Molar and Molecular Models of Performance for Rewarding Brain Stimulation

Yannick-André Breton

A Thesis
In the Department
of
Psychology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Psychology) at
Concordia University
Montreal, Quebec, Canada

August, 2013

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Yannick-André Breton

Entitled: Molar and molecular models of performance for rewarding brain stimulation

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Psychology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Richard Courtemanche — Chair

C. Randy Gallistel — External Examiner

Lea Popovic — External to Program

Andreas Arvanitogiannis — Examiner

Rick Gurnsey — Examiner

Peter Shizgal — Thesis Supervisor

Approved by

C. Andrew Chapman
Chair of Department or Graduate Program Director

29/10/2013

Joanne Locke
Dean of Faculty

Abstract

Molar and Molecular Models of Performance for Rewarding Brain Stimulation

Yannick-André Breton, Ph.D.

Concordia University, 2013

This dissertation reports analyses of performance for rewarding brain stimulation in a three-part sequential task. A session of self-stimulation was composed of three trial types, during which the strength and opportunity cost of electrical stimulation were kept constant. Within a trial, a lever allowed animals to harvest brain stimulation rewards as soon as the lever had been held for a set, cumulative amount of time. When the time spent holding reached this criterion, the lever retracted and a burst of rewarding electrical stimulation was delivered. A flashing house light and 10s inter-trial interval signalled the start of a new trial. Rats were presented with strong/inexpensive/certain stimulation on one trial, a randomly selected strength, cost and risk on the next trial, and weak/inexpensive/certain stimulation on the third trial of a sequence. The sequence then repeated. Rewards during the second trial of this sequence were delivered with cued probabilities ranging from 0.5 to 1.0. The current thesis evaluates the ability of a previously published (molar) model of performance during a trial to accurately detect the effect of risk on payoff but not reward intensity. Although animals were less willing to work for stimulation trains that may not be delivered than those delivered with certainty, risk did not change the relative reward produced by stimulation. We also present evidence on a fine time scale that self-stimulating rats develop a model of their world. The first pause made

as a trial began was a function of the payoff the animal had yet to receive, indicating that rats had a model of the triad sequence. Analysis of the conditions under which pauses were uncharacteristic also provides evidence of what this model might be. Analysis of the fine scale of performance provides evidence that animals had a model of the stability of trial conditions. Finally, we present a (molecular) model of performance for brain stimulation rewards in real-time. Our results demonstrate that rats develop a model of the testing paradigm and can adjust to changes in reward contingencies with as few as one exemplar.

Acknowledgements

Sir Isaac Newton famously wrote in his letter to Robert Hooke that "if I have seen further, it is by standing on the shoulders of giants." Obviously, I wouldn't dare compare myself to the father of classical mechanics, or this thesis to the *Principia Mathematica*. I would, however, like to start by thanking all of those who made this possible, all of those who made this bearable, and all of those who contributed to the thinking that I attempted (and in some rare instances, accomplished) to communicate here.

First, I owe a great deal to Dr. Peter Shizgal and the support he has provided over the course of my graduate studies. He has forgotten more about reward than I could ever learn in a lifetime. Our weekly, bi-weekly, and monthly meetings, sometimes spent strolling the grounds of the University, have provided most of the reasoning that has gone into the experiments and analyses described here. While we're at it, I would like to thank the Shizgal lab members, present and past, for all the help you gave along the way. Rebecca Solomon graciously allowed me to use the data she collected as part of her own graduate work for many of the studies presented here. I'd also like to especially thank Dr. Kent Conover, who has been absolutely instrumental to my learning MATLAB.

On a similar note, our multi-way, multi-timezone teleconferences with Dr. Peter Dayan and his brilliant student, Ritwik Niyogi, have provided so much fodder for intellectual discourse that, given enough time and energy, a completely separate dissertation could also be written. I look forward to continuing our amicable collaboration.

I have been blessed with an outstanding committee, whose contributions cannot be ignored. Thank you Drs. Andreas Arvanitogiannis, Rick Gurnsey, and Lea Popovic for your insightful questions on the many drafts this document has gone through. Thanks also go to Dr. C. Randy Gallistel for his idea to use Bayesian

methods in chapter 4—the exercise has been truly instructive and will continue to serve me well in future endeavours.

