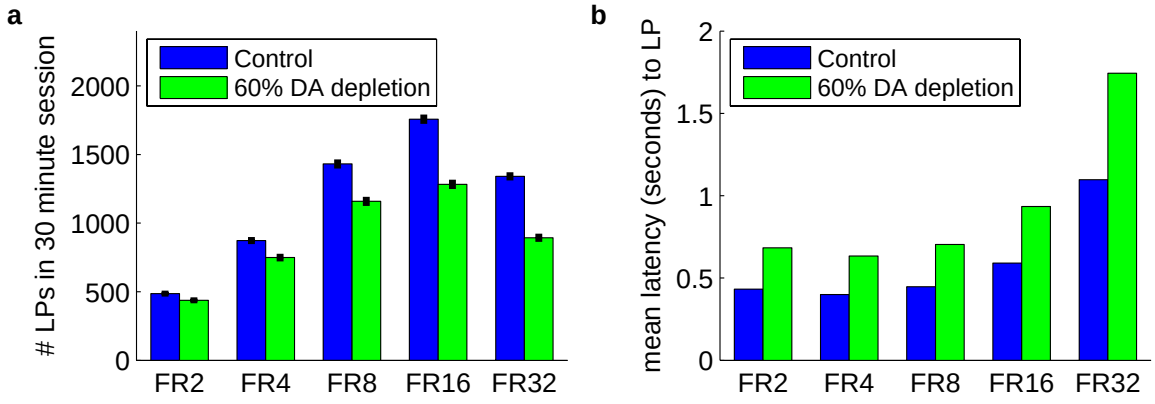

Figure 5.6: Steady state behavior of the Actor/Critic model trained on various fixed ratio (FR) tasks, before and after dopamine depletion. Each bar represents the mean (black: standard error of the mean) of 20 sessions, each 30 minutes long. **a**. The number of leverpresses per session (measured in the absence of new learning) is reduced after 60% dopamine depletion (modeled by reducing $\bar{R}$ to 40% its original value). This effect is more pronounced in higher FR schedules, due to a larger proportion of time spent leverpressing (see Chapter 4, section 4.3). **b**. The effect of this manipulation on the mean latency to leverpress (standard error of the mean negligible) is to increase it by 60% regardless of the schedule requirement.

$\frac{1}{\sqrt{a}}$, for any reinforcement schedule[11].

Figure 5.6 illustrates this effect for fixed ratio schedules, for comparison with the dopamine depletion experiments of Salamone and colleagues (see Chapter 4, figure 4.1). As in Chapter 4, $\bar{R}$ was decreased to 40% of its original value in order to simulate 60% dopamine depletion, and behavior was measured in the absence of further learning in order to distill the effects of dopamine on action selection from those on learning. Similar to the results in Chapter 4, in our online learning model the effect of this manipulation on the overall number of leverpresses in a 30 minute session differs from one FR schedule to another. This is due to the different proportion of time spent leverpressing and eating in each schedule. A more direct measure of mean latency to leverpress shows that, as expected, (a) leverpress latencies increase with the schedule requirement, and (b) the effect of the change in $\bar{R}$ on the mean latency to leverpress is to increase it by a factor of $\frac{1}{\sqrt{0.4}} \approx 1.6$.

5.4 Discussion

In this chapter, we have provided an online reinforcement learning algorithm that uses local information and stochastic experience with a free operant task in order to learn a policy specifying which action to take, and with what vigor to perform it, so as to maximize the accrued net rate of reward. In the interests of biological and behavioral plausibility, we sought an algorithm that allows for soft-max-like (rather than uniform) behavioral stochasticity and generalizes learning from one action latency to similar latencies for the same action. Further, we required that changes in the net rate of reward have an immediate effect on response rates. Here, we have shown that one solution that meets these constraints is an Actor/Critic architecture that explicitly represents an action-latency selection policy parameterized as a mixture of unimodal distributions of latencies. In our algorithm, the Actor learns the parameters of this policy by following the gradient of the expected future rewards, which are learned by the Critic using an online version of relative value iteration.

Our algorithm is 'model-free', that is, it uses experience with the task to sample its properties (the probability of transitioning between and receiving rewards at different states) and construct gradients based on expected long-term rewards, without ever explicitly representing the structure of the task or waiting for the long-term consequences of actions to fully unfold. As such, it is suitable as a model of learning in the habitual controller, which has indeed been associated with an Actor/Critic mechanism in the basal ganglia (see Chapter 1, section 1.4). Like rats, the algorithm can learn a near-optimal policy for ratio or interval schedules within few hundreds of trials. It further demonstrates the differences between ratio and interval learning dynamics: interval schedule learning reaches a stable state faster than ratio schedule learning (that is, within as few as 200 trials[12]), albeit with more variability inherent in the steady state solution.

[11] In interval schedules, the behavioral effect of such a manipulation will not be strictly proportional to $\frac{1}{\sqrt{a}}$, because the altered response latencies affect the stationary distribution of states, and through it the prominence of different components of the Actor's policy in the overall behavior. In ratio schedules changes in response latency do not affect the stationary distribution of states.

[12] For comparison, rats in our variable interval behavioral experiments (Chapter 6) showed stable performance after ten days of training with 30 trials each. Training our algorithm according to the gradual protocol of that experiment (ie, 30 trials continuous reinforcement, then 30 trials on VI2, followed by 60 trials on VI15 and only then VI30 training) speeds learning in the model even further (results not shown).

5.4.1 Relationship to previous work

Although stemming from a wealth of literature on online reinforcement learning, our algorithm is novel in the way in which it deals with both learning in an average reward framework and a continuous action space. Online average reward reinforcement learning is challenging in its own right, but coupled with continuous action selection the challenges are amplified, mainly because of the sensitivity of ARL to the method of exploration (Mahadevan, 1996), and the limitation that a continuous action space places on this.

Previous online ARL algorithms

The first online stochastic ARL algorithm, "$\mathcal{R}$-learning", was suggested by Schwartz (1993a, 1993b) as an analog of online discounted Q-learning that learns a policy that maximizes the rate of rewards rather than the discounted sum of future rewards. The algorithm learns state-action $\mathcal{R}(S,a)$ values and an average reward $\bar{R}$, for an MDP in which discrete actions are chosen at discrete time-points. Analogous to Q-learning, the algorithm's update equations are stochastic approximations to the Dynamic Programming consistency equation for $\mathcal{R}$-values. Singh (1994) suggested a similar algorithm for learning both state and state-action values. Following the relative value iteration Dynamic Programming method, he was the first to suggest clamping a recurrent reference state or state-action value to zero, to aid convergence.

Gosavi (2004a, 2004b) described a Q-learning ARL algorithm for semi-Markov Decision Processes. The algorithm, "Q-P-learning", is based on policy-iteration. It first estimates the reward rate for a given deterministic policy by choosing actions according to this policy with $p = 1 - \varepsilon$ (and uniformly otherwise) and registering the obtained rewards and the passage of time for the policy chosen actions. Then, using this reward rate, it selects actions uniformly and estimates the Q-values of the state-action pairs off-policy. Finally, the policy is improved based on the new Q-values and the process is repeated. Gosavi (2004b) also proved the convergence of stochastic relative value iteration for SMDPs, given a recurrent state and given initialization of the estimated reward rate close to the true reward rate (Abounadi et al. (2000) proved similar convergence for MDPs), thus guaranteeing the convergence of Q-P-learning.

Previous algorithms for ARL with continuous action spaces

The first implementation to combine ARL and function approximation was by Tadepalli and Ok (1996). They approximated state values using a locally linear function which interpolates between exemplar states. Two characteristics of their algorithm are noteworthy in comparison to ours: first, their algorithm uses experience with the environment less efficiently than ours, because only those trials in which an exemplar state is visited are used for learning state values. Second, their algorithm is 'model-based' as, in addition to learning state values, it uses a Bayesian network to learn a model of the environment.

Kimura and Kobayashi (1998a, 1998b) extended Williams' (1992) 'REINFORCE' framework to an Actor/Critic framework. In their algorithm, the Actor represents a parameterized (approximated) policy that is

updated using gradient 'eligibilities' as in REINFORCE, and the Critic provides the 'reinforcement baseline' to the Actor. In contrast to our algorithm, their Actor follows the gradient of the obtained rewards, rather than utilizing the state value estimates learned by the Critic.

Baxter and Bartlett (1999) proposed a specific parameterization of a stochastic policy such that the obtained average reward is a differentiable function of the policy parameters, and proposed an algorithm that performs gradient ascent directly on the average reward, to solve the ARL problem. Their algorithm was not applicable to our problem because it requires a policy that is smooth over the state space, as well as ('model-based') knowledge of the stationary distribution of the Markov chain corresponding to each policy. For large state spaces, Baxter and Bartlett (1999) suggested an approximation to the actual gradient, which was later proven by Kakade (2001) to lead to an average reward which is indeed close to optimal.

It should be noted that, apart from mitigating the problem of action selection, directly approximating a policy rather than action values can also be justified theoretically. Sabes (1993) showed that minimizing an error on Q-values does not necessarily lead to a near-optimal policy (see also Baxter & Bartlett, 1999). Because action selection (and not action evaluation) is ultimately important for maximizing rewards, Sabes advocates minimizing the error on the policy directly, as in REINFORCE. Sabes and Jordan (1996) then suggested an alternative to REINFORCE for learning a parameterized policy. Rather than climbing the gradient of the expected reward, they suggested minimizing the Kullback-Leibler divergence between the distribution of actions as described by the policy and the optimal distribution (as defined by the rewards) through stochastic gradient descent. Their algorithm uses a temperature to set the balance between increasing action choice entropy (exploration) and maximizing reward. By annealing the temperature down during learning, they showed faster and more reliable convergence than REINFORCE.

Finally, the algorithm most similar to ours is "SMART" — an online 'model-free' ARL semi-Markov learning algorithm suggested by Das et al. (1999). Different from our algorithm, 'SMART' is an 'Indirect Actor' method in which a neural network approximator is used to generalize learning of Q-values over a continuous state space. Other differences are that "SMART" uses ε-greedy action selection within a discrete action space and that the reward rate is estimated only once every epoch, from rewards and state-transition intervals accumulated over the previous epoch.

Convergence guarantees

Convergence proofs for stochastic average reward RL have considerably lagged behind the analysis of stochastic discounted RL. Two factors make the case of ARL more complicated: the fact that relative values are defined only up to an additive constant, and the need to estimate the reward rate (Mahadevan, 1996). Although the foundations for the model-based version of ARL and its relationship to Dynamic Programming were provided by Puterman (1994) and Bertsekas (1995a), the extension of their convergence analyses to stochastic online learning has proved difficult. The use of function approximation further limits the convergence properties of our algorithm — even in discounted RL, most value-learning algorithms can fail to converge when using value function approximation (Baird & Moore, 1999; Sutton et al., 2000).

Abounadi et al. (2000) provided the first convergence proofs for two stochastic ARL algorithms based on relative value iteration, under a number of conditions on the learning rate, the frequency of samples of the different states, and the unichain properties of the MDP. While these established the theoretical soundness of stochastic ARL, their proof holds for either synchronous updating of all state-action pairs, or an otherwise evenly distributed updating process.

Baird (1995) suggested a "Residual Learning" algorithm that combines gradient descent on the Bellman residual with conventional TD updating, and is guaranteed to converge even with state-value function approximation. The main drawback of this method is that in stochastic MDPs *two* samples of the successor state are required in order to obtain an unbiased estimate of the (gradient-based) update, so $V(S)$ can only be updated on every second visit to S. For large or continuous state spaces, the time between two visits to the same state might be too long for this algorithm to be a feasible online learning method. Furthermore, convergence of the value estimates is ensured only if the training distribution is fixed, ie, when the policy is not improved until the value estimates converge. Baird and Moore (1999) suggested a family of algorithms ("Value and Policy Search"; VAPS) that converge even when the policy is constantly updated, as long as it is a usually greedy policy with respect to the current value estimates. This approach involves stochastic gradient descent on an average error function weighted by state-visitation frequencies, and is guaranteed to converge to a local minimum if values are updated only at the end of trials. It is suitable for any stochastic policy, as long as states are sampled on-policy and the policy is smooth in the function parameters. Unfortunately, like "Residual Learning", this method requires two samples of successor states for learning in stochastic MDPs.

Finally, Sutton et al. (2000) provided the first convergence proof for policy iteration with a general differentiable parameterized policy and an associated function approximation of the Q-values, for both ARL and discounted RL. Their algorithm climbs a gradient of the expected reward, which is approximated using the Q-values, and so is guaranteed to converge to a local maximum if the values are allowed to converge before the policy is improved (as in REINFORCE). Because both the policy and the value function are approximated, a compatibility condition is required for convergence: the value function must be linear in the features of the derivative of the log of the policy function. As an Actor/Critic implementation of REINFORCE with policy function approximation, our learning algorithm also has similar convergence guarantees. Our implementation updates both Actor and Critic at every trial, and so is prone to oscillations (Baird, 1995), however, we used a high learning rate for the Critic compared to that of the Actor, which has been suggested to improve the convergence of such simultaneous Actor/Critic learning (Konda & Tsitsiklis, 2003).

5.4.2 Interval vs. ratio schedules

Decades of free-operant conditioning experiments have investigated the difference between behavior in interval and ratio schedules, with the observation that animals respond faster on ratio schedules attracting the most attention. Our results highlight two other potential differences: in our model the rate of convergence of learning and the properties of the steady state behavior are different for the two types of schedules.

That interval schedules are easier to learn may not be surprising: it is commonly thought that a ratio sched-

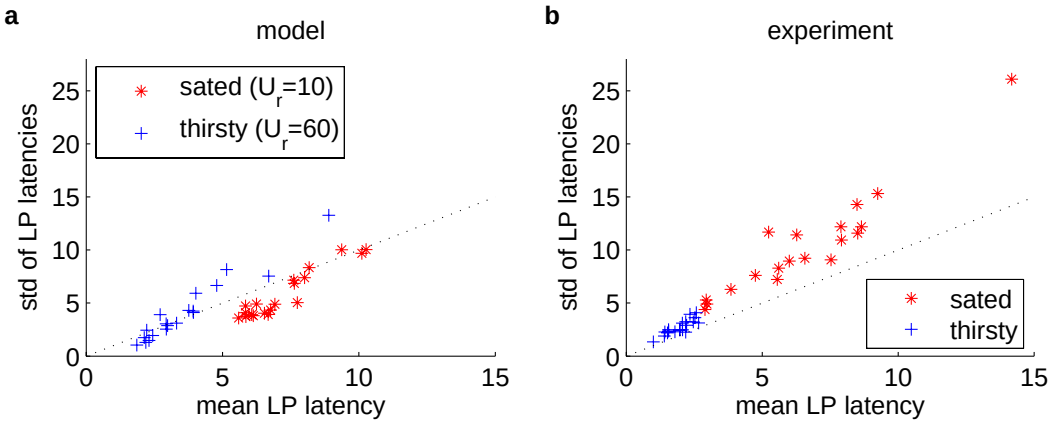

Figure 5.7: Between subject and within subject variability in interval schedule behavior: model and experimental results. **a**. Mean and standard deviation (std) of the latency to leverpress[13] in different runs of the online learning algorithm. Red stars: 19 runs with a low utility reinforcer ($U_r = 10$) simulating sated rats; Blue crosses: 19 runs with a high utility reinforcer ($U_r = 60$) simulating thirsty rats; Black dotted: unity line (mean=std). To ensure that responding is assessed at steady state, each run consisted of $2,000$ trials, with responding measured in the last $1,000$. **b**. Mean and standard deviation of leverpress latency for each of 38 rats trained on a VI30 schedule of reinforcement (see Chapter 6 for details). Latencies were assessed based on the last 150 trials (5 training sessions) out of a total of 450 training trials (15 training sessions – see methods in Chapter 6), by measuring all inter-response intervals ending with a leverpress, excluding the interval between the first nosepoke after a reinforced leverpress and the next leverpress (in order to exclude eating time). Red stars: 19 rats trained that were not food or water deprived (sated); Blue crosses: 19 rats trained after 22 hours of water deprivation (thirsty); Black dotted: unity line (mean=std). In both model and simulation, thirsty rats showed higher response rates (shorter latencies to leverpress) compared to sated rats. Furthermore, there was considerable variability between the response rates of different rats and different runs, as well as high within-subject and within-run variability in steady-state responding. Note that the model parameters were not fit to the behavioral data so a quantitative fit should not be expected.

ule is difficult to learn because with slow responding rewards are almost never received, whereas in interval schedules sporadic responding is more highly rewarded. However, the optimal solution to the reinforcement learning problem in Chapter 2 suggested a contrary view: ratio schedules can be represented using a small number of states, for which the transition dynamics are simple enough (specifically, they are time-independent) that an analytical solution to the Bellman equation can be derived. In comparison, finding an optimal solution to interval schedules is complicated due to non-linear transition probabilities between continuous states. Despite this, our online learning implementation showed that interval schedules are indeed learned faster than ratio schedules. But, rather than appealing to issues of frustration or extinction (which no doubt also exist, but are not exemplified in our simulations), in our model faster learning of interval schedules can be attributed to the relative insensitivity of the optimization criterion (the net rate of reward achieved) to the specific rate of responding. In comparison to ratio schedules, the solution space in interval schedules is larger and less constrained by the schedule dynamics, lending itself well to a stochastic approximation learning method.