The Centre for Studies in Behavioural Neurobiology, and especially its technical officer, David Munro, and system manager, Steve Cabillio, are owed much gratitude. Without them, the PREF3 computerized system for data acquisition would still be a collection of Med-Associates levers and plexiglas sheets in a closet somewhere. In fact, I'd like to extend my thanks to the students of the CSBN, who provided the necessary moral support when times were tough, the necessary scientific scepticism when bouncing ideas, and the necessary companionship when came time to celebrate. It has been an honour and a pleasure to work alongside you all.

Family, friends, and partner—thank you. Thank you mom, dad, Sacha, Michael, Andrew, Mark, Mehdi, Stephanie, Dean, Sherri, Amélie, Erin, and, obviously, *Tiana*, for dealing with my craziness while writing this. I really couldn't ask for a better support system if I wanted to. Thank you for being there, for not slapping me silly, and, when warranted, for figuratively slapping sense into me.

Many thanks to Dr. A. David Redish for his understanding, and the entire Redish lab for their help. I promise to get right back to post-doctoral work as soon as humanly possible.

Finally, I would like to thank some of the people who paid for this. Financial support was provided by a *Bourse de doctorat en Recherche* grant from the *Fonds Québécois de la Recherche sur la Nature et les Technologies*.

One last note: if I forgot to mention you, but should have, my apologies. Writing a thesis is a lot like what I'm told is involved in tending to a newborn: your sleep-wake cycle is disrupted, your eating patterns are erratic, you obsess over little things, and you worry about screwing up in some major way. There's also some crying in the middle of the night, but that's beside the point. It's easy to lose track of everyone who had a hand in your success. Thank you!

Dedication

For Miguelina Gomez-Walsh and Jean-Daniel Breton.

Table of Contents

List of Figures

Chapter 1

Background

A rat sits idly by the corner of his cage, carefully grooming his upper body in comforting darkness. He licks one paw, then the other, running it along the mound of acrylic atop his head. The implant has been part of his proprioception for some time. Midway through his grooming bout, a large flashing light stops him in his tracks. As the flashing stops, the rat leaps to the lever and begins patiently holding, periodically tapping it. Nothing can tear him away from the manipulandum. He steadies the lever down with his snout, teeth digging into the metal paddle, paws slipping. Soon, he receives his reward. This rat is neither hungry, nor thirsty, nor in a position to gain access to an oestrous female. The reward this animal receives can only be detected on an oscilloscope: he is tirelessly working for a brief burst of cathodal pulses that will be delivered via implanted electrodes to the lateral hypothalamic level of the medial forebrain bundle.

The phenomenon of brain stimulation reward was first discovered by Olds & Milner (1954). In their preparation, animals would quickly return to a location that had been paired with the delivery of electrical stimulation. Before long, instrumental methods were used to investigate the behavioural effects of electrical stimulation.

Electrical stimulation provides at least three distinct advantages over natural rewards like food and water. First, there is no satiety for electrical stimulation (Olds, 1958). An animal may cease to be hungry, or cease to desire a specific type of food, but they do not cease to seek out strong and easily acquired electrical stimulation trains. Second, the electrode is implanted within a substrate that is identifiable in principle and can compete with, summate with (Conover and Shizgal, 1994; Conover et al., 1994), and act as an economic substitute for (Green and Rachlin, 1991) natural rewards. Food, water, and sex may activate the circuitry of valuation at some point

in the animal's pursuit of each, but only electrical stimulation provides a probe into a common evaluative circuitry. Finally, the rewarding effect produced by electrical stimulation can be tightly controlled experimentally. Although more food may be more valuable to a hungry animal than less food, the exact degree to which the animal values every ounce of food is much less under experimental control than the signal that is artificially injected into the medial forebrain bundle.

1.1 A brief history of the psychophysics of reward

Brain stimulation reward offers the exemplar *par excellence* of motivated behaviour. We can tightly control the subjective rewarding impact of the stimulation by way of the current and pulse duration, which set the spread of activation, the pulse frequency, which sets the induced firing rate, and the train duration, which sets the activation period. There is no satiety, so animals will work tirelessly for many hours without filling up on coulombs of electricity. The electrode is in an identifiable substrate; as a result, it is possible to derive (at least in principle) some characteristics of the neurons responsible for the rewarding effect of stimulation.