The second related observation is that the steady state solution in interval schedules is more variable than that in ratio schedules. This is in contrast to the text-book claim that both random interval and random ratio schedules generate regularly spaced behavior (eg. Domjan, 2003). This is illustrated in Figure 5.7a. In

interval schedules, leverpress latencies are variable both between different runs of our simulation (as seen in the spread of the means of the latencies to leverpress), and within the steady state portion of each run (as evident from the high standard deviation of leverpress latencies). Note that, although a Gamma distribution of latencies is characterized by a standard deviation that is smaller than the distribution mean (mean $= \alpha\theta$, std $= \sqrt{\alpha}\theta$, with $\alpha \geq 1$ in our model), the simulation results show that the standard deviation of response latencies is in fact equal to or higher than the mean response latency, probably due to continuing changes in the Gamma distribution from which behavior is drawn. Mean latencies in ratio schedules were similar for all runs, and their within-run standard deviation was as expected based on a Gamma distribution (not shown). This suggests two predictions: first, steady-state behavior should be more variable (across animals) in interval schedules than in ratio schedules, and second, response rates should show higher *within-subject* variability in interval schedules compared to ratio schedules. The former prediction stems from the large solution space in interval schedules. The second is due to the fluctuating nature of the stochastic solution, which is perhaps caused by an interaction between the schedule dynamics and the different time-courses of policy learning and reward rate learning.

A partial confirmation of this prediction is suggested by Figure 5.7b, which shows the means and standard deviations of leverpress latencies (inter-response times) for each of the 38 rats trained in the experiments described in Chapter 6. Measurements taken from the last five days of training on a VI30 schedule of reinforcement confirm that there is much variability between rats and within each rat's behavior. Indeed the rats' behavior seems more variable than that of our learning algorithm. This is perhaps due to the difficulty in measuring exact response latencies in the rat data — since only leverpress and nosepoke actions are registered, the inter-response intervals may sometimes include other actions whose execution time would be wrongly attributed to the latency to leverpress. Note that the results are portrayed separately for rats trained when sated or when thirsty (the reinforcement used for thirsty rats was a sucrose solution). The mean leverpress latency for each of the thirsty rats is shorter than that for each of the sated rats, demonstrating the prominent effect of motivation on response rates. In accord with Chapter 3, modeling the difference in motivational state as a difference in the utility of the reinforcer replicated this result (Figure 5.7a): 19 runs of our simulation[14] with a reward utility $U_r = 10$ (simulating satiety) resulted in longer response latencies than 19 runs with $U_r = 60$ (simulating hunger).

Note that the pattern of results in Figure 5.7a is not predicated on a specific choice of model parameters or training duration. Qualitatively similar results were obtained when training the model for only 450 trials and measuring response statistics from the last 150 trials (similar to the experimental data). However, some discrepancies between the model and the experimental results may be attributed to a poor fit of the model parameters. For instance, the apparent lower variability in the low utility simulations may have resulted from slower learning (effectively, a lower learning rate) due to the smaller prediction errors in this case[15]. The experimental rats displayed the opposite trend with the responding of sated rats more variable than that of thirsty rats. A higher tendency to choose actions other than leverpressing and nosepoking could be at the

[14] Each simulation represents a different rat, although, different from real rats, the initial conditions and parameters of all the simulations were identical.

[15] To partly overcome this, the learning rate parameters we used for the 'sated' simulations were $\eta_v = 0.1$, $\eta_m = \eta_\alpha = \eta_\theta = 0.05$ rather than $\eta_v = 0.05$, $\eta_m = \eta_\alpha = \eta_\theta = 0.01$ otherwise.

heart of this result, and a poor calibration of the internal reward (the negative unit cost) for 'Other' in our model is perhaps the reason why our simulation did not adopt such alternative responding.

Of course, these results provide only partial confirmation of our prediction, the crucial test for which is a comparison between responding in ratio schedules to that in interval schedules. At present we do not have comparable experimental data from random ratio training, although casual observations of rats leverpressing on ratio schedules are in line with our prediction.

5.4.3 Limitations of the learning algorithm

The algorithm we have proposed has several limitations. First, because we presuppose a parametric shape for the policy, it can learn the true optimal policy only if the optimal policy conforms to this parametric family. Based on both experimental observations regarding the distribution of inter-response latencies and an analysis of the optimal solution of the reinforcement learning problem, we chose to parameterize the policy as a mixture of Gamma distributions. As discussed previously, this family includes the optimal policy for ratio schedules, but not necessarily for interval schedules or other schedules that have a non-linear time-dependent structure. This does not mean, however, that a mixture of Gamma distributions cannot provide a good *approximation* of the optimal policy even in these more complicated schedules. Our algorithm should thus be evaluated based on how well it approximates the optimal policy in a wide variety of problems, rather than whether it can represent the true optimal policy for any specific case. We note also that the ubiquity of Gamma distributed response times in experimental data (Rohrer & Wixted, 1994; Ratcliff et al., 1999, to give but two examples from different domains) suggests that a similar limitation might exist in biological learning and action selection.

A second, more crippling, limitation is due to the use of an Actor/Critic architecture that explicitly learns a behavioral policy, rather than the values of different actions at different states. Action values represent real aspects of the problem (ie, the amount of future expected reward) on a true ratio scale. In contrast, a policy orders actions according to their preferability, not necessarily maintaining anything but an ordinal relationship between them. Thus, if two actions have very similar values, Q-value based action selection will tend to perform both. Conversely, the Actor/Critic algorithm climbs gradients toward an optimal *deterministic* policy (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), and so it will eventually perform the better of the two actions exclusively. Information regarding how much worse are the suboptimal actions is lost, and an action that is only slightly worse than the optimal one will not be treated differently than one that would lead to substantially fewer long-run rewards. As a result, whereas in Q-value based action selection a temperature parameter can explicitly control the degree of exploration, in Actor/Critic methods there is no such control, and exploration of suboptimal actions gradually stops. This is especially detrimental in terms of adapting to changes in the task: because the policy does not continue to sample suboptimal actions, changes in the values of these will go unnoticed.

This limitation has two consequences, a technical one regarding the stability of the algorithm, and a behavioral one regarding the fit of the model to experimental observations of suboptimal behavior. Take, for

example, a policy in which the probability of taking the action 'Other' is close to zero (as was frequently the result of our learning algorithm). Because of this, states in which (a_{prev} = Other) will rarely be sampled, and so their values will not be learned accurately. In practice this may be of little consequence — these states are not reached because they are not advantageous, and so their values are not important. However, this can affect the stability of the learning process: because in ARL the values of states are constantly updated *relative* to each other, the relative value of these rarely-visited states can be grossly inaccurate, in which case a rare visit to them can result in a very large prediction error that will destabilize the learning process[16].

Behaviorally, in Chapter 2 we discussed the excessive nosepoking of rats trained on interval schedules, which seemed to occur even though the rats knew that no reward was available to be harvested. Casual observations of rats trained on ratio schedules seem to indicate that this phenomenon is specific to interval but not ratio schedule behavior. One possible reason for this is suggested by the Q-values of these actions in the two schedules. In interval schedules, leverpress and nosepoke actions are similarly useful (or useless) in times in which a reward is not baited. As a result, in a schedule in which the baiting probability is quite low (for instance, the RI30 schedule we have used), the difference between the Q-values of leverpressing and nosepoking will be quite small. An algorithm which learns these values from stochastic (noisy) sampling of the task dynamics may find it difficult to learn such a small difference in value. This would result in similar rates of leverpressing and nosepoking in interval schedules, as is seen in the experimental data. In ratio schedules where the values of the two actions are very different, this problem would not exist.

Our algorithm does not learn Q-values, and thus is not prone to this problem. Excessive nosepoking is not seen, and furthermore, 'Other' actions are not performed. This may be optimal, but it is not in accord with the experimental observations. One solution is to use a hybrid learning process which learns a Q-value for each action in each state (averaging over the different latencies with which it is performed), and a parameterized policy (a Gamma distribution) for the latencies. In this algorithm the mixture parameters $m(S,a)$ are replaced by $Q(S,a)$-values, and action selection involves two stages. A soft-max over the actions values is used to choose which action to perform, with the corresponding Gamma distribution determining the latency with which the action is performed. While not adding to the complexity of the algorithm (the Q-values can be learned using a similar prediction error) or to the number of learned parameters, this allows for better control of exploration and for a representation of the relative value of different actions that is true to the rewards that are consequent on their performance. Such an algorithm might thus capture the difference between interval and ratio schedules in terms of excessive nosepoking in the absence of rewards.

Finally, a conceptual limitation of our algorithm is that incorporating the net rate of reward into action selection is not normative or necessary. It is not surprising that action selection in a model-free algorithm does not necessitate global terms such as an overall reward rate. However, immediate responding to changes in the net rate of reward are a tantalizing medium for the effects of motivational states on response rates.

[16]To partly alleviate this, in our implementation we accumulated value updates over the course of all the actions leading to a reward, and updated according to their sum only once a reward was obtained. This prevented a constant drift (proportional to η_v^2) of the values of the visited states away from those of the rarely visited states. However, this could not prevent a drift resulting from the gradual growth of the estimated average reward rate during the early stages of learning, which could still cause a large discrepancy in state values.

Experimental results suggest that these effects are immediate and do not require new value or policy learning even in the habitual system (see Chapters 3 and 4), supporting a direct online dependence of action selection on aspects other than learned values or policies. The potential link between response rates and tonic levels of dopamine is a further reason to preserve the online effect of the net rate of reward on latency selection.

5.4.4 Conclusions and future directions

In sum, we have provided an online framework that implements online learning and action selection of response rates in free operant situations, in which response rates can be immediately affected by changes in motivational state or in the tonic level of dopamine. Our algorithm is based on the classic Actor/Critic framework, in which a temporal difference error signal is used to train values in the Critic and policies in the Actor.

In terms of modeling, it would be interesting to incorporate into our framework more complex aspects of temporal difference learning, such as learning with eligibility traces (TD(λ) learning). As mentioned, another natural extension is to combine Q-value learning for actions with policy learning for latencies. A third area of future investigation is the impact of using a state space that is *not* suitable for the schedule, on the resulting behavior. For instance, optimal behavior on concurrent-interval schedules dictates that the probability of switching increase the longer the animal has not sampled one of the options. Contrary to this, animals' behavior shows a constant switch probability. Daw and Touretzky (2001) suggested that this results from a failure to take into account the time that has elapsed since each schedule was sampled, ie, from a failure to represent the correct Markovian state space. Incorporating only partial information into the state space transforms the Markov decision process into a partially observable Markov decision process (POMDP). In fact, it is likely that most real-world problems are solved as POMDPs, if only because the dynamics of the environment are rarely fully Markovian. Finding the optimal solution to POMDPs is notoriously intractable (Cassandra et al., 1994; Kaelbling et al., 1998). Our online framework provides a naïve solution, one that treats the problem as if it were a simple MDP, rather than keeping track of a belief about the real state of the world, as is necessary in order to find the true optimal policy for a POMDP. Given the simplicity of this (wrong) solution, it is interesting to know to what extent it can account for animal behavior in free-operant schedules.

Last, we have applied our algorithm to both ratio and interval schedules of reinforcement, and have stressed the differences between the learning processes and the resulting behavioral policies in these two types of schedules. An interesting proposition is that these differences may, in some way, help explain why behavior on random interval schedules seems to habitize faster than ratio schedule behavior (Dickinson & Nicholas, 1983b; Dickinson et al., 1983; Dickinson, 1985; Dickinson et al., 1995). While a normative explanation of other factors that promote or hinder habitization of responding has been provided in terms of relative uncertainty in habitual and goal-directed controllers (Daw et al., 2005), the different rates of habitization of ratio and interval schedules has not been explained. Reduced uncertainty of the habitual controller in interval schedules as a result of fast convergence of the net reward rate (as compared to ratio schedules), may provide the necessary link to a normative answer.

Motivation is what gets you started, habit is what keeps you going – "Fitness First" gym

Chapter 6

Motivation and habitual behavior: An empirical investigation

Abstract: In recent years much experimental investigation into the effects of motivation on instrumental behavior has shown that motivation affects goal-directed behavior through determining the utility of its outcomes. How motivation influences habitual behavior is still rather unclear. The defining feature of habitual behavior, its insensitivity to changes in the value of its outcomes, seems to suggest motivational insensitivity. However, using a normative model of the effects of motivation on habitual action selection, we have suggested (in Chapter 3) that motivation can affect habitual responding through an outcome-general 'energizing' influence. Here, we attempt to test this hypothesis experimentally. Because previous studies have only considered motivational downshifts, they cannot distinguish between 'energizing' effects and effects of generalization decrement as a result of a motivational shift. To remedy this, we investigated the effects of motivational up-shifts and side-shifts on habitual behavior. In Experiment 1, hungry rats were trained to leverpress for sucrose solution. Following a side-shift from food- to water-deprivation, rats showed less leverpressing in extinction compared to non-shifted controls, although a subsequent consumption test found no differences in sucrose consumption between thirsty and hungry groups. In Experiment 2, undeprived rats were trained to leverpress for either sucrose solution or sucrose pellets. A post-training up-shift from satiety to water-deprivation did not affect leverpressing in extinction, regardless of outcome identity, although free consumption of sucrose solution, but not of pellets, was enhanced. Together, these results suggest that, as hypothesized in Chapter 3, a motivational shift affects habitual behavior through a combined effect of 'generalized drive' and generalization decrement, but not through determining the value of the outcome. The absence of an outcome-specific effect is in line with cached value-based rather than forward-model based control of habitual behavior.

6.1 Introduction

A wealth of recent experiments has been devoted to the motivational control of conditioned behavior, revealing a rich and intricate tapestry of effects (for a review see Dickinson & Balleine, 2002). Contrary to the intuition that motivation directly modulates the propensity to behave, as has been suggested for Pavlovian responses (eg, a hungry dog will salivate to a stimulus predictive of food, while a sated one will not; Mackintosh, 1974), motivational states have been shown to affect moderately trained, goal-directed, instrumental actions indirectly, by determining the incentive value of the outcome of the behavior (Adams & Dickinson, 1981; Balleine, 1992; Lopez et al., 1992; Dickinson & Balleine, 1990, 1994; Dickinson et al., 1995; Balleine et al., 1995; Balleine & Dickinson, 1998, 2000; Dickinson & Balleine, 2002). This dependence on outcome value to mediate the effect of motivational states is, at least to some extent, explicit, in that the altered value needs to be experienced for the effects of a motivational shift to be manifest. For example, behavior aimed at acquiring food will be enhanced when hungry rather than sated, only if the subject has experienced the enhanced incentive value (desirability) of the food in this state (a process termed "incentive learning", eg, Dickinson & Dawson, 1988, 1989; Balleine, 1992, 2000). Such effects have been demonstrated for a wide range of motivational shifts – from hunger to thirst (Dickinson & Dawson, 1988, 1989), from thirst to hunger (Dickinson & Balleine, 1990; Balleine, 1992), from hunger to satiety (Balleine, 1992; Balleine et al., 1995; Balleine & Dickinson, 2000), from satiety to hunger (Balleine, 1992), from thirst to satiety (Lopez et al., 1992) and even using aversive motivational states (Henderson & Graham, 1979).

These results parallel those from a different line of experimentation, in which the value of one specific outcome is manipulated after training has commenced, without changing the general motivational state of the animal. This is done, for instance, by conditioning taste aversion to the outcome through pairing its consumption with the induction of gastric illness (eg, Adams & Dickinson, 1981; Adams, 1982), or by inducing specific satiety through prefeeding of the outcome (eg., Balleine & Dickinson, 1988, 2000; Killcross & Coutureau, 2003). When behavior is then tested in extinction (ie, with no rewards available, in order to probe the effects of the manipulation on previously encoded relationships rather than on new learning), these studies show that given appropriate incentive learning, outcome devaluation causes a reduction in performance of goal-directed instrumental actions compared to non-devalued controls (Balleine & Dickinson, 1991).

Interestingly, these latter studies also demonstrate that with *more extensive training* instrumental behavior can become *independent* of the value of its consequent outcome (Dickinson, 1985). Thus, when instrumental actions are over-trained, behavior becomes *insensitive* to such post-training outcome devaluation, and subjects perform the trained action at comparable rates whether the outcome is devalued or not, regardless of incentive learning (ie, even when they are given explicit experience with the outcome's modified value, eg. Adams, 1980; Dickinson, 1994; Killcross & Coutureau, 2003). This change in outcome sensitivity has been postulated to result from a shift in the underlying associative structure controlling behavior, from response-outcome (R-O) goal-directed (Tolman, 1949b), to stimulus-response (S-R, Thorndike, 1911) *habitual* control (Dickinson, 1985; Dickinson & Balleine, 1994; Dickinson et al., 1995; Balleine & Dickinson, 1998; Killcross & Coutureau, 2003, but see Colwill & Rescorla, 1985, 1986, 1988).

Daw et al. (2005) relate these two forms of behavior to forward-model and cached-value strategies for action control, respectively, and show why such a pattern of devaluation sensitivity is expected for each controller. Furthermore, by assuming that the brain bases action selection on the more accurate of the two controllers (that is, the one that can evaluate candidate actions with less uncertainty), they provide a normative explanation for why goal-directed behavior is dominant after moderate training, but extensive training induces a transition to habitual control. Their model also suggests a new interpretation of the effects of 'incentive learning'. According to this, outcome devaluation causes the goal-directed (but not the habitual) system to be less certain of its action value estimates, and as a result control is delegated to the habitual system. 'Incentive learning' reduces the goal-directed system's uncertainty, shifting control back. The devaluation-insensitivity of responding prior to 'incentive learning' actually results from *habitual control* rather than from insensitivity on the part of goal-directed action selection.

Similar to the psychological S-R accounts, this model assumes that habitual behavior is inherently insensitive to outcome revaluation. However, in Chapter 3, we postulated that despite its insensitivity to the revaluation of specific outcomes, habitual behavior can be sensitive to motivational manipulations, albeit in an outcome-general way. Specifically, building on the model we presented in Chapter 2, we suggested that changes in motivational state should affect the rates of all habitual actions, in a form of 'generalized drive' or 'energizing' effect. In this chapter, we describe the results of two experiments aimed at testing this hypothesis directly.

6.1.1 The possible effects of motivation on behavior

First, let us step back and consider all the theoretically possible ways by which a motivational shift can affect behavior (be it habitual or otherwise). Perhaps most intuitive is a *direct modulation* of the propensity to act, which could be dependent on the identity of the outcome, as has been postulated in Pavlovian behavior (Dickinson & Balleine, 2002). An alternative *outcome-specific* form of motivational control is by determining the value of the goals of behavior ('incentive motivation'; Tolman, 1949a, 1949b), such that behavior toward more valuable or desirable goals would be enhanced, while that for less desired goals would be reduced. Important for our discussion, both these effects would be manifest in the so called "directing" aspect of motivational control, which we will henceforth refer to as outcome-specific motivational control.

A second possible route to controlling behavior, in terms of the 'energizing' aspects of motivation, was proposed by Hull in his Generalized Drive hypothesis (eg, Hull, 1943; Brown, 1961; Bolles, 1967). According to this, motivational states generate a certain 'drive' which is applicable to many kinds of on-going behavior. Thus, as a result of *reduced generalized drive*, sated rats may be less inclined to perform *any* pre-potent action in the experimental situation, including leverpressing. Importantly, this effect is outcome-general, and depends only on the internal motivational state, such that generalized drive would, in general, be lower for satiety than for hunger.

Last, post-training shifts in motivational state can influence behavior as a result of a *generalization decre-*

	Down-Shift (hunger → satiety)	**Up-Shift** (satiety → thirst)	**Side-Shift** (hunger → thirst)
Outcome-Specific (directing effect)	↓	↔ or ↑ pellets solution	↓ or ↔ pellets solution
Generalized Drive (energizing effect)	↓	↑	??
Generalization Decrement	↓	↓	↓

Table 6.1: Possible effects of motivation on behavior, and their prediction for instrumental behavior tested after motivational down-shifts, side-shifts and up-shifts. Arrows illustrate a reduction, increase, or no predicted change in rate of behavior as compared to unshifted controls. Note that the prediction regarding the drive effect for the side-shift is underdetermined, as there is no clear independent measure of the relative drive induced by hunger versus thirst. For simplicity, and as both effects are presumably outcome-specific and thus have similar predictions for behavior, direct control and control through determining the incentive motivation of the outcome have been combined.

ment from the learned context (which may include the motivational state which was in effect during training) to the context in effect when behavior is tested (Brown, 1961; Dickinson & Balleine, 2002). Note that this effect is not only outcome-independent, but is also not dependent on the identity of the motivational state. Thus, generalization decrement predicts that *any shift in motivation*, even an up-shift from a non-motivated to a motivated state (such as from satiety to hunger), will induce a generalization decrement, which will result in a *reduction* in conditioned behavior (Brown, 1961).

Note that these potential effects of a motivational shift are not at all mutually exclusive. However, they may predict different directions of change of behavior, as a result of their different dependencies on the identity of the outcomes and motivational states. Table 6.1.1 illustrates the predictions of these effects for qualitatively different motivational shifts - a down-shift from a deprived to an undeprived state (eg, from hunger to satiety), an upshift from an undeprived to a deprived state (eg, from satiety to thirst), and a side-shift between two different deprivation states (eg, from hunger to thirst). Predictions are illustrated for behavior trained with either sucrose pellets (relevant to food-deprivation only) or sucrose solution (relevant both to a food deprived and a water deprived state).

Clearly, a motivational down-shift confounds the above three modes of influence, as they all predict a reduction in behavior. Thus, to contrast and differentiate between these qualitatively different possible effects of motivational shifts on habitual behavior, in the present study we tested the effects of motivational side-shifts and up-shifts on extensively trained lever pressing. The design roughly followed Balleine's (1992) study of the effect of motivational shifts on moderately trained behavior, albeit with more extensive training using a training protocol specifically aimed at promoting habitual behavior. In accordance with Adams (1982), Dickinson and Nicholas (1983b), and Holland (2004), we trained rats to perform a single action to obtain a single outcome, using a random-interval schedule of reinforcement, and training for fifteen sessions.

Since we are primarily interested in instrumental control, we furthermore sought to minimize any influence of Pavlovian motivational effects on our results, for instance via control exerted by the context (eg, a mo-

	Training (15 sessions)	**Test** (extinction)
Exp. 1	Hungry: LP (VI30) → sucrose solution	Hungry or Thirsty
Exp. 2	Sated: LP (VI30) → sucrose pellets or solution	Sated or Thirsty

Table 6.2: Design of Experiments 1 and 2. LP=lever press; VI=variable interval schedule or reinforcement.

tivational shift can affect Pavlovian approach to the lever which would then indirectly affect instrumental leverpressing, Dickinson & Nicholas, 1983a; Dickinson & Dawson, 1987). A standard route to eliminating the effect of these is to compare responding on two different levers trained in the same context for two different outcomes (one being the motivational target). This equates Pavlovian contextual influences on responding on the two levers, and allows the isolation of the instrumental motivational control (eg. Dickinson & Dawson, 1989). Unfortunately however, training with two actions and two outcomes has been shown to produce behavior resistant to habitization (Holland, 2004; see also Daw et al., 2005 for a normative explanation of this). Therefore, following Coutureau and Killcross (2003), we used post-training 'context extinction' sessions. In these, the animals were put in the training chamber in the absence of levers and of outcomes, to extinguish possible Pavlovian associations between the experimental context and the outcome, while avoiding extinction of the leverpressing instrumental behavior.

6.2 Experiment 1: Motivational side-shift

In the first experiment, food-deprived rats were extensively trained to leverpress for a sucrose solution outcome. They were then tested under extinction conditions while either food-deprived (group HUNGRY) or water-deprived (group THIRSTY). Importantly, the sucrose solution was a motivationally relevant outcome in both these deprived states, as a result of its liquid as well as nutritional properties. This design (see Table 6.2) allowed us to differentiate between an outcome-dependent form of motivational control, according to which there should be no change in the behavior, since as the outcomes were relevant to both the training and the test motivational states, and the generalization decrement hypothesis which predicts a decrease in conditioned behavior as a result of the motivational shift itself. The effect of 'generalized drive' could not be determined in this experiment, and were therefore the focus of Experiment 2.

Because previous work has demonstrated the importance of incentive learning for motivational control of moderately trained behaviors (Adams & Dickinson, 1981; Dickinson & Dawson, 1988, 1989; Balleine, 1992), we also assessed the effects of incentive learning, by exposing one group of rats (group THIRST+IL) to the sucrose solution under thirst, before instrumental training began (similar to Balleine, 1992). Inclusion of this group controlled for possible effects of an unknown outcome value, on habitual behavior. Note that we used a shift from hunger to thirst, rather than from thirst to hunger, as previous studies have shown that water deprived rats reduce consumption of the dry lab chow, and thus, over time, also become food-deprived (Dickinson & Balleine, 1990). By depriving the rats of water for only one day prior to the test, it is likely that the dominant deprivation was that of water rather than food.

6.2.1 Materials and Methods

Subjects and apparatus

Twenty male Sprague Dawley rats (Harlan Laboratories, Jerusalem, Israel) approximately three months old, weighing 329-473 grams (mean 362g) were housed 3-4 to a cage, in a vivarium maintained on a 12-hour light-dark cycle (lights on 15:00-3:00). All behavioral training and testing occurred during the dark portion of the cycle. Animals were allowed one month familiarization with the vivarium before training began. During training, rats were maintained on a 22-hour food restriction schedule, with tap water available *ad lib* in the home cage. Under this schedule, food (standard lab chow) was provided in the home cage for two hours each day, always after the daily treatment/session. The rats were weighed twice a week to ensure that their body weight did not decrease below 90% of their free-feeding weight. During testing and incentive learning stages, some of the rats were shifted to a 23-hour water restriction schedule (see below), with food freely available, while the rest were maintained on the food restriction schedule. All animal research was carried out according to the guidelines of the Institutional Animal Care and Use Committee of Tel Aviv University. All efforts were made to minimize the number of animals used and their suffering.

Behavioral training and testing was conducted in four operant chambers (Campden Instruments, Loughborough, UK) fitted with a recessed food magazine and two retractable levers. The levers were 4 cm wide and were positioned 2.8 cm from the side walls, 7.5 cm from either side of the food magazine, and 5 cm from the grid floor. Only the left lever was used in this experiment, and the right lever remained retracted at all times. The chambers could be illuminated by a house-light located at the ceiling. Access to the food magazine was through a hinged Perspex panel, the opening of which activated a microswitch. A peristaltic pump (RS Components, Northants, UK) attached to a silicon tube inaccessible to the rats, delivered approximately 0.25ml of 20% sucrose solution (hand mixed) into the food magazine, over a period of 1 second. The operant chambers were housed in sound-attenuating boxes, and ventilating fans were mounted on the side of each box. Equipment programming and data recording were computer controlled by ABET software (Lafayette Instrument Co., Indiana, USA). Incentive learning and familiarization with the sucrose solution took place in the home cages, where the solution was placed in a small plastic dish. The consumption test took place in individual feeding cages made of white plastic measuring 13x13x30cm, with wire-mesh ceilings, which were fitted with plastic bottles containing sucrose solution. The bottles and their location in the wire mesh were similar to those used for providing water in the home cage.

Procedure

Handling and incentive learning On Days 1-3, the rats were individually handled for about 2 minutes daily. The rats were first divided into three groups: group HUNGRY would be tested under food deprivation, groups THIRSTY and THIRST+IL would be tested under water deprivation, with group THIRST+IL also receiving an incentive learning treatment prior to training. A 22-hour food restriction schedule for groups HUNGRY and

THIRSTY and a 23-hour water-restriction schedule for group THIRST+IL began one day prior to handling.

To reduce neophobia to the sucrose solution, after each daily handling session and prior to the daily feeding, rats were pre-exposed to sucrose solution in their home cages for ten minutes. A small plastic dish containing 20% sucrose solution was placed in the home cage. All rats were observed to consume solution in the ten-minute exposure. For group THIRST+IL this stage also constituted the incentive learning pre-exposure stage, as they were exposed to sucrose solution in a water-deprived state. Approximately 30 minutes after this exposure, water bottles were put in their home cages for an hour. This procedure was repeated until the last handling/pre-exposure day, after which all rats were returned to an *ad lib* water and 22-hour food restriction schedule, which would be maintained throughout all subsequent training (until day 20).

Magazine training. On Days 4-5, rats were trained to consume sucrose solution from the food magazine in the operant chamber, with the lever retracted. The session began with the onset of the house-light, which remained on for the entire session. Sucrose solution was delivered into the food magazine on a random time schedule, with a uniformly variable delay of 30-90 seconds (mean 60 seconds). On Day 4, the magazine flap was taped back so that the magazine was constantly open and the sucrose solution was easily reachable. Training continued until twenty-five outcomes were delivered, after which the session ended and the house-light was turned off. On Day 5, the session ended after twenty outcomes had been collected (as measured by the insertion of the rats' head into the food magazine), or until twenty-five outcomes had been delivered. The number and timing of head insertions into the food magazine were recorded.

Leverpress training. On Days 6-18, rats were trained to press the operant lever in order to obtain sucrose solution in a free operant procedure. The beginning of each session was signaled by the onset of the house-light and the insertion of the left lever. On Day 6, each press on the lever delivered an outcome into the food magazine. Individual shaping was used in this session to assist in acquisition of the leverpress response. On Days 7-18, a variable interval schedule of reinforcement was introduced, in which an outcome was delivered only for the first leverpress after the programmed interval had elapsed. The mean interval was 2 seconds on Day 7, 15 seconds on Day 8, and 30 seconds for the following twelve training sessions. Reinforcement was equally likely in the interval starting from half the nominal value of the schedule and ending after 150% of the nominal value of the schedule had elapsed. Immediately after the delivery of an outcome, a new interval within these boundaries was drawn, and the interval timer was started. All sessions terminated when thirty outcomes had been delivered, except the session on day 6 which ended when thirty outcomes had been collected (as measured by insertion of the rat's head into the food magazine prior to the delivery of the next outcome), or when forty outcomes had been delivered. The lever was then withdrawn, and the house-light was turned off. Days 12 and 15 included two training sessions each, separated by at least one hour, such that overall, a total of fifteen leverpress training sessions were given over thirteen days. Both leverpresses and head insertions into the food magazine (nosepokes) were recorded throughout training[1].