The first manipulations of intracranial self-stimulation simply assessed whether the response rate was affected. The logic was simple: if a particular manipulation boosted the value of rewards, it followed that animals would work more vigorously for brain stimulation under those conditions. Although initial studies by Olds (1956) mapping the rewarding effect across brain regions employed a rough estimate of the percentage of session time spent responding, subsequent studies focused on changes in self-stimulation rates that resulted from a particular manipulation. For example, Brady (1957) measured response rates for brain stimulation reward as a function of the duration of a food or water deprivation, and found significant increases the longer the rat had been deprived. These types of measures may, at best, detect that a variable

affects self-stimulation performance, and little else. At worst, rate measures fail to reveal important effects or identify effects that may not be related to the motivation for brain stimulation rewards.

It was not long before critics of the non-parametric approach began voicing concerns. In their seminal article, Hodos & Valenstein (1962) pointed out that rate alone may not be an accurate measure of the rewarding impact of brain stimulation reward. Although others had parametrically varied the degree of training on self-stimulation (Bindra and Mendelson, 1962) and electrical stimulation parameters (Ward, 1959), this paper provided a solid argument for parametrically varying the strength of the stimulation, by way of the stimulation current, comparing performance for septal and posterior hypothalamic stimulation. The road had been paved for a parametric analysis of brain stimulation reward (Gallistel et al., 1974; Edmonds et al., 1974; Edmonds and Gallistel, 1974), allowing researchers to determine whether manipulations of the circuitry underlying reward valuation and action selection affected how the injected signal was impacted by lesions (Murray and Shizgal, 1991; Waraczynski, 2006) and pharmacological agents (Franklin, 1978; Hernandez et al., 2008).

The paradigm that emerged from this approach, the curve-shift paradigm (Miliaressis et al., 1986), was intended to allow researchers to determine whether a manipulation has affected the circuitry that underlies an animal's goal-directed behaviour. By assessing the rate of responding at various stimulation strengths, varying either the spread of activation via the current, or the injected spike rate via the pulse frequency, a manipulation that simply reduces responding can be distinguished from a manipulation that alters the animal's motivation to seek out rewarding stimulation. In essence, performance will vary from floor to ceiling levels along with a given stimulation parameter (pulse current, pulse duration, pulse frequency, and train duration). The stimulation parameter that drives half-maximal performance (often referred to

as M50) provides a meaningful comparison to that collected under a different set of conditions. For example, if cocaine reduces the pulse frequency that supports half-maximal performance without affecting the rat's maximum response rate (Hernandez et al., 2008), then cocaine boosts the animal's pursuit of non-maximal rewards at a given programmed rate of reinforcement. If Pimozide-induced dopamine depletion increases the pulse frequency that drives half-maximal performance without affecting the rat's maximum response rate (Phillips and LePiane, 1986), then Pimozide reduces the animal's motivation to seek out rewards. In the original study by Hodos and Valenstein (1962), although the rate of responding for septal stimulation was lower, overall, than posterior hypothalamic stimulation, the current required to produce a threshold level of responding was also lower for septal stimulation than posterior hypothalamic stimulation. Although rats responded less vigorously for septal stimulation, posterior hypothalamic stimulation required stronger stimulation in order to drive performance to a similar level.

The curve-shift paradigm is not without its problems. Fouriezos et al. (1990) evaluated the effect of increasing task difficulty (adding weight to the manipulandum) on self-stimulation thresholds derived from parametrically varying the pulse frequency of a 500ms stimulation train. The authors found that as lever loads increased from 0 to 45g, the rate of self-stimulation for the highest pulse frequencies decreased, but the pulse frequency required to drive half-maximal performance increased. Consequently, weighted levers, and possibly other challenges unrelated to reward valuation *per se*, are capable of changing the stimulation strength required to drive a threshold level of performance.