One rat in group HUNGRY refused to consume the sucrose solution from the food magazine, and was dropped from the study. Another two rats, one in group THIRSTY and one from group HUNGRY, required extra days

[1] See Chapter 2 (sections 2.1.2 and 2.2) for an analysis of the training behavior from Experiments 1 and 2.

for magazine training and/or lever press training, and thus received only seven sessions of training on the 30sec random interval schedule, instead of twelve. Results omitting these rats were no different from those reported below, and thus they were not omitted from the study. This resulted in final Ns of 6 rats in the HUNGRY and THIRST+IL groups and 7 rats in the THIRSTY group.

Context extinction. On Days 19-20, all rats underwent three sessions of context extinction. The sessions began with the onset of the house-light and were terminated twenty minutes later with its offset. Throughout these sessions, the levers were retracted and no outcomes were given. Two sessions were given on Day 19, separated by at least one hour, and one session was given on Day 20. Note that these sessions did not extinguish the learned leverpress behavior, as the lever was not available for pressing.

Motivational shift. Immediately after the last context extinction session (i.e., on Day 20), food and water were returned to the home cages for 24 hours. The next day (Day 21), water was removed from the home cages of groups THIRSTY and THIRST+IL, and food was removed from the home cages of group HUNGRY. This procedure was intended to ensure that the rats on a water-deprivation schedule in test did not also remain food-deprived as a result of low consumption of the dry lab chow in the absence of water. Thus all rats were allowed to consume food and water *ad lib* for one day before the appropriate test deprivation state was established.

Extinction and Consumption tests. On Day 22, after at least twenty-two hours of food/water deprivation, rats were tested for leverpressing behavior in extinction. The test session started with the onset of the house-light and the insertion of the left lever, and terminated fifteen minutes later with the retraction of the lever and the offset of the house-light. Lever pressing had no programmed consequences in this session, and no outcomes were delivered throughout. Leverpresses and head insertions into the food magazine (nosepokes) were recorded. Immediately after the extinction test, a consumption test ensued. Rats were placed in individual cages and allowed to drink sucrose solution freely from a plastic bottle, for an hour. The amount of sucrose consumed was computed by weighing the bottles before and after the test.

6.2.2 Results

Analysis of the results was performed using one-way analysis of variance (ANOVA), with a main factor of group and an alpha level of 0.05. In the last session of leverpress training, the three groups exhibited similar response rates (mean (SD) leverpress rates for groups THIRSTY, THIRST+IL and HUNGRY were 20.4(5.6), 23.9(8.5) and 17.7(5.5) presses per minute, respectively, $F(2,16) = 1.3$; mean (SD) rates of nosepoke rates were 13.7(3.9), 18.8(10.6) and 16.4(10.0) magazine entries per minute, respectively, $F(2,16) < 1$). The consumption test showed that the amount of sucrose consumed was similar in the three groups (Figure 6.1a; $F(2,16) = 1.17$), suggesting that the sucrose solution could be regarded as a motivationally relevant outcome for both the food-deprived and the water-deprived groups.

The results of prime interest are from the extinction test. Figures 6.1b and 6.1c show the leverpress and mag-

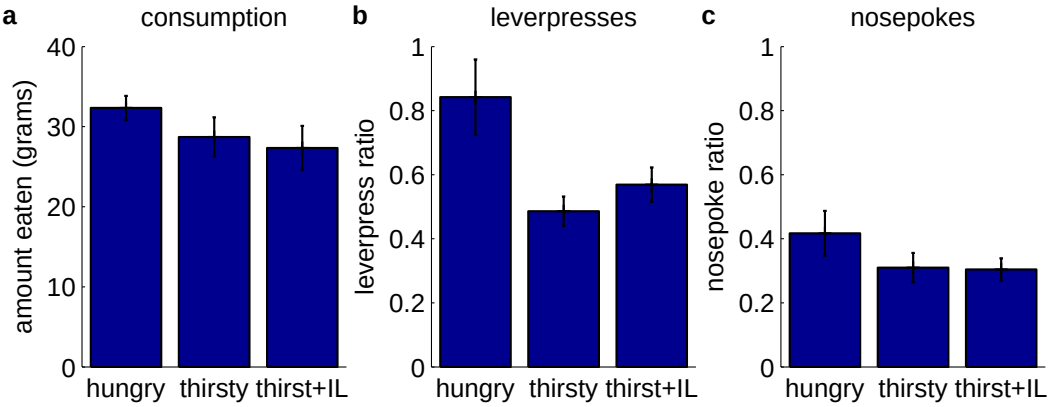

Figure 6.1: Experiment 1 results. **a.** Mean amount (in grams) of sucrose solution consumed by each of the three groups in the consumption test. **b.** Mean rate of leverpresses per minute in each group in the extinction test, expressed as a percentage of the leverpress rate in the last training session. **c.** Mean rate of nosepokes per minute in each group in the extinction test, expressed as a percentage of the nosepoke rate in the last training session.

azine entry results from this test stage. To reduce between-subject variability, response rates were expressed as a percentage of the responding in the last training session. As can be seen, there were no differences between the groups in the rate of magazine entry during the test (Figure 6.1c; $F(2,16) = 1.5$). In contrast, water-deprived rats showed *less* leverpress responding than food-deprived rats, regardless of whether they had been given an opportunity for incentive learning or not (Figure 6.1b; $F(2,16) = 5.755$, $P < 0.05$ for main effect of group; Student Newman-Keuls post-hoc comparisons showed that both the THIRSTY and THIRST+IL groups were significantly different from the HUNGRY group, however, they were not significantly different from each other). Although the consumption results showed no significant differences between treatment groups, there was a small tendency of the shifted rats to consume less sucrose solution than the non-shifted controls. To assess whether the effect of the motivational shift on leverpressing behavior was confounded by this tendency, an analysis of covariance (ANCOVA) was performed on the leverpress data using the amount consumed as covariate. This analysis yielded a similar main effect of group ($F(2,15) = 5.26$, $P < 0.05$).

These results show that rats that had undergone a motivational shift between training and testing *reduced* their instrumental leverpress performance compared to rats whose motivational state had not been shifted, although they did not reduce their subsequent consumption of the sucrose solution. This effect was prominent regardless of an incentive learning treatment, and thus cannot be attributed to a lack of knowledge of the high incentive value of the sucrose solution under water-deprivation. The finding that the difference between the groups was significant also when using the amount of sucrose consumed as a covariate, indicates that differences in the 'desirability' of the outcome cannot account for the reduced rate of leverpressing in the water-deprived groups. Moreover, this reduction cannot be explained by a competition between leverpresses and nosepokes because, if anything, the shifted group tended to perform slightly less magazine behavior than the control group. The reduction in leverpressing is thus in line with the generalization decrement hypothesis, rather than an outcome-value dependent motivational effect of responding.

6.3 Experiment 2: Motivational up-shift

Experiment 1 provided support for the generalization decrement hypothesis, however, it could not provide evidence for or against the 'generalized drive' hypothesis because the relative drive associated with hunger and thirst could not be assessed (and, in any case, may have been equated). Experiment 1 also could not provide conclusive evidence against the outcome specific hypothesis, because only one outcome was used. To further investigate the role of these in the motivational control of habitual behavior, in Experiment 2 we studied the effects of a motivational up-shift on behavior trained with one of two different outcomes (see Table 6.2). Rats that were neither food deprived nor water deprived (ie, sated) were trained to leverpress for either sucrose pellets or sucrose solution, and tested under extinction conditions, with half the rats tested in a water-deprived state (groups PEL/THIRSTY and SOL/THIRSTY) and half the rats tested in a nondeprived state (groups PEL/SATED and SOL/SATED). Importantly, as a result of its liquid property, the shift to water deprivation was intended to enhance the value of the sucrose solution for group SOL/THIRSTY compared to that for non-deprived rats (group SOL/SATED), while not affecting the value of sucrose pellets, which should be similarly desirable for shifted and non-shifted rats (groups PEL/THIRSTY and PEL/SATED, respectively). Outcome-specific motivational control of behavior would thus predict an increase in behavior in the shifted rats trained with solution, but no change in the behavior of the rats trained for pellets.

The motivational upshift also provides a testable prediction regarding the effect of 'generalized drive'. Because water deprivation should induce a higher drive state compared to satiety, if 'generalized drive' plays a significant role in the motivational control of habitual behavior, water-deprived rats should elevate their responding compared to non-deprived rats, and this effect should be outcome-general. In contrast, the generalization decrement hypothesis predicts the opposite direction of change — a reduction in behavior as a result of the motivational upshift, again, regardless of the identity of the outcome for which the animals are working, or its relevance to the deprived motivational state.

6.3.1 Materials and Methods

Subjects and apparatus

Twenty male Sprague Dawley rats (Harlan Laboratories, Jerusalem, Israel) approximately three months old, weighing 323-389 grams (mean 361g) were housed four to a cage, in the same vivarium described in Experiment 1. Animals were allowed one month familiarization with the vivarium before behavioral training began. Throughout training, rats were maintained with tap water and standard lab chow available *ad lib*. During testing half of the rats were shifted to a 23-hour water restriction schedule (see below), with food freely available. All animal research was carried out according to the guidelines of the Institutional Animal Care and Use Committee of Tel Aviv University. All efforts were made to minimize the number of animals used and their suffering.

The apparatus was the same one used in Experiment 1. In addition to the use of the peristaltic pump to deliver sucrose solution, a pellet dispenser delivered 45-mg, "dust-free" sucrose pellets (Noyes, Sandown Chemical Limited, Hampton, England) into the same food magazine. Familiarization with sucrose pellets or sucrose solution took place in the home cages, where pellets or solution were placed in a small plastic dish. The consumption test took place in the individual feeding cages described in Experiment 1, which were either fitted with plastic bottles containing sucrose solution or contained small plastic dishes in which sucrose pellets were placed.

Procedure

Handling. On Days 1-3, the rats were individually handled for about 2 minutes daily. The rats were divided into two groups — group PEL would be trained with one sucrose pellet as an instrumental outcome, while group SOL would be trained with approximately 0.27ml 20% sucrose solution as an outcome. The pure sucrose content of the pellet and the solution outcomes were approximately the same. To reduce neophobia to the sucrose solution and sucrose pellets, after each daily handling session rats were pre-exposed to their respective assigned outcome in their home cages for ten minutes. A small plastic dish containing either 20% sucrose solution or approximately 30 sucrose pellets was placed in the home cage. The dish was removed from the cage after each rat was observed to consume the pellets or solution.

Magazine training. On Days 4-5, the rats were trained to consume their respective outcomes from the food magazine in the operant chamber, as described for Experiment 1.

Leverpress training. On Days 6-18, the rats were trained to leverpress in order to obtain their respective outcome, in a free operant procedure, as described for Experiment 1. One rat in the SOL group did not acquire magazine entry behavior and was dropped from the study. Another two rats in the SOL group required extra days for magazine training and/or lever press training, and thus received only six and nine sessions of training on the 30sec interval schedule, respectively, instead of twelve. Results omitting these rats were no different from those reported below, and so they were not omitted from the study.

Context extinction. On Days 19-21, rats underwent three sessions of context extinction (one session a day), as described for Experiment 1.

Motivational shift. Immediately after the last context extinction session (i.e., on Day 21), water was removed from the home-cages of half of the rats in each of the two groups. This resulted in four groups: groups SOL/THIRSTY (N=5) and PEL/THIRSTY (N=5) — to be tested when water deprived, and groups SOL/SATED (N=4) and PEL/SATED (N=5) — to be tested undeprived.

Extinction and Consumption tests. On Day 22, after at least twenty-two hours of water deprivation, the rats were tested for leverpressing behavior in extinction as described for Experiment 1. Immediately after the extinction test, a consumption test ensued. Rats in groups SOL/THIRSTY and SOL/SATED were placed in individual cages and allowed to drink sucrose solution freely from a plastic bottle, for an hour. The

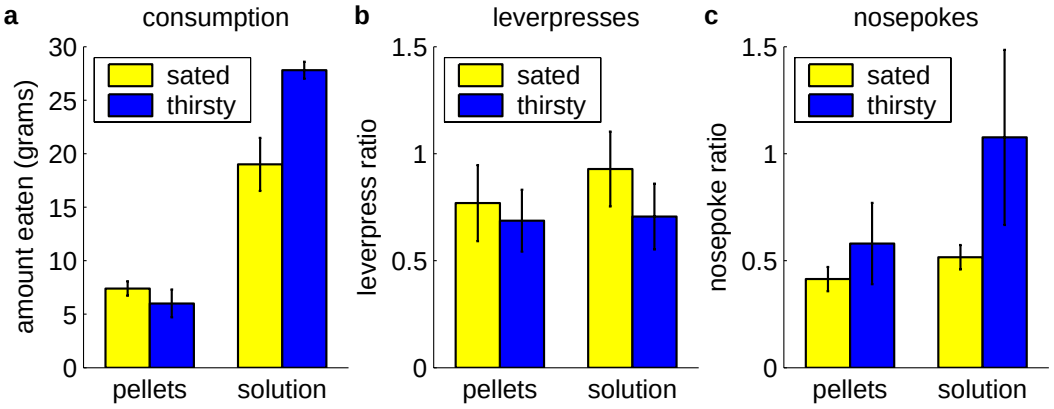

Figure 6.2: Experiment 2 results. **a.** Mean weight (in grams) consumed by each of the four groups in the consumption test. **b.** Mean rate of leverpresses per minute in each group in the extinction test, expressed as a percentage of the leverpress rate in the last training session. **c.** Mean rate of nosepokes per minute in each group in the extinction test, expressed as a percentage of the nosepoke rate in the last training session.

amount of sucrose consumed was computed by weighing the bottles before and after the test. Rats in groups PEL/THIRSTY and PEL/SATED were placed in individual cages and allowed to consume sucrose pellets freely from a small plastic dish in the cage, for an hour. The amount consumed was computed by weighing the cage (without the rat) before and after the test.

6.3.2 Results

Analysis of the results was performed using two-way analysis of variance (ANOVA), with main factors of motivation and outcome identity and an alpha level of 0.05. In the last session of instrumental training, both PEL groups showed similar response rates (mean (SD) rates of lever pressing in groups PEL/SATED and PEL/THIRSTY were 7.6(3.8) and 7.6(3.2) presses per minute, respectively; mean (SD) rates of magazine entry were 6.7(4.3) and 8.1(4.1) entries per minute, respectively), and both SOL groups showed similar response rates (mean (SD) rates of lever pressing in groups SOL/SATED and SOL/THIRSTY were 6.8(3.2) and 6.8(3.0) presses per minute, respectively; mean (SD) rates of magazine entry were 4.8(1.6) and 5.7(2.9) entries per minute, respectively). Statistical analysis showed no significant main effects or interactions (all $Fs < 1.7$, $Ps > 0.2$). However, as the groups trained with sucrose solution as an outcome showed a tendency to respond less than those trained with sucrose pellets, and to reduce between-subject variability, response rates in the extinction test were expressed as a percentage of response rates in the last training session.

The results from the consumption test confirmed that the motivational upshift had indeed produced the desired effect on the values of the different outcomes. Figure 6.2a shows the amount of outcome (sucrose pellets or sucrose solution) consumed in the consumption test, by all four groups. Clearly, more sucrose solution was consumed (as measured by weight) than were sucrose pellets. However, the result of interest is that the PEL/THIRSTY groups consumed the same amount of pellets as the PEL/SATED group, whereas the

SOL/THIRSTY group consumed *more* solution than the SOL/SATED group. This suggests that the sucrose solution was more desirable for water-deprived rats as compared to non-deprived rats, while sucrose pellets had similar value for both thirsty and sated rats. These observations were confirmed by statistical analysis, which showed significant main effects of outcome ($F(1,15) = 151.4$, $P < 0.0001$) and motivation ($F(1,15) = 7.4$, $P < 0.5$), and a significant outcome x motivation interaction ($F(1,15) = 14.1$, $P < 0.005$).