As a result of the inadequacies of the curve-shift paradigm in identifying the stage of processing at which a manipulation acts to alter reward seeking, Shizgal (Arvanitogiannis and Shizgal, 2008) developed a computational model of brain stimulation. The proportion of time allocated to self stimulation activities was assessed

as a function of the number of pulses delivered in a train of fixed duration and the experienced rate of reinforcement. As a result, manipulations that affect the translation of injected pulse rate can be distinguished from those that affect the translation of payoff into performance. The current thesis builds on this tradition by validating a molar computational model of performance for brain stimulation reward that assesses performance, indexed by the proportion of time an animal invests in harvesting rewards, as a function of the pulse frequency of the stimulation delivered and the amount of time the animal must invest to obtain such a reward. Furthermore, we propose a molecular model of performance and derive its molar predictions.

Although the psychophysics of brain stimulation reward have focused on molar measures of performance, collapsing performance across an entire trial or session, concerns about the obscuring effect of molar measures were voiced early-on. In their critique of response-rate measures, Hodos and Valenstein (1962) conceded that measures that preceded reinforcement rate, based on the proportion of session time spent responding, were insensitive to the pattern of responding. Molar measures generally cannot take into account the pattern and syntax of performance on every trial.

1.1.1 The patterning of behaviour

The tradition most obviously concerned with characterizing the molecular pattern of behaviour is reinforcement learning. In this account, performance reflects the learned value of various states and the actions that may be taken in those states. For example, pressing a lever for some small period of time dt allows the rat to be that much closer to a reward state that follows lever-pressing. Not pressing the lever for that period of time does not bring it closer to a reward, but that non-pressing state may have some value of its own. The rat implements a policy, or a probability of selecting an action in a particular state, that will maximize the rate at which it will be rewarded, based on the value of anticipated states and associated possible actions

that it has learned over the course of the trial, session, and experiment. This thesis takes inspiration from all these traditions in developing a molecular computational model of performance for brain stimulation reward.

1.2 Action selection

The problem of how animals select actions among all those available, and how this process may be implemented in the brain, has been approached most comprehensively by learning theorists interested in problems of conditioning. A deep philosophical tradition exists indeed regarding the power of associative learning in directing behaviour. From Aristotelian ideas describing the mind at birth of being a blank slate, to John Locke's conceptualization of human empiricism, associative learning has provided a useful methodological framework for understanding how agents select actions. That said, it is simplistic to assume that all actions are selected solely on the basis of conditioned associations between stimuli, responses, and outcomes. Even if they were, their representation is not likely to involve only associations to outcomes, and selection itself is not likely a reflexive system. Human phenomenology certainly suggests that many actions are selected following protracted deliberation, and it may not be particularly far-fetched to presume that similar, though much less complex, deliberative processes are at work in non-human animals, even those with continuously-growing upper and lower incisors.

The idea that two parallel systems compete for the selection of an action is not new. Greek mythology describes a dichotomy between Apollonian (god of reason) and Dionysian (god of intoxication) modes of thinking. William James (Boring, 1950) believed that a dual process governed behaviour, one which was purely associative and one which relied on reasoning and insight. Others (Kahneman, 2011) have since refined the distinction between a fast, automatized, habitual, intuitive and

affect-dependent system, and a slow, flexible, deliberative, reasoned and cognition-dependent system. The two systems may not be as unique to human experience as might be supposed at first glance. Rats faced with a difficult decision will pause at a choice point in a maze and move their heads back and forth (Tolman, 1948), implying a deliberative process (Johnson et al., 2007); this back-and-forth movement is more prominent early in training than later on and more prominent on early laps of a maze than later laps (Johnson and Redish, 2007). The problem of action selection across a wide range of animals is arguably likely to involve a dual process of habitual and deliberative systems, both competing for action selection.

1.2.1 Classical conditioning

The simplest description of action selection involves classical (Pavlovian) conditioning. In this paradigm, the animal learns an association between one stimulus—the conditional stimulus (CS)—and another stimulus—the unconditional stimulus (US)—thereby producing a conditional response (CR). Over multiple trials, an animal is presented with pairings between the CS, which has a neutral valence, and the US, which induces a reflexive unconditional response (UR) on its own. Eventually, the CS alone is capable of eliciting a CR.

It may be difficult to infer that an animal is "selecting" a response in this case, as it would be easy to assume the conditional response is a reflexive behaviour. Evidence to the contrary can be found in investigations of Pavlovian-to-instrumental transfer. In this case, a CS, such as light, is paired with an appetitive US, such as sucrose, and an instrumental action is subsequently paired with a different outcome. General Pavlovian-to-instrumental transfer occurs when the presentation of the CS, which had never before been presented in the instrumental setting, increases the rate of responding for a reinforcer that was not paired with the CS (Estes, 1943). Moreover, classically conditioned actions may compete with instrumental actions. For example,

an animal may avoid a lever that shuts off a loud noise if it is physically close to a cue that has been paired with the noise (Breland and Breland, 1961). These findings suggest that the association learned during classical conditioning affects action selection on a higher level than the simple reflex.

In fact, one can trace the birth of computational reinforcement-learning models to classical conditioning. In an effort to explain the inability of a new stimulus, CS2, to acquire the ability to produce a CR when the US had already been paired (pre-conditioned) with a different CS1, Rescorla and Wagner (1972) developed a model of classical conditioning based on violated expectation. When CS1 is pre-conditioned with the US, the US is perfectly predicted by CS1; subsequent pairings of CS2 with CS1 and the US fail to produce a learned association between CS2 and the US because no expectation is violated. The Rescorla-Wagner model implies that the degree to which new learning occurs depends on the level of learning that has already taken place.

1.2.2 Instrumental conditioning

The most direct description of action selection occurs in free-operant instrumental conditioning, during which an animal learns the relationship between an action and a desirable or undesirable consequence of that action. The animal chooses to spend its time working for experimenter-programmed rewards or for the rewards it derives from all other actions it may perform in the operant chamber. Normative models of action selection assume that the partition of time the rat makes between operant responding and everything else reflects a partition of time the rat deems optimal. The dimension along which the rat is optimizing is operationalized as utility; if the rat spends half of its time lever-pressing, then the inference is that the rat derives maximum utility from allocating 50% of his time to experimenter-programmed rewards and 50% of his time to non-experimenter related rewards.

The foundations for this analysis run deep in the behaviourist tradition in psychology. Allison (1983) characterized operant performance as the result of an economic contract between subject and experimenter. For example, the experimenter will deliver one pellet of food for every three lever presses the rat makes, and the rat is free to allocate its time to fulfilling that economic contract as it deems fit. Similar microeconomic analyses are made of human behaviour: an individual is free to allocate his or her time to pursuing the fruits of labour, or to pursuing the fruits of leisure. The labourer balances the amount of time spent working with the amount of time spent away from work such that the overall utility of a particular partitioning of work and leisure is subjectively optimal. Just as a labourer cannot pursue both goals at once, the rat must necessarily trade off time it would otherwise spend grooming and resting for the time it must spend pursuing experimenter-programmed rewards. As the rat can't hold the lever and groom at the same time, it necessarily sacrifices one for the other, and must select the action that maximizes the utility of a particular partitioning of operant performance and everything else.

Microeconomic accounts of animal behaviour (Kagel et al., 1995) propose that animals partition their time in an operant setting according to an underlying utility-maximizing rule. Just as a graduate student with a fixed budget to buy bread and peanut butter is presumed to maximize the best combination of these goods in their shopping basket, a lever-pressing rat is thought to allocate their fixed "budget" of lever presses to the goals of eating and drinking according to the subjectively optimal combination of food and water (Green and Rachlin, 1991). In a single operant context, because the goods that can be acquired by lever-pressing (for example, electrical stimulation) are only partly substitutable for those that can be acquired from extraneous activities (like grooming), the rat's "investment" into pursuit of experimenter-controlled and extraneously-delivered rewards changes as the temporal and energetic budget is manipulated. Assuming perfect information, were the two types of rewards

completely fungible, the underlying utility-maximizing rule would have to be a step function: as soon as one was even slightly better than the other, the rat would spend its time pursuing one to the exclusion of the other. The same is true of any savvy graduate student: if, while holding all other factors equal, the cost of one peanut butter jar was even slightly lower than the cost of another, they would exclusively buy that which drained the budget less.

1.2.3 Experimental control

In the traditional variable-interval schedule of reinforcement used to assess instrumental conditioning, animals are rewarded following the first response after a lever is armed. The levers are armed at intervals drawn from an exponential distribution, thereby providing no time signal about when they are likely to be armed, since the probability they will be armed is constant over time. Such a schedule is more likely to produce steady responding (Skinner et al., 1997), providing experimenters with a large quantity of data.