Despite of the higher motivational relevance of the sucrose solution to the water-deprived rats, the leverpress results from the extinction test (Figure 6.2b) show that the SOL/THIRSTY group *did not* press the lever any more than the other groups. If anything, the water-deprived rats trained with sucrose solution tended to press the lever *less* than the non-deprived rats. This result, however, did not reach statistical significance (all $Fs < 1$ in a two-way ANOVA).

As can be seen in Figure 6.2c, results for the nosepoke behavior were more reflective of the value of the different outcomes to the different groups, as there was a tendency of the SOL/THIRSTY rats to perform more food magazine entries than the other groups. However, this failed to reach significance (all $Ps > 0.15$ in a two-way ANOVA). Nonparametric tests used as a result of the difference of variances between shifted and non-shifted groups yielded similarly non-significant results. That the outcome-specific effect on magazine entry behavior did not reach significance may be due to the low number of animals in each group. It can also be a consequence of the context extinction procedure — Coutureau and Killcross (2003), who applied six sessions of context extinction, also failed to show the previously well-documented effect of outcome devaluation by specific-satiety prefeeding on magazine behavior (even after extensive training, eg., Killcross & Coutureau, 2003), a negative finding that they attributed to the addition of the context extinction stage.

6.4 Discussion

Balleine (1992) investigated the effects of motivational up-shifts and down-shifts on moderately trained goal-directed leverpressing for food, and found that post-training motivational shifts modulate goal-directed behavior by means of determining the value of the goal of the behavior. Here, we conducted a similar investigation into the control of extensively trained leverpressing by motivational states. Our aim was to elucidate the motivational control of habitual behavior, and test our predictions from Chapter 3.

In Experiment 1, we trained rats extensively to leverpress for a sucrose solution while hungry, and then employed a side-shift from food deprivation to water deprivation, to assess the effects of this shift on leverpressing in extinction, in comparison to that of non-shifted rats. Although a consumption test showed that the outcome was similarly desirable in the thirsty as in the hungry state, rats shifted to thirst pressed the lever significantly less than unshifted controls. This was not the result of the shifted group working for a reward with an unknown utility, as pre-exposure to the sucrose solution in the thirsty state, which was intended to allow learning of the high value of the solution in this state (incentive learning), did not alleviate this, and leverpressing in the pre-exposed group was no different than that of a non-preexposed group. The rate of

magazine behavior, however, was not significantly lower in the shifted groups compared to the unshifted group.

Experiment 2 employed an upshift from a non-deprived to a water-deprived state, and compared leverpressing in extinction of rats trained extensively with either sucrose pellets or sucrose solution as an outcome. A consumption test showed the solution outcome to be more valued in the deprived compared to the non-deprived state, while the pellets were similarly desirable in both states. Regardless of the different outcome values, lever pressing in both shifted groups was not significantly different from that of unshifted controls. In contrast, the rate of nosepoke behavior did show a tendency to follow the new outcome values, although this result did not reach statistical significance.

Taken together, these results suggest several conclusions regarding the effects of motivation on habitual behavior. First and foremost, outcome-specific motivational effects were not seen in either experiment. Thus our results show that, in accord with the view that habitual behavior is controlled by a value caching mechanism (Chapter 1; Daw et al., 2005), the motivational control over extensively trained leverpressing is *not* mediated by alterations of the value of the outcome. This is perhaps not surprising given that we chose to investigate behavior that is specifically known to be independent of the value of its outcome (Dickinson, 1985).

Second, the results of both experiments are consistent with a combined effect of 'generalized drive' and generalization decrement on habitual leverpressing. In Experiment 1, drive was probably more or less equated, so an effect of generalization decrement as a result of the motivational shift could be clearly observed. That such an effect is not seen in the shifted groups in Experiment 2 can be explained by assuming the *summation* of two opposing effects, a 'generalised drive' effect elevating responding in the water deprived groups, and a generalization decrement effect reducing responding in these same rats.

One might argue that the lack of outcome-dependent effects of motivation in Experiment 2 is due to the absence of an incentive learning stage, as this has been shown to have a crucial role in moderately trained behavior. However, two lines of evidence lead us to believe that this is not the case. First, previous outcome devaluation studies have shown that incentive learning does not affect extensively trained behavior (Adams, 1980, 1982; Dickinson et al., 1995). Furthermore, in the absence of incentive learning, moderately trained behavior does not show either the generalization decrement or the generalized drive effects seen here (compare, for instance, response rates in control groups versus shifted groups with no incentive learning in: Balleine, 1992; Balleine et al., 1995; Dickinson et al., 1995; Lopez et al., 1992).

We used a training protocol specifically aimed at producing habitual behavior, and indeed showed that, contrary to the outcome-mediated motivational control of goal-directed behavior, the motivational control of habitual behavior is outcome independent. Furthermore, the prominent role for generalization decrement is in accord with an S-R value-caching view of habitual control, as this view stresses the importance of the state (or stimulus) S in the presence of which the behavior was learned, in determining the value of an action, and eliciting conditioned habitual behavior. Changes in S (which, in free-operant tasks, is usually identified with contextual stimuli) will have a profound effect on the response R, if this is indeed the controlling structure. It

would not be surprising to find a lesser role for generalization decrement in goal-directed behavior, as there the control structure is based on the relationship between responses and their specific predicted outcomes, rather than the contextual stimuli. It is interesting that in both cases a prominent route to motivational control of behavior is through the traditional constituents of the association thought to control behavior: in the so called R-O controlled goal-directed behavior, motivation asserts control through the "O", while in S-R controlled habitual behavior, control is asserted through the "S".

Interestingly, the pattern of magazine responding and its susceptibility to motivational influences was at least qualitatively different from that of leverpressing. In contrast to the outcome independence of leverpressing, in both experiments the rate of magazine behavior tended to follow the value of the outcome in the different motivational states, as assessed in the consumption tests. This is in agreement with several recent studies which show direct, outcome-specific effects of devaluation or motivational shifts on magazine behavior, regardless of the amount of training (eg, Balleine & Dickinson, 1991; Balleine, 1992; Killcross & Coutureau, 2003; see also Daw et al., 2005).

The apparent outcome-specific effects of motivation on nosepoking behavior, even in face of extensive training that normally renders instrumental behavior insensitive to outcome value, have been taken to suggest that magazine behavior is not predominantly an instrumentally controlled behavior, but rather a Pavlovian approach-type response (Balleine & Dickinson, 1991; Balleine, 1992, 2000; Dickinson & Balleine, 2002; Dayan & Balleine, 2002). This has also been one of the major lines of evidence for the direct motivational control of Pavlovian responses. Two studies contest this interpretation. Balleine and colleagues (Balleine et al., 1995; Corbit & Balleine, 2003) trained rats to perform a chain of two purely instrumental responses (leverpressing followed by chain pulling) in order to obtain food that was made available in a flap-less food magazine, such that a further magazine-flap response was not required. They found that after moderate training (or training on a ratio schedule of reinforcement known to produce behavior relatively resistant to habitization), the response more proximal to the reward was directly affected by a motivational shift, whereas the effect of the motivational shift on the distal response necessitated incentive learning. Moreover, when trained to leverpress for food in the same flap-less operant chamber, leverpressing itself showed such immediate motivational effects (although incentive learning further augmented these; Balleine et al., 1995, and see also Rescorla, 1994).

These results suggest that a major factor in determining susceptibility to direct outcome-dependent motivational influence, even in the light of extensive training and in the absence of an opportunity for incentive learning, may be the *proximity* of the response in question to the outcome, rather than its Pavlovian or instrumental nature. According to this view, instrumental actions that are proximal to the outcome may be directly sensitive to the value of the outcome, while distal actions necessitate incentive learning in order to show the effects of motivation early in training, and are rendered insensitive to outcome value with extensive training (Daw et al., 2005). In this sense, our magazine entry results suggest that direct motivational control of proximal actions can also be seen with motivational up-shifts and side-shifts. Interestingly, there may be evidence for an analogous proximal-distal distinction in the effects of outcome devaluation on purely

Pavlovian responses. Measuring several different conditioned responses to a cue predicting food, Holland and Straub (1979) showed that post-conditioning outcome devaluation resulted in a significant reduction of responses more proximal to the outcome, while distal responses were not affected.

To summarize, although motivation seems to control goal-directed behavior by determining the incentive value of its consequent outcomes, we have shown here that motivation does not impose any form of outcome-specific control on habitual behavior. However, our results show that motivational states can indeed control the vigor of habitual behavior through 'generalized drive' or 'energizing' effects, and, furthermore, that shifts in motivational states can influence habitual behavior as a result of generalization decrement, supporting the hypotheses borne from our theoretical analysis of habitual behavior.

Chapter 7

Contributions and future directions

This thesis makes four main contributions. The first is a normative framework for free operant behavior. The model presented in Chapter 2 provides the first principled explanation for the relationship between response rates in self-paced scenarios and parameters such as reward magnitude, reward frequency and contingency. Key to this is that the availability of rewards, the rate of rewards and the costs of responding can all depend on the speed of responding. Further, Chapter 5 suggests how such normative behavior might emerge through resource-constrained, on-line learning.

The other three contributions of the thesis arise (i) through drawing out the direct implications of this framework for factors such as motivation that are manipulated in psychological experiments; (ii) through using the framework to clarify the complexities associated with the control of one particular type of instrumental behavior, namely habitual responding; and (iii) through identifying a key signal in the model with tonic levels of dopamine and thereby accounting for a huge wealth of data on effects associated with this neuromodulator.

The second contribution of this thesis is a normative account of the effects of motivation on response rates. In Chapter 3 we start from the prosaic premise that motivational states influence the utilities to a subject of different outcomes. Incorporating this into the framework of Chapter 2 shows that the normative (ie, optimal) effects of changes in motivational states should be twofold. The first of these is an outcome-specific effect by which the propensity to perform actions depends on their motivationally-determined utility (ie, there is a higher propensity to perform those actions that will yield outcomes with higher utility), as in the traditional 'directing' effect of motivation. The second effect, which corresponds to the classic but controversial 'energizing' effect of motivation, is an outcome-general effect in which, through modulation of the expected net rate of reward, motivation influences the vigor of all pre-potent actions.

A key insight is that how these two effects are manifest in action selection depends on the computational characteristics of the decision-making system controlling behavior. The two main psychological classes of behavioral control are goal-directed control, which has been modeled using a forward model that computes

the predicted outcomes of actions, and habitual control, which traditionally involves cached state or state-action values that summarize information about previously experienced outcomes. We claim that since decision-making in the goal-directed system is based on online forward prediction of the consequences of different actions, this system can be immediately sensitive to the 'directing' aspects of motivation. In contrast, because action selection in the habitual system is based on previously learned cached values, it can only be sensitive to the 'energizing' effects of motivation. Thus the third contribution of this thesis is a theory which clarifies the motivational control of habitual responding and unites previously contradicting suggestions: habits are indeed insensitive to some aspects of motivation, but they are *not* wholly motivation-insensitive. Initial results from two experiments detailed in Chapter 6 provide empirical support for this theoretical result.

Finally, the fourth contribution is the extension of theories regarding the involvement of dopamine in instrumental and Pavlovian learning and action selection, such that they explain not only the role of phasic signals but also that of tonic levels of dopamine. In Chapter 4 we hypothesized that tonic levels of dopamine encode the expected net reward rate. This suggestion not only provides a first link between this important mode of dopaminergic transmission and reinforcement learning, but, importantly, it provides the first normative explanation of why dopamine has such profound effects on response vigor. That higher levels of dopamine enhance response rates while low levels of dopamine cause lethargy, can be seen as a direct consequence of the online effect no responding of the net reward rate (that is, the opportunity cost of time) associated with different tonic levels of dopamine.

Together, these results provide a framework for self-paced instrumental experiments within which animal behavior is demonstrably near-optimal. That response rates are higher when more reward is available is perhaps not surprising. But our analysis also shows that the fact that a rat works faster when food is delivered on a ratio schedule, compared to an interval schedule with a similar reward frequency, is not a haphazard quirk but rather an optimal adjustment of the cost-benefit tradeoff to the schedule contingency. Similarly, that hungry rats might not only behave faster when procuring food, but also when working for water, is optimal as it takes into account the opportunity cost of time that is shared between all actions, regardless of their specific consequences.

However, deviations from optimality still exist, and are particularly interesting in that they can shed light on constraints that our model has not taken into account. A prime example of such a deviation is the phenomenon of Pavlovian-instrumental transfer, in which the rate of instrumental responding is influenced by a cue predicting the availability of rewards that are *not* contingent on any actions. Optimally, such 'free' rewards should not affect the cost/benefit tradeoff determining optimal instrumental responding. However, this relies on correct classification of different rewards into those earned instrumentally and those independent of one's actions. This distinction can be difficult to make, especially as instrumental outcomes may be delayed or contingent on a long sequence of actions. Errors or heuristics used in this process may lie at the heart of phenomena like Pavlovian-instrumental transfer.

Suboptimality is also apparent in behavior on concurrent free-operant schedules. For example, in concurrent

variable-interval/variable-ratio schedules, animals show matching behavior, allocating more responding to the interval schedule than is optimal. In concurrent variable interval schedules, rats and pigeons display a constant probability of switching between the two arms, although optimally the switching probability should depend on the time since the previous switch. These suboptimalities may be due to limitations on the ability of animals to learn and to represent adequately the correct Markovian state-space for a given schedule of reinforcement. However, inadequate state representations are not necessarily indicative of suboptimal control, but rather they may arise as a result of lack of information, due to the influence of prior beliefs, or perhaps because simpler state representations allow more efficient learning, which, in the context of an experimental session of limited length, counteracts the loss of reward as a result of using the wrong state space. In any case, as we detail below, the issue of state-space learning poses a long list of burning questions for future research.

Indeed, directions for future work abound: extending the model such that state transitions can occur during the latency of responses (which would make our model applicable to scenarios such as avoidance learning and Pavlovian conditioning), dealing with partial observability and timing noise, empirically testing further the effects of generalized drive and dopaminergic manipulations, and clarifying the role of tonic dopamine in aversive conditioning and its relationship to opponency with serotonin. These and others have been detailed in the relevant chapters (see sections 2.4.3, 4.4.1-4.4.3 and 5.4.4), and so we will not repeat them here.

However, several important questions arise from this thesis as a whole. First, much empirical and theoretical work has been devoted the inter-temporal choice, discounting and the relationship between these and rational behavior on the one hand, and pathologies such as drug addiction on the other hand. Our framework can be seen as a continuous extension of inter-temporal choice: in our model every choice of latency is, in fact, a choice between an earlier but more costly reward, and a later, larger one (due to lower vigor costs). Of course, the opportunity cost of lost time, and the fact that latency can also affect the probability of reward, make this choice more complex than the common 'low reward now/high reward later' scenario used in intertermporal choice experiments. However, the application of our framework to discrete trial inter-temporal experiments in which the vigor costs of responding are equated, can isolate and flesh out the contribution of the the opportunity cost of time to such action selection, and perhaps shed light on the controversies in that field.

A second important avenue for future research regards the multiplicity of timescales that are relevant for the net rate of reward. In average reward reinforcement learning, the reward rate should pertain to the whole Markov decision process. However, the world is not one Markov decision process, but rather can be divided into separate tasks (some of them Markovian and most of them not), which are not necessarily executed in sequence, but rather can be simultaneous or nested one within another. Furthermore, tasks are frequently non-stationary with respect to time. Given all these complexities, what should the net reward rate correspond to? If different reward rates (and policies) are learned for different tasks, how are task boundaries clearly defined? And is there a hierarchical structure in which these reside that could assist in decision making? By suggesting, in Chapter 3, that different net reward rates should be associated with different motivational states, we implicitly made an assumption of task separation and of the maintenance of a multiplicity of net

reward rates. Furthermore, in Chapter 4 we suggested that such a mapping between task and net reward rate can affect changes in tonic levels of dopamine when tasks change. These implicit assumptions and vague suggestions should be clarified both experimentally and theoretically, if we are to understand how the brain deals with multiplicity of net rates of reward.

Last, in our simulations we have used different state spaces for ratio and interval schedules. This is because ratio and interval schedules pose different Markov decision problems, with different solutions, but also different state spaces. Behavioral results suggest that animals indeed represent or learn these schedules differently, because they show different rates of responding in the two schedules even when reinforcement rates are matches (yoked). An intriguing area for future research is how the relevant state space is learned (online, from stochastic experience with the environment), and how this learning process interacts with the reinforcement learning process in which values are attributed to states, and with innate beliefs (priors). This fundamental learning process, which has received little (if any) attention in the literature, is brought to the forefront by our model. Treating state-space representation as a learning problem also makes partially observability (as a result of a state-space that is ill matched to the actual states underlying the task) inherent in the decision process. In fact, partially observable state spaces are inherent in many real-world tasks even without this added complexity, for instance, as a result of timing noise. In this thesis, I have eschewed issues of partial observability on optimal prediction and action selection. The picture of free operant decision making, and how this is affected by motivation, will not be complete until these issues are taken into consideration.

References

Aberman, J. E., & Salamone, J. D. (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience*, *92*(2), 545-552.

Abounadi, J., Bertsekas, D., & Borkar, V. S. (2000). Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, *40*(3), 681–698.

Adams, C. D. (1980). Post conditioning devaluation of an instrumental reinforcer has no effects on extinction performance. *Quarterly Journal of Experimental Psychology*, *32*, 447-458.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *34B*, 77-98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *33B*, 109-121.

Ainslie, G. (1975). Specious reward: a behavioural theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*, 463-496.

Amalric, M., & Koob, G. F. (1987). Depletion of dopamine in the caudate nucleus but not in nucleus accumbens impairs reaction-time performance in rats. *Journal of Neuroscience*, *7*(7), 2129–2134.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th international conference on machine learning.* San Francisco, CA: Morgan Kaufman.

Baird, L., & Moore, A. W. (1999). Gradient descent for general reinforcement learning. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems.* MIT Press.

Baird, L. C. (1993). *Advantage updating* (Tech. Rep. No. WL-TR-93-1146). Dayton, OH: Wright-Patterson Air Force Base.

Baird, L. C., & Klopf, A. H. (1993). *Reinforcement learning with high-dimensional, continuous actions* (Tech. Rep.). Wright Laboratory, Wright-Patterson Air-Force Base.

Balleine, B. W. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*(3), 236-250.

Balleine, B. W. (1994). Asymmetrical interactions between thirst and hunger in Pavlovian-instrumental transfer. *The Quarterly Journal of Experimental Psychology B*, *47*(2), 211–231.

Balleine, B. W. (2000). Incentive processes in instrumental conditioning. In R. Mowrer & S. Klein (Eds.), *Handbook of contemporary learning theories* (p. 307-366). Mahwah, NJ: Lawrence Erlbaum Associates.

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol Behav*, *86*(5), 717–730.

Balleine, B. W., & Dickinson, A. (1988). The role of incentive learning in instrumental outcome revaluation by sensory specific satiety. *Animal Learning and Behavior*, *26*, 46-59.

Balleine, B. W., & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology*, *43B*(3), 279-296.

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

Balleine, B. W., & Dickinson, A. (2000). The effect of lesions of the insular cortex on instrumental conditioning: Evidence for a role in incentive memory. *Journal of Neuroscience*, *20*(23), 8954-8964.

Balleine, B. W., Garner, C., Gonzalez, F., & Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(3), 203-217.

Balleine, B. W., Killcross, A. W., & Dickinson, A. (2003). The effects of lesions of the basolateral amygdala on instrumental conditioning. *Journal of Neuroscience*, *23*(2), 666-675.

Balleine, B. W., & Killcross, S. (2006). Parallel incentive processing: an integrated view of amygdala function. *Trends in Neuroscience*, *29*(5), 272–279.

Barrett, J. E., & Stanley, J. A. (1980). Effects of ethanol on multiple fixed-interval fixed-ratio schedule performances: Dynamic interactions at different fixed-ratio values. *Journal of the Experimental Analysis of Behavior*, *34*(2), 185-198.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232). Cambridge: MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 834-846.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1989). Sequential decision problems and neural networks. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (p. 686-693). Cambridge, MA: MIT Press.

Baum, E. B., & Smith, W. D. (1997). A Bayesian approach to relevance in game playing. *Artificial Intelligence*, *97*, 195–242.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242.

Baum, W. M. (1982). Choice, changeover, and travel. *Journal of the Experimental Analysis of Behavior*, *38*, 35–49.

Baum, W. M. (1989). Quantitative prediction and molar description of the environment. *The Behavior Analyst*, *12*, 167–176.

Baum, W. M. (1993). Performances on ratio and interval schedules of reinforcement: Data and theory. *Journal of the Experimental Analysis of Behavior*, *59*, 245–264.

Bautista, L. M., Tinbergen, J., & Kacelnik, A. (2001). To walk or to fly? How birds choose among foraging modes. *Proceedings of the National Academy of Sciences*, *98*(3), 1089-1094.

Baxter, J., & Bartlett, P. L. (1999). *Direct gradient-based reinforcement learning: I. Gradient estimation algorithms* (Tech. Rep.). Research School of Information Sciences and Engineering, Australian National University.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Beninger, R. J. (1983). The role of dopamine in locomotor activity and learning. *Brain Res Rev*, *6*, 173-196.

Bergstrom, B. P., & Garris, P. A. (2003). 'Passive stabilization' of striatal extracellular dopamine across the lesion spectrum encompassing the presymptomatic phase of Parkinson's disease: a voltammetric study in the 6-OHDA lesioned rat. *J Neurochem*, *87*(5), 1224-1236.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiol Behav*, *81*, 179–209.

Berridge, K. C. (2007). The debate over dopamine's role in reward:The case for incentive salience. *Psychopharmacology (Berl)*, *191*(3), 391–431.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Review*, *28*, 309-369.

Bertsekas, D. P. (1995a). *Dynamic programming and optimal control (Vol II).* Athena Scientific Belmont, Mass.

Bertsekas, D. P. (1995b). *Dynamic programming and optimal control (Vol I).* Athena Scientific Belmont, Mass.

Bertsekas, D. P. (1998). A new value iteration method for the average cost dynamic programming problem. *SIAM Journal of Control and Optimization*, *36*, 742–759.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming.* Athena Sc.

Bindra, D. (1974). A motivational view of learning, performance and behavior modification. *Psychological Review*, *81*(3), 199-213.

Bitterman, M. E. (1965). Phyletelic differences in learning. *American Psychologist*, *20*, 396–410.

Blundell, P., Hall, G., & Killcross, S. (2001). Lesions of the basolateral amygdala disrupt selective aspects of reinforcer representation in rats. *Journal of Neuroscience*, *21*(22), 9018–9026.

Blundell, P., Hall, G., & Killcross, S. (2003). Preserved sensitivity to outcome value after lesions of the basolateral amygdala. *Journal of Neuroscience*, *23*(20), 7702–7709.

Boelens, H., & Kop, P. F. M. (1983). Concurrent schedules: Spatial separation of response alternatives. *Journal of the Experimental Analysis of Behavior*, *40*(1), 35–45.

Bolles, R. C. (1967). *Theory of motivation.* Harper & Row.

Bradshaw, C. M., Ruddle, H. V., & Szabadi, E. (1981). Relationship between response rate and reinforcement frequency in variable interval schedules: II. Effect of the volume of sucrose reinforcement. *Journal of the Experimental Analysis of Behavior*, *35*, 263–270.

Bradshaw, C. M., Szabadi, E., & Bevan, P. (1978). Relationship between response rate and reinforcement frequency in variable-interval schedules: The effect of concentration of sucrose reinforcement. *Journal of the Experimental Analysis of Behavior*, *29*, 447–452.

Bradtke, S. J., & Duff, M. O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 393–400). MIT Press.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 681–684.

Brown, J. S. (1961). *The motivation of behavior.* New York: McGraw-Hill.

Brown, P. L., & Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior*, *11*(1), 1–8.

Brown, V. J., & Bowman, E. M. (1995). Discriminative cues indicating reward magnitude continue to determine reaction time of rats following lesions of the nucleus accumbens. *European Journal of Neuroscience*, *7*(12), 2479–2485.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci Biobehav Rev*, *26*(3), 321–352.

Carr, G. D., & White, N. M. (1987). Effects of systemic and intracranial amphetamine injections on behavior in the open field: A detailed analysis. *Pharmacol Biochem Behav*, *27*, 113-122.

Cassandra, A. R., Kaelbling, L. P., & Littman, M. (1994). Acting optimally in partially observable stochastic domains. In *Proceedings of the twelfth national conference on artificial intelligence (aaai-94)* (Vol. 2, pp. 1023–1028). Seattle, Washington: AAAI Press/MIT Press.

Catania, A. C., Matthews, T. J., Silverman, P. J., & Yohalem, R. (1977). Yoked variable-ratio and variable-interval responding in pigeons. *Journal of the Experimental Analysis of Behavior*, *28*, 155–161.

Chéramy, A., Barbeito, L., Godeheu, G., Desce, J., Pittaluga, A., Galli, T., Artaud, F., & Glowinski, J. (1990). Respective contributions of neuronal activity and presynaptic mechanisms in the control of the in vivo release of dopamine. *J Neural Transm Suppl*, *29*(183-193).

Chesselet, M. F. (1990). Presynaptic regulation of dopamine release: Implications for the functional organization of the basal ganglia. *Annals of the New York Academy of Science*, *604*, 17-22.

Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *Journal of Neuroscience*, *6*(3), 613–619.

Cleveland, J. M. (1999). Inter-response time sensitivity during discrete-trial and free operant concurrent variable-interval schedules. *Journal of the Experimental Analysis of Behavior*, *72*(3), 317–339.

Cohen, J. D., Braver, T. S., & Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Curr Opin Neurobiol*, *12*(2), 223–229.

Cole, M. (1999). Molar and molecular control in variable-interval and variable-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *71*, 319–328.

Colwill, R. M., & Motzkin, D. K. (1994). Encoding of the unconditioned stimulus in Pavlovian conditioning. *Animal Learning & Behavior*, *22*, 384–394.

Colwill, R. M., & Rescorla, R. A. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(4), 520-536.

Colwill, R. M., & Rescorla, R. A. (1986). Associative structures in instrumental learning. *The Psychology of Learning and Motivation*, *20*, 55-104.

Colwill, R. M., & Rescorla, R. A. (1988). The role of response-reinfrcement associations increases throughout extended instrumental training. *Animal Learning & Behavior*, *16*(1), 105-111.

Colwill, R. M., & Triola, S. M. (2002). Instrumental responding remains under the control of the consequent outcome after extended training. *Behavioural Processes*, *57*, 51-64.

Corbit, L. H., & Balleine, B. W. (2000). The role of the hippocampus in instrumental conditioning. *Journal of Neuroscience*, *20*(11), 4233–4239.

Corbit, L. H., & Balleine, B. W. (2003). Instrumental and Pavlovian incentive processes have dissociable effects on components of a heterogeneous instrumental chain. *Journal of Experimental Psychology: Animal Behavior Processes*, *29*(2), 99-106.

Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. *Journal of Neuroscience*, *25*(4), 962–970.

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *Journal of Neuroscience*, *21*(9), 3251–3260.

Corbit, L. H., Ostlund, S. B., & Balleine, B. W. (2002). Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *Journal of Neuroscience*, *22*(24), 10976–10984.

Correa, M., Carlson, B. B., Wisniecki, A., & Salamone, J. D. (2002). Nucleus accumbens dopamine and work requirements on interval schedules. *Behavioral and Brain Reearchs*, *137*, 179-187.

Cousins, M. S., Atherton, A., Turner, L., & Salamone, J. D. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. *Behavioral and Brain Reearchs*, *74*, 189-197.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, *146*, 167-174.

Dallery, J., & Soto, P. L. (2004). Herrnstein's hyperbolic matching equation and behavioral pharmacology: Review and critique. *Behavioral Pharmacology*, *15*(7), 443–459.

Das, T. K., Gosavi, A., Mahadevan, S., & Marchalleck, N. (1999). Solving semi-Markov decision problems using average reward reinforcement learning. *Managment Science*, *45*(4), 560–574.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603-616.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Daw, N. D., Niv, Y., & Dayan, P. (2006). Actions, policies, values, and the basal ganglia. In E. Bezard (Ed.), *Recent breakthroughs in basal ganglia research* (p. 111-130). Nova Science Publishers Inc.

Daw, N. D., & Touretzky, D. S. (2000). Behavioral results suggest an average reward TD model of dopamine function. *Neurocomputing*, *32*, 679-684.

Daw, N. D., & Touretzky, D. S. (2001). Operant behavior suggests attentional gating of dopamine system inputs. *Neurocomputing*, *38-40*, 1161-1167.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*(11), 2567–2583.

Dawson, G. R., & Dickinson, A. (1990). Performance on ratio and interval schedules with matched reinforcement rates. *The Quarterly Journal of Experimental Psychology B*, *42*, 225-239.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems.* MIT Press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*(2), 285–298.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153–1160.

de Borchgrave, R., Rawlins, J. N. P., Dickinson, A., & Balleine, B. W. (2002). Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. *Experimental Brain Research*, *144*(1), 50–68.

Denk, F., Walton, M. E., Jennings, K. A., Sharp, T., Rushworth, M. F., & Bannerman, D. M. (2005). Differential involvement of serotonin and dopamine systems in cost-benefit decisions about delay or effort. *Psychopharmacology*, *179*(3), 587-596.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *308*(1135), 67-78.

Dickinson, A. (1994). Instrumental conditioning. In N. Mackintosh (Ed.), *Animal learning and cognition* (p. 45-79). San Diego, California: Academic Press.

Dickinson, A. (1997). Bolles's psychological sylllogism. In M. E. Bouton & M. S. Fanselow (Eds.), *Learning, motivation, and cognition: The functional behaviorism of Robert C. Bolles* (pp. 345–367). Washington DC: American Psychological Association.

Dickinson, A., & Balleine, B. W. (1990). Motivational control of instrumental performance following a shift from thirst to hunger. *The Quarterly Journal of Experimental Psychology*, *24B*(4), 413-431.

Dickinson, A., & Balleine, B. W. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*, 1-18.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Learning, motivation and emotion* (Vol. 3, p. 497-533). New York: John Wiley & Sons.

Dickinson, A., Balleine, B. W., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning and Behavior*, *23*(2), 197-206.

Dickinson, A., & Dawson, G. R. (1987). Pavlovian processes in the motivational control of instrumental performance. *The Quarterly Journal of Experimental Psychology*, *39B*, 201-213.

Dickinson, A., & Dawson, G. R. (1988). Motivational control of instrumental performance: The role of prior experience with the reinforcer. *The Quarterly Journal of Experimental Psychology*, *40B*(2), 113-134.

Dickinson, A., & Dawson, G. R. (1989). Incentive learning and the motivational control of instrumental performance. *The Quarterly Journal of Experimental Psychology*, *41B*(1), 99-112.

Dickinson, A., & Nicholas, D. J. (1983a). Irrelevant incentive learning during instrumental conditioning: the role of the drive-reinforcer and response-reinforcer relationships. *Quarterly Journal of Experimental Psychology*, *35B*, 249-263.

Dickinson, A., & Nicholas, D. J. (1983b). Irrelevant incentive learning during training on ratio and interval schedules. *Quarterly Journal of Experimental Psychology*, *35B*, 235–247.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of instrumental training contingency on susceptibility to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *35B*, 35-51.

Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine agonists. *Behavioral Neuroscience*, *114*(3), 468-483.

Dickinson, A., Squire, S., Varga, Z., & Smith, J. (1998). Omission learning after instrumental pretraining. *The Quarterly Journal of Experimental Psychology*, *51B*(3), 271-286.