However, animals may take advantage of the nature of these infinite-hold variable-interval schedules of reinforcement. The animal can sacrifice a small number of rewards by waiting sufficiently long, since the lever will continue to be armed for so long as a reward has not been harvested. As a result, there will be little need for the self-stimulating rat to trade off the time spent working for electrical rewards that come at a low rate with that spent engaged in other activities. By simply waiting, obtaining the fruits of leisure, the probability that a reward is waiting will increase, and the rat can therefore almost guarantee that it can also obtain the fruits of lever-pressing while it waits for rewards (Arvanitogiannis, 1997).

One way to control for this is to ensure that the rewards from self-stimulation must be traded off with the rewards from other activities by enforcing a free-running, zero-hold variable interval schedule (Conover and Shizgal, 2005; Breton et al., 2009).

In this procedure, rewards are only delivered when the lever is depressed at the end of the interval. If it is not, a new interval is drawn and the rat has missed an opportunity to obtain rewards. As in the wild, unless the rat is actively foraging, the fruits of its labours may spoil, but while the rat is foraging, it cannot pursue other goals. On this view, the time spent in one activity truly imposes an opportunity cost that the rat cannot avoid.

The nature of exponential distributions of latencies is such that a large number of samples must be drawn before the mean of that exponential distribution can be known with any confidence. However, previous work (Breton et al., 2009) has shown that when the price (the reciprocal of the rate of reinforcement) is held constant over long periods of time, its lower evaluability makes rats much more insensitive to changes in price and produces inconsistencies in behavioural allocation. When the price changes often, its value is much more salient, and no inconsistencies are observed.

Experimental control over both the trade-off imposed by and the evaluability of the rate of reinforcement—or more accurately, its reciprocal, the price—can be achieved using a schedule of reinforcement inspired by behavioural ecology. In the fixed cumulative handling time schedule, animals must invest a fixed amount of time, accumulated over possibly many bouts, into acquiring the reward on offer. One hallmark of such a schedule is that the number of rewards that can be harvested is directly proportional to the amount of time spent engaged in the activity (unlike traditional variable-interval schedules). Another is that the fixed time requirement can, at least in principle, be extracted from a single exemplar (unlike free-running variable interval schedules). As a result, eliminating the variability in price facilitates measurement of the rate at which animals can learn the work requirement in effect during a trial.

1.2.4 "Cognitive" action selection

Not all action selection is the direct result of a learned association between stimuli, responses, and rewards. Tolman famously proposed (Tolman, 1948) that maze learning in the rat proceeded, not by a sequence of associative response-reward pairings, but instead by the establishment of a cognitive map by which animals could navigate. Using a metaphor that many today would read as quaint, Tolman likened the learning of complex mazes to requests to a map control room rather than a telephone switchboard. Rats that were free to explore a maze before it was baited with food learned where the food was located much more quickly than rats without this experience, suggesting that spatial navigation decisions can be based on non-associative components. Furthermore, when sated but thirsty rats were free to explore a Y-maze baited with food in one arm and water in the other, they approached the water location quickly when they were subsequently made hungry but not thirsty. Those that had not had the previous free exploration phase approached the food-baited arm much more slowly. If a simple process had associated the location with a representation of reward magnitude, one would expect pre-trained rats to make incorrect responses when their homoeostatic state changed. Some process must have occurred during the pre-training phase, above and beyond the reward magnitude to be found at each baited end: a process mapping the identity of the reward to be found to its spatial location.

What does spatial navigation have in common with instrumental conditioning? While it may be natural to describe the solution of a spatial navigation task in terms of maps, it is equally possible to describe instrumental conditioning tasks in terms of maps, too. When an animal is afforded the opportunity to acquire rewards, some process may map lever-pressing actions not only to the absolute reward intensity that can be derived from that lever, but also to the identity of the reward. To wit, rats

that are made ill from a food delivered in an operant setting may subsequently cease responding for the food (Dickinson et al., 1995; Belin et al., 2009), suggesting there is a representation of the identity of the reward to come.