Domjan, M. (2003). *The principles of learning and behavior* (5th ed.). Belmont, California: Thomson/Wadsworth.

Dragoi, V., & Staddon, J. E. R. (1999). The dynamics of operant conditioning. *Psychological Review*, *106*(1), 20-61.

Egelman, D. M., Person, C., & Montague, P. R. (1998). A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, *10*(5), 623–630.

Estes, W. K. (1948). Discriminative conditioning II: Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *Journal of Experimental Psychology*, *38*, 173-177.

Evenden, J. L., & Robbins, T. W. (1983). Increased dopamine switching, perseveration and perseverative switching following *d*-amphetamine in the rat. *Psychopharmacology*, *80*, 67-73.

Everitt, B. J., & Robbins, T. W. (1992). Amygdala-ventral striatal interactions and reward-related processes. In J. Aggleton (Ed.), *The amygdala: Neurological aspects of emotion, memory, and mental dysfunction* (p. 401-429). John Wiley & Sons.

Faure, A., Haberland, U., Condé, F., & Massioui, N. E. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *Journal of Neuroscience*, *25*, 2771-2780.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: Evidence from a reversal learning paradigm. *Brain*, *126*(8), 1830–1837.

Felton, M., & Lyon, D. O. (1966). The post-reinforcement pause. *Journal of the Experimental Analysis of Behavior*, *9*, 131–134.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement.* Appleton-Century-Crofts.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898-1902.

Fletcher, P. J., & Korth, K. M. (1999). Activation of 5-HT1B receptors in the nucleus accumbens reduces amphetamine-induced enhancement of responding for conditioned reward. *Psychopharmacology*, *142*, 165-174.

Floresco, S. B., West, A. R., Ash, B., Moore, H., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience*, *6*(9), 968-973.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683.

Foster, T. M., Blackman, K. A., & Temple, W. (1997). Open versus closed economies: Performance of domestic hens under fixed-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *67*, 67-89.

Friston, K. J., Tononi, G., Reeke, G. N. J., Sporns, O., & Edelman, G. M. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, *59*(2), 229-243.

Gallagher, M., McMahan, R. W., & Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, *19*(15), 6610–6614.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Science USA*, *101*(36), 13124–13131.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review*, *107*, 289-344.

Gallistel, C. R., Stellar, J. R., & Bubis, E. (1974). Parametric analysis of brain stimulation reward in the rat: I. The transient process and the memory-containing process". *J Comp Physiol Psychol*, *87*, 848-860.

Geisler, S., & Zahm, D. S. (2005). Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *The Journal of Comparative Neurology*, *490*(3), 270–294.

Gewirtz, J. C., & Davis, M. (2000). Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. *Learning & Memory*, *7*(5), 257–266.

Gibbon, J. (1977). Scalar Expectancy Theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279-325.

Gibbon, J. (1992). Ubiquity of scalar timing with poisson clock. *Journal of Mathematical Psychology*, *36*(2), 283-293.

Giertler, C., Bohn, I., & Hauber, W. (2003). The rat nucleus accumbens is involved in guiding of instrumental responses by stimuli predicting reward magnitude. *European Journal of Neuroscience*, *18*(7), 1993–1996.

Giertler, C., Bohn, I., & Hauber, W. (2004). Transient inactivation of the rat nucleus accumbens does not impair guidance of instrumental behaviour by stimuli predicting reward magnitude. *Behavioral Pharmacology*, *15*(1), 55–63.

Gonon, F. G. (1988). Nonlinear relationship between impulse flow and dopamine released by rat midbrain dopaminergic neurons as studied by in vivo electrochemistry. *Neuroscience*, *24*(1), 19–28.

Gosavi, A. (2004a). A reinforcement learning algorithm based on policy iteration for average reward: Empirical results with yield management and convergence analysis. *Machine Learning*, *55*(1), 5–29.

Gosavi, A. (2004b). Reinforcement learning for long-run average cost. *European Journal of Operational Research*, *155*, 654–674.

Goto, Y., & Grace, A. A. (2005). Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nature Neuroscience*, *8*, 805-812.

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience*, *41*(1), 1-24.

Hall, J., Parkinson, J. A., Connor, T. M., Dickinson, A., & Everitt, B. J. (2001). Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behavior. *European Journal of Neuroscience*, *13*, 1984-1992.

Hatfield, T., Han, J. S., Conley, M., Gallagher, M., & Holland, P. (1996). Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *Journal of Neuroscience*, *16*(16), 5256–5265.

Hauber, W., Bohn, I., & Giertler, C. (2000). NMDA, but not dopamine D_2, receptors in the rat nucleus accumbens are involved in guidance of instrumental behavior by stimuli predicting reward magnitude. *Journal of Neuroscience*, *20*(16), 6282–6288.

Heimer, L., & Wilson, R. D. (1975). The subcortical projections of the allocortex: Similarities in the neural associations of the hippocampus, the piriform cortex, and the neocortex. In M. Santini (Ed.), *Golgi centenial symposium proceedings* (p. 177-193). New York: Raven.

Henderson, R. W., & Graham, J. (1979). Avoidance of heat by rats: Effects of thermal context on the rapidity of extinction. *Learning and Motivation*, *10*, 351-363.

Hernandez, G., Hamdani, S., Rajabi, H., Conover, K., Stewart, J., Arvanitogiannis, A., & Shizgal, P. (In Press). Prolonged rewarding stimulation of the rat medial forebrain bundle: Neurochemical and behavioral consequences. *Behavioral Neuroscience*.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267-272.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243-266.

Herrnstein, R. J. (1990). Rational choice theory: Necessary but not sufficient. *American Psychologist*, *45*(3), 356–367.

Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. *The American Economic Review*, *81*(2), 360–364.

Herrnstein, R. J. (1997). *The matching law: Papers in psychology and economics.* Harvard University Press.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, *24*(1), 107–116.

Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *The Journal of Economic Perspectives*, *5*(3), 137-156.

Heyman, G. M., & Herrnstein, R. J. (1986). More on concurrent interval-ratio schedules: a replication and review. *Journal of the Experimental Analysis of Behavior*, *46*(3), 331–351.

Holland, P. C. (1984). Origins of behavior in Pavlovian conditioning. *The Psychology of Learning and Motivation*, *18*, 129-174.

Holland, P. C. (1997). Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. *Animal Learning and Behavior*, *25*(4), 373–399.

Holland, P. C. (1998). Amount of training affects associatively-activated event representations. *Neuropsychopharmacologyy*, *37*, 461–469.

Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(2), 104-117.

Holland, P. C., & Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Science*, *3*(2), 65–73.

Holland, P. C., & Gallagher, M. (2003). Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned-stimulus potentiated feeding and Pavlovian-instrumental transfer. *European Journal of Neuroscience*, *17*, 1680–1694.

Holland, P. C., & Gallagher, M. (2004). Amygdala-frontal interactions and reward expectancy. *Curr Opin Neurobiol*, *14*(2), 148–155.

Holland, P. C., & Gallagher, M. (2006). Different roles for amygdala central nucleus and substantia innominata in the surprise-induced enhancement of learning. *Journal of Neuroscience*, *26*(14), 3791–3797.

Holland, P. C., & Rescorla, R. A. (1975a). Second-order conditioning with food unconditioned stimulus. *J Comp Physiol Psychol*, *88*(1), 459–467.

Holland, P. C., & Rescorla, R. A. (1975b). The effets of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*(4), 355-363.

Holland, P. C., & Straub, J. J. (1979). Differential effects of two ways of devaluing the unconditioned stimulus after Pavlovian appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(1), 65-78.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(304-309).

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 249-270). Cambridge: MIT Press.

Houston, A. I., & McNamara, J. (1981). How to maximize reward rate on two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior*, *35*, 367–396.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* New York: Appleton-Century-Crofts.

Ikemoto, S., & Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Res Rev*, *31*, 6-41.

Jackson, D. M., Anden, N., & Dahlstrom, A. (1975). A functional effect of dopamine in the nucleus accumbens and in some other dopamine-rich parts of the rat brain. *Psychopharmacologia*, *45*, 139-149.

Jalali, A., & Ferguson, M. J. (1990). A distributed asynchronous algorithm for expected average cost dynamic programming. In *Proceedings of tthe 29th conference of decision and control.*

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*, 535-547.

Joel, D., & Weiner, I. (1994). The organization of the basal ganglia-thalamocortical cicuits: open interconnected rather than closed segregated. *Neuroscience*, *63*, 363-379.

Joel, D., & Weiner, I. (1999). Striatal contention scheduling and the split circuit scheme of basal ganglia-thalamocortical circuitry: From anatomy to behaviour. In R. Miller & J. Wickens (Eds.), *Conceptual advances in brain research: Brain dynamics and the striatal complex* (p. 209-236). Harwood Academic Publishers.

Kacelnik, A. (1997). Normative and descriptive models of decision making: time discounting and risk sensitivity. In G. R. Bock & G. Cardew (Eds.), *Characterizing human psychological adaptations: Ciba Foundation symposium 208* (p. 51-70). Chichester: Wiley.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kakade, S. (2001). Optimizing average reward using discounted rewards. In *Proceedings of the 14th annual conference on computational learning theory* (pp. 605–615).

Kawagoe, R., Takikawa, Y., & Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience*, *1*(5), 411–416.

Kearns, M., & Singh, S. (1998). Near-optimal reinforcement learning in polynomial time. In *Proceeding of the 15th international conference on machine learning* (pp. 260–268). San Francisco, CA: Morgan Kaufmann.

Keesey, R. E., & Kling, J. W. (1961). Amount of reinforcement and free-operant responding. *Journal of the Experimental Analysis of Behavior*, *4*, 125–132.

Kehoe, E. J. (1977). *Effects of serial compound stimuli on stimulus selection in classical conditioning of the rabbit nictitating membrane response.* Unpublished doctoral dissertation, University of Iowa.

Killcross, S., & Blundell, P. (2002). Associative representationsof emotionally significant outcomes. In S. Moore & M. Oaksford (Eds.), *Emotional cognition. from brain to behaviour* (Vol. 44, p. 35-73). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal conrtex of rats. *Cerebral Cortex*, *13*, 400-408.

Killeen, P., & Sitomer, M. (2003). MPR. *Behav Processes*, *62*(1-3), 49–64.

Killeen, P. R. (1995). Economics, ecologies and mechanics: The dynamics of responding under conditions of varying motivation. *Journal of the Experimental Analysis of Behavior*, *64*, 405-431.

Killeen, P. R. (1998). The first principle of reinforcement. In C. D. L. Wynne & J. E. R. Staddon (Eds.), *Models of action: Mechanisms for adaptive behavior* (pp. 127–156). London: Lawrence Erlbaum Associates.

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274–295.

Killeen, P. R., Hanson, S. J., & Osborne, S. R. (1978). Arousal: Its genesis and manifestation as response rate. *Psychological Review*, *85*(6), 571–581.

Kim, J., & Ragozzino, M. E. (2005). The involvement of the orbitofrontal cortex in learning under changing task contingencies. *Neurobiol Learn Mem*, *83*(2), 125–133.

Kimura, H., & Kobayashi, S. (1998a). Reinforcement learning for continuous action using stochastic gradient ascent. In *Proceedings of the 5th international conference on intelligent auntonomous systems* (pp. 288–295).

Kimura, H., & Kobayashi, S. (1998b). An analysis of Actor/Critic algorithms using eligibility traces: Reinforcement learning with imperfect value functions. In *roceedings of the 15th international conference on machine learning* (pp. 278–286).

Kita, H., & Kitai, S. T. (1990). Amygdaloid projections to the frontal cortex and the striatum in the rat. *J Comp Neurol*, *298*(1), 40–49.

Kobayashi, Y., & Okada, K.-I. (2007). Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. *Annals of the New York Academy of Science*.

Konda, V. R., & Tsitsiklis, J. N. (2003). On actor-critic algorithms. *SIAM Journal of Control Optimization*, *42*(4), 1143–1166.

Konorski, J. (1967). *Integrative activity of the brain: An interdisciplinary approach.* Chicago: University of Chicago Press.

Krettek, J. E., & Price, J. L. (1977). Projections from the amygdaloid complex to the cerebral cortex and thalamus in the rat and cat. *J Comp Neurol*, *172*(4), 687–722.

Lauwereyns, J., Watanabe, K., Coe, B., & Hikosaka, O. (2002a). A neural correlate of response bias in monkey caudate nucleus. *Nature*, *418*(6896), 413-417.

Lauwereyns, J., Watanabe, K., Coe, B., & Hikosaka, O. (2002b). A neural correlate of response bias in monkey caudate nucleus. *Nature*, *418*(6896), 413–417.

Le Moal, M., & Simon, H. (1991). Mesocorticolimbic dopaminergic network: functional and regulatory roles. *Physiological Review*, *71*, 155-234.

Lee, H. J., Youn, J. M., O, M. J., Gallagher, M., & Holland, P. C. (2006). Role of substantia nigra-amygdala connections in surprise-induced enhancement of attention. *Journal of Neuroscience*, *26*(22), 6077–6081.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopaminergic neurons during learning of behavioral reactions. *J Neurophys*, *67*, 145-163.

Ljungberg, T., & Enquist, M. (1987). Disruptive effects of low doses of *d*-amphetamine on the ability of rats to organize behaviour into functional sequences. *Psychopharmacology*, *93*, 146-151.

Lodge, D. J., & Grace, A. A. (2005). The hippocampus modulates dopamine neuron responsivity by regulating the intensity of phasic neuron activation. *Neuropsychopharmacology*, *Advanced online publication*.

Lodge, D. J., & Grace, A. A. (2006). The laterodorsal tegmentum is essential for burst firing of ventral tegmental area dopamine neurons. *Proceedings of the National Acadeny of Science USA*, *103*(13), 5167-5172.

Lopez, M., Balleine, B., & Dickinson, A. (1992). Incentive learning and the motivational control of instrumental performance by thirst. *Animal Learning & Behavior*, *20*(4), 322-328.

Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 225–247.

Lyon, M., & Robbins, T. W. (1975). The action of central nervous system stimulant drugs: a general theory concerning amphetamine effects. In *Curr dev psychopharmacol* (p. 80-163). New York: Spectrum.

Mackintosh, N. J. (1974). *The psychology of animal learning.* New York, NY: Academic Press.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms and empirical results. *Machine Learning*, *22*, 1-38.

Mahadevan, S. (1998). Partially observable semi-Markov decision processes: Theory and applications in engineering and cognitive science. In *Aaai fall symposium on planning with partially observable markov decision processes* (p. 113-120).

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman & Co.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*.

Mazur, J. A. (1983). Steady-state performance on fixed-, mixed-, and random-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *39*(2), 293-307.

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends in Neuroscience*, *26*(8), 423-428.

McDonald, A. J. (2003). Is there an amygdala and how far does it extend? an anatomical perspective. *Annals of the New York Academy of Science*, *985*, 1–21.

McDowell, J. J., & Wixted, J. T. (1986). Variable-ratio schedules as variable-interval schedules with linear feedback loops. *Journal of the Experimental Analysis of Behavior*, *46*, 315–329.

McSweeney, F. K., Melville, C. L., Buck, M. A., & Whipple, J. E. (1983). Local rates of responding and reinforcement during concurrent schedules. *Journal of the Experimental Analysis of Behavior*, *40*(1), 79–98.

Millan, J.-D. R., Posenato, D., & Dedieu, E. (2002). Continuous-action Q-learning. *Machine Learning*, *49*, 247–265.

Miller, R., & Wickens, J. R. (1991). Corticostriatal cell assemblies in selective attention and in representation of predictable and controllable events. *Concepts in Neuroscience*, *2*(1).

Mingote, S., Weber, S. M., Ishiwari, K., Correa, M., & Salamone, J. D. (2005). Ratio and time requirements on operant schedules: effort-related effects of nucleus accumbens dopamine depletions. *European Journal of Neuroscience*, *21*, 1749-1757.

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, *379*, 449-451.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, *377*, 725-728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*(5), 1936-1947.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*, 760-767.

Montague, P. R., McClure, S. M., Baldwin, P. R., Phillips, P. E. M., Budygin, E. A., Stuber, G. D., Kilpatrick, M. R., & Wightman, R. M. (2004). Dynamic gain control of dopamine delivery in freely moving animals. *Journal of Neuroscience*, *24*(7), 1754–1759.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, *43*(1), 133–143.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057–1063.

Murschall, A., & Hauber, W. (2006). Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance. *Learning & Memory*, *13*, 123-126.

Nelson, A., & Killcross, S. (2006). Amphetmine exposure enhances habit formation. *Journal of Neuroscience*, *26*(14), 3805–3812.

Nicola, S. M., Surmeier, J., & Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual Reviews in Neuroscience*, *23*, 185–215.

Niv, Y. (2007). Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Annals of the New York Academy of Science*, *1104*, 357–376.

Niv, Y., Daw, N. D., & Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *NIPS 18* (p. 1019-1026). MIT Press.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2005). Motivational effects on behavior: Towards a reinforcement learning model of rates of responding. In *Cosyne.* Salt Lake City, Utah.

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*, 6.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Science*, *10*(8), 375–381.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, *33*(5), 815–826.

Ohyama, T., Horvitz, J. C., Drew, M. R., Gibbon, J., Malapani, C., & Balsam, P. D. (2000). Conditioned and unconditioned behavioral-cognitive effects of a dopamine antagonist in rats. *Behavioral Neuroscience*, *114*(6), 1251–1255.

Ohyama, T., Horvitz, J. C., Kitsos, E., & Balsam, P. D. (2001). The role of dopamine in the timing of Pavlovian conditioned keypecking in ring doves. *Pharmacol Biochem Behav*, *69*(3-4), 617–627.

Packard, M. G., & White, N. M. (1991). Dissociation of hippocampus and caudate nucleus memory systems by posttraining intracerebral injection of dopamine agonists. *Behavioral Neuroscience*, *105*(2), 295–306.

Parent, A., & Hazrati, L. N. (1993). Anatomical aspects of information processing in primate basal ganglia. *Trends Neurosci*, *16*(3), 111-116.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*(1), 61–73.

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*(4), 587–607.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532-552.

Peele, D. B., Casey, J., & Silberberg, A. (1984). Primary of interresponse-time reinforcement in accounting for rate differences under variable-ratio and variable interval schedules. *Journal of experimental psychology: Animal behavior processes*, *10*(2), 149-167.

Phillips, P. E. M., Stuber, G. D., Heien, M. L. A. V., Wightman, R. M., & Carelli, R. M. (2003). Subsecond dopamine release promotes cocaine seeking. *Nature*, *422*, 614-618.

Phillips, P. E. M., & Wightman, R. M. (2004). Extrasynaptic dopamine and phasic neuronal activity. *Nature Neuroscience*, *7*, 199.

Pliskoff, S. S., & Fetterman, J. G. (1981). Undermatching and overmatching: The fixed-ratio changeover requirement. *Journal of the Experimental Analysis of Behavior*, *36*(1), 21–27.

Precup, D., Sutton, R. S., & Singh, S. P. (1998). Theoretical results on reinforcement learning with temporally abstract options. In *European conference on machine learning* (pp. 382–393).

Puterman, M. L. (1994). *Markov Decision Processes: Discrete stochastic dynamic programming.* John Wiley & Sons, Inc. New York, NY, USA.

Ratcliff, R., Zandt, T. V., & McKoon, G. l. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300.

Reber, A. S. (1985). *The Penguin dictionary of psychology.* Penguin Books.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci*, *22*(4), 146–151.

Reed, P., Soh, M., Hildebrandt, T., DeJongh, J., & Shek, W. Y. (2000). Free-operant performance on variable interval schedules with a linear feedback loop: no evidence for molar sensitivities in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *26*(4), 416–427.

Reed, P., & Wright, J. E. (1988). Effects of magnitude of food reinforcement on free operant response rates. *Journal of the Experimental Analysis of Behavior*, *49*, 75–85.

Rescorla, R. A. (1982). Some consequences of associations between the excitor and the inhibitor in a conditioned inhibition paradigm. *Journal of Experimental Psychology: Animal Behavior Processes*, *8*(3), 288-298.

Rescorla, R. A. (1994). A note on depression of instrumental responding after one trial of outcome devaluation. *The Quarterly Journal of Experimental Psychology*, *47B*(1), 27-37.

Rescorla, R. A., & Solomon, R. L. (1967). Two process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*(3), 151-182.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II* (p. 64-99). Appleton-Century-Crofts.

Robbins, T. W., & Everitt, B. J. (1996). Neurobehavioural mechanisms of reward and motivation. *Curr Opin Neurobiol*, *6*, 228-236.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (Submitted). Dopamine neurons encode the more valued option when deciding between immediate versus delayed gratification. *Nature Neuroscience*.

Roesch, M. R., & Olson, C. R. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. *Science*, *304*(5668), 307–310.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*(5), 511–524.

Roitman, M. F., Stuber, G. D., Phillips, P. E. M., Wightman, R. M., & Carelli, R. M. (2004). Dopamine operates as a subsecond modulator of food seeking. *Journal of Neuroscience*, *24*(6), 1265-1271.

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-intiated arm movements. *The Journal of Neurophysiology*, *63*, 592-606.

Rushworth, M. F. S., Kennerley, S. W., & Walton, M. E. (2005). Cognitive neuroscience: resolving conflict in and over the medial frontal cortex. *Curr Biol*, *15*(2), R54–R56.

Sabes, P. N. (1993). Approximating Q-values with basis function representations. In *Proceedings of the 1993 connectionist models summer school.* Hillsdale, NJ: Erlbaum.

Sabes, P. N., & Jordan, M. I. (1996). Reinforcement learning by probability matching. In *Advances in neural information processing systems 8.* Cambridge, MA: MIT Press.

Sah, P., Faber, E. S. L., Lopez De Armentia, M., & Power, J. (2003). The amygdaloid complex: anatomy and physiology. *Physiological Review*, *83*(3), 803–834.

Salamone, J. D., Aberman, J. E., Sokolowski, J. D., & Cousins, M. S. (1999). Nucleus accumbens dopamine and rate of responding: Neurochemical and behavioral studies. *Psychobiology*, *27*(2), 236-247.

Salamone, J. D., & Correa, M. (2002). Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behavioural Brain Research*, *137*, 3-25.

Salamone, J. D., Correa, M., Farrar, A., & Mingote, S. M. (2007). Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology (Berl)*, *191*(3), 461–482.

Salamone, J. D., Wisniecki, A., Carlson, B. B., & Correa, M. (2001). Nucleus accumbens dopamine depletions make animals highly sensitive to high fixed ratio requirements but do not impair primary food reinforcement. *Neuroscience*, *5*(4), 863-870.

Samuels, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*, 210–229.

Santamaria, J. C., Sutton, R. S., & Ram, A. (1998). Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive Behavior*, *6*(2), 163–217.

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *Journal of Neuroscience*, *23*(30), 9913-9923.

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, *1*(2), 155–159.

Schoenbaum, G., Nugent, S. L., Saddoris, M. P., & Setlow, B. (2002). Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport*, *13*(6), 885–890.

Schoenbaum, G., & Roesch, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron*, *47*(5), 633–636.

Schoenbaum, G., Setlow, B., Nugent, S. L., Saddoris, M. P., & Gallagher, M. (2003). Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. *Learning & Memory*, *10*(2), 129–140.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1-27.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during succesive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900-913.

Schultz, W., Apicella, P., Scarnati, P., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*, 4595-4610.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599.

Schwartz, A. (1993a). Thinking locally to act globally: A novel approach to reinforcement learning. In *Proceedings of the fifth annual conference of the cognitive science sociery.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwartz, A. (1993b). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning* (p. 298-305). Morgan Kaufmann.

Sellings, L. H. L., & Clarke, P. B. S. (2003). Segregation of amphetamine reward and locomotor stimulation between nucleus accumbens medial shell and core. *Journal of Neuroscience*, *23*(15), 6295–6303.

Shull, R. L., & Pliskoff, S. S. (1967). Changeover delay and concurrent schedules: Some effects on relative performance measures. *Journal of the Experimental Analysis of Behavior*, *10*, 517–527.

Singh, S. P. (1994). Reinforcement learning algorithms for average-payoff Markovian decision processes. In *Proceedings of the 12th aaai* (pp. 700–705). MIT Press.

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *Journal of General Psychology*, *12*, 66–77.

Smith, A. J. (2002a). Applications of the self-organising map to reinforcement learning. *Neural Networks*, *15*(8-9), 1107–1124.

Smith, A. J. (2002b). *Dynamic generalisation of continuous action spaces in reinforcement learning: A neurally inspired approach.* Unpublished doctoral dissertation, Division of Informatics, Edinburgh University, UK.

Sokolowski, J. D., & Salamone, J. D. (1998). The role of accumbens dopamine in lever pressing and response allocation: Effects of 6-OHDA injected into core and dorsomedial shell. *Pharmacol Biochem Behav*, *59*(3), 557-566.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychological Review*, *81*, 119-145.

Soto, P. L., McDowell, J. J., & Dallery, J. (2005). Effects of adding a second reinforcement alternative: implications for herrnstein's interpretation of r(e). *Journal of the Experimental Analysis of Behavior*, *84*(2), 185–225.

Soto, P. L., McDowell, J. J., & Dallery, J. (2006). Feedback functions, optimization, and the relation of response rate to reinforcer rate. *Journal of the Experimental Analysis of Behavior*, *85*(1), 57–71.

Staddon, J. (1992). Rationality, melioration, and law-of-effect models for choice. *Psychological Science*, *3*, 136–141.

Staddon, J. E. R. (2001). *Adaptive dynamics.* Cambridge, Mass.: MIT Press.

Staddon, J. E. R., & Motheral, S. (1978). On matching and maximizing in operant choice experiments. *Psychological Review*, *85*, 436–444.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(5678), 1782–1787.

Suri, R. E. (2002). Td models of reward predictive responses in dopamine neurons. *Neural Networks*, *15*(4-6), 523–533.

Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*, 871-890.

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning*, *3*, 9-44.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135-170.

Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the ninth annual conference of the cognitive science society* (p. 355-378). Hillsdale,NJ: Erlbaum.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (p. 497-537). MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* MIT Press.

Sutton, R. S., Mcallester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems 12* (pp. 1057–1063). MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Swanson, L. W. (2003). The amygdala and its place in the cerebral hemisphere. *Annals of the New York Academy of Science*, *985*, 174–184.

Tadepalli, P., & Ok, D. (1996). Scaling up average reward reinforcement learning by approximating the domain models and the value function. In *Proceedings of the 13th international conference on machine learning.*

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*, *92*, 2520-2529.

Takikawa, Y., Kawagoe, R., Itoh, H., Nakahara, H., & Hikosaka, O. (2002). Modulation of saccadic eye movements by predicted reward outcome. *Experimental Brain Research*, *142*(2), 284–291.

Takikawa, Y. K., Kawagoe, R. K., Itoh, H. K., Nakahara, H. K., & Hikosaka, O. K. (2002). Modulation of saccadic eye movements by predicted reward outcome. *Experimental Brain Research*, *142*(2), 284-291.

Taylor, J. R., & Robbins, T. W. (1984). Enhanced behavioural control by conditioned reinforcers following microinjections of *d*-amphetamine into the nucleus accumbens. *Psychopharmacology*, *84*, 405-412.

Taylor, J. R., & Robbins, T. W. (1986). 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens *d*-amphetamine. *Psychopharmacology*, *90*(390-397).

Taylor, K. M., Mello, B. K., Horvitz, J. C., & Balsam, P. D. (2006). Amphetamine alters timed behavior but not time perception. In *Society for neuroscience abstracts* (Vol. 32: 572.15).

Thorndike, E. L. (1911). *Aminal intellignece: Experimental studies.* New York: Macmillan.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, *23*(32), 10402-10410.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642-1645.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tolman, E. C. (1949a). There is more than one kind of learning. *Psychological Review*, *56*, 144-155.

Tolman, E. C. (1949b). The nature and functioning of wants. *Psychological Review*, *56*, 357-369.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*(6729), 704–708.

Tremblay, L., & Schultz, W. (2000). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*(4), 1877–1885.

Tsitsiklis, J. N., & Van Roy, B. (2002). On average versus discounted reward temporal-difference learning. *Machine Learning*, *49*(2), 179-191.

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, *303*(5666), 2040–2042.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*(15), 4019–4026.

Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. A. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neuroscience*, *27*(8), 468–474.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43-48.

Walton, M. E., Bannerman, D. M., Alterescu, K., & Rushworth, M. F. S. (2003). Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. *Journal of Neuroscience*, *23*(16), 6475–6479.

Walton, M. E., Bannerman, D. M., & Rushworth, M. F. S. (2002). The role of rat medial frontal cortex in effort-based decision making. *Journal of Neuroscience*, *22*(24), 10996–11003.

Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. S. (2006). Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks*, *19*(8), 1302–1314.

Watanabe, M., Cromwell, H., Tremblay, L., Hollerman, J., Hikosaka, K., & Schultz, W. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. *Experimental Brain Research*, *140*(4), 511-518.

Watanabe, M., Cromwell, H. C., Tremblay, L., Hollerman, J. R., Hikosaka, K., & Schultz, W. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. *Experimental Brain Research*, *140*(4), 511–518.

Watkins, C. J. C. H. (1989). *Learning with delayed rewards.* Unpublished doctoral dissertation, Cambridge University, Cambridge, UK.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Wearden, J. H., & Burgess, I. S. (1982). Matching since baum (1979). *Journal of the Experimental Analysis of Behavior*, *38*(3), 339–348.

Weiner, I., & Joel, D. (2002). Dopamine in schizophrenia: Dysfunctional information processing in basal ganglia-thalamocortical split circuits. In G. D. Chiara (Ed.), *Handbook of experimental pharmacology vol. 154/II, dopamine in the CNS II* (p. 417-472). Berlin: Springer-Verlag.

White, D. J. (1963). Dynamic programming, Markov chains, and the method of successive approximations. *J Math Anal Appl*, *6*, 373–376.

Wickens, J. (1990). Striatal dopamine in motor activation and reward-mediated learning: Steps towards a unifying model. *Journal of Neural Transmission*, *80*, 9-31.

Wickens, J., & Kötter, R. (1995). Cellular models of reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 187-214). MIT Press.

Williams, B. A. (1994). Reinforcement and choice. In *Animal learning and cognition* (p. 81-108). Academic Press.

Williams, D. R., & Williams, H. (1969). Auto-maintenance in pigeon: Sustained pecking despite contingent nonreinforcement. *Journal of the Experimental Analysis of Behavior*, *12*, 511–520.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256.

Wilson, C., Nomikos, G. G., Collu, M., & Fibiger, H. C. (1995). Dopaminergic correlates of motivated behavior: Importance of drive. *Journal of Neuroscience*, *15*(7), 5169-5178.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*, 483-495.

Wise, R. A., Spindler, J., deWit, H., & Gerberg, G. J. (1978). Neuroleptic-induced "anhedonia" in rats: pimozide blocks reward quality of food. *Science*, *201*(4352), 262–264.

Wyvell, C. L., & Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: Enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *Journal of Neuroscience*, *20*(21), 8122–8130.

Yerkes, R. M., & Morgulis, S. (1909). The method of Pawlow in animal psychology. *Psychological Bulletin*, *6*, 257–273.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of the dosolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 505–512.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.

Zald, D. H., & Kim, S. W. (1996). Anatomy and function of the orbital frontal cortex, i: anatomy, neurocircuitry; and obsessive-compulsive disorder. *J Neuropsychiatry Clin Neurosci*, *8*(2), 125–138.

Zuriff, G. E. (1970). A comparison of variable-ratio and variable-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, *13*, 369-374.