This mapping may even be applicable to higher-level representations of the task. In traditional studies employing brain stimulation reward as the operant reinforcement, trials are presented in a repeating sequential order, from those delivering strong stimulation to those delivering weak stimulation. After a trial providing weak stimulation, the sequence repeats, and a new trial of very strong stimulation begins. Anecdotal evidence can be found for periodic, unexpectedly vigorous responding on trials for which stimulation should not be particularly motivating, even when there is no other cue. The trials have a fixed duration and follow a fixed sequence. When the stimulation is sufficiently weak, and a sufficient amount of time has passed, the reward on offer on the next trial will be very strong. By the end of a trial of fixed duration delivering sufficiently weak stimulation, an animal may come to anticipate the value of lever-pressing on the next trial. It may be possible—and we present formal evidence here—that experimental subjects are at least capable of a higher-order representation of how trials progress within a session. In other words, the control over which actions to perform and how long for which to engage in them is not only driven by strict associative learning, but also by a cognitive map of task demands based on statistical regularities in the environment that can be detected.

1.3 Computational models

In order to forage efficiently, animals must balance the anticipated costs and benefits of pursuing one option with those involved in pursuing all others. The basic question of action selection in this sense regards how animals make the trade-off and what the key determinants are that affect the goal to be pursued.

Quantitative models allow precise predictions about the effect of manipulations on performance. These models can come from the normative tradition, deriving what action should be selected from first principles, or from a more descriptive tradition, deriving what action will be selected from experimental findings. Of course, there is considerable overlap between these two approaches, because it is important first to identify what animals should do in order to determine where they stray from optimality.

At shorter time scales, the most common normative description of action selection is provided by temporal difference reinforcement learning models. The agent is conceived as an integrated pair consisting of a "critic," which determines the mapping of trial states to the total temporally-discounted sum of future rewards, and an "actor," which determines the optimal policy to implement and the sequence of actions that will bring about the greatest discounted sum of rewards at the greatest rate. Furthermore, two types of reinforcement learning are possible: learning about the temporally discounted value that is associated with a trial state (model-free learning), and learning about the mapping between trial states themselves (model-based learning). In one case, the rat maintains only a representation of the common currency required to judge the desirability of actions in particular states, while in the other, the rat maintains both the value of the state and qualitative aspects of how states and actions are interrelated. The relevant time scale for these paradigms is on an action-state pair level, and temporal difference reinforcement learning models usually describe the action selection problem in terms of punctate events: if the rat is in trial state S at time t, the action to be selected is that which will lead to a state S' at time $t+1$ that will bring about a larger net total reward in the long run.

Behaviour on an operant task can be parsed on multiple time scales. The Matching Law (Herrnstein, 1974), for example, states that the ratio of rates of responding on one operant compared to another will match the ratio of the rates of

reinforcement on each. If pecking a key delivers rewards at a rate of one grain per second, and another at a rate of one grain per four seconds, pigeons will peck at the first key at four times the rate of its responding at the second, and vice-versa for the second key. This phenomenon was described by Herrnstein (1961) in pigeons responding on concurrent variable interval schedules, and was soon generalized to the single-operant case. Assuming a constant amount of behaviour to allocate between a single operant and extraneous activities, the rate of operant responding is a function of only the rate at which it provides rewards and the rate at which extraneous activities provide rewards. Thus, the matching law provides a means of scaling the value of rewards by assuming that the rate of reinforcement from extraneous activities is constant and there is a constant amount of behaviour to be allocated to each. If a manipulation has altered the perceived rate of reinforcement, it will also change the response rate, such that a rate of reinforcement required to drive performance to a given level will be altered. The assumptions have not gone unchallenged in the case of brain stimulation rewards (Conover et al., 2001a) and although a change in reinforcement rate will produce a change in response rate, the converse cannot be said: changes in threshold rates of reinforcement required to drive performance to a given level are not necessarily the product of altered perception of the rate of reinforcement.

Such a description of action selection is intrinsically molar, and does not explain how time is partitioned among competing goals in an ongoing sense. Melioration is one attempt at explaining the first principle that guides ongoing action selection, though it has not been uncontroversial (Williams and Royalty, 1989; Vaughan, 1981). Maximization is another attempt (Green et al., 1983), though it, too, has had its detractors. Neither has been entirely successful in unequivocally explaining performance in a wide range of operant schedules of reinforcement. One proposal (Gallistel et al., 2001) has been that feed-forward mechanisms provide the self-stimulating rat with an interval over which to compute the expected rate of reinforcement, which,

when combined with the subjective reward magnitude, provides an expected income. Stay durations at each option—and therefore the selected action—are stochastic realizations of a stay-termination process that depends on the expected income from each option.

The following sections will elaborate on these two different accounts of performance for rewards in the context of brain stimulation rewards: normative models, based on temporal-difference reinforcement learning, and a descriptive model, based on Herrnstein's Matching Law.

1.3.1 Markov decision processes

Typical temporal-difference reinforcement learning (TDRL) models boil the problem of action selection down to a Markov decision process (MDP) in which the rat attempts to maximize its total net reward, subject to costs it incurs by selecting one action in a given state of the world rather than all others. Strictly speaking, a MDP refers to any control process—such as action selection—in which the only relevant state to consider is the current state. In other words, the probability of moving from one state to any other state is independent of any previous state. In its discrete formulation, a MDP simply requires that the probability of moving from state S at time t (S_t) to state S' at time $t+1$ (S'_{t+1}) depends only on S_t, or

$$\mathcal{P}[S_t|S_{t-1}, S_{t-2}, \ldots, S_0; a_{t-1}, a_{t-2}, \ldots, a_0] = \mathcal{P}[S_t|S_{t-1}, a_{t-1}].$$

The original problem was formulated by Richard Bellman (1957) in solving his shortest path problem. Suppose there are multiple roads that lead from location A (New York) to location Z (Los Angeles), which may each pass through intermediate destinations. The problem is to find the shortest path from A to Z. Each point on the map is connected by roads of varying length. Although an individual may collect

some amount of reward (R) immediately upon visiting a point on the map (a state S), taking a particular road (an action) will incur a cost (C) related to how long the road is. The net reward from travelling from point A to Z on the map, assuming a sequence of actions given by π, will be the sum of the net rewards ($R - C$) obtained by following that particular route. The task, then, is to find the policy π which will maximize the expected total net reward from visiting all the states and completing all the actions inherent in the policy. Starting at state S_0, the policy will tell the decision-maker to take action a_0, which brings it to state S_1, where the policy will tell the decision-maker to take action a_1, which brings it to state S_2, and so on, until we reach the desired absorbing state S_n. The total expected net reward from following policy π in state S_0 will be the sum of all the rewards and costs incurred by visiting all the states and implementing all the actions in the sequence:

$$V_\pi(S_0) = R(S_0) - C(a_0) + R(S_1) - C(a_1) + \ldots + R(S_n) - C(a_n).$$

This sum can be re-written recursively. The total expected net reward from following policy π in any state S will be the immediate net reward, summed with the total expected net reward from following policy π in the state that follows it from taking action a. Following policy π in state S_0 will lead to state S_1; supposing we were to start in state S_1 rather than S_0, the total net reward from following policy π is

$$V_\pi(S_1) = R(S_1) - C(a_1) + R(S_2) - C(a_2) + \ldots + R(S_n) - C(a_n),$$

so the total net reward from following policy π in any given state is

$$V_\pi(S) = R(S) - C(a) + V_\pi(S'),$$

where S' is, in this deterministic case, the state that results from executing policy π

(taking action a in state S). This recursively formulated objective can be expressed as a set of linear equations: there is one V_π for each possible state S that can be visited, which imposes a set of linear constraints on what V_π could be. By solving for the system of n linear equations (for each $V_\pi(S)$) with n unknowns (for each S), it is possible to solve the total net rewards that arise from executing policy π.

The solution to the problem, then, is to find the policy for which this recursively defined "value function" (the total net rewards obtained by executing the policy) is maximal. In other words, if we define $V^*(S)$ as the value of state S when executing an optimal policy, then

$$V^*(S) = \max_\pi \{V_\pi(S)\}.$$

The value function for state S, when executing the optimal policy, will provide an immediate net reward, and will tell the agent to execute the action that leads to a future state with maximal expected net reward when taking the cost into account. In other words,

$$V^*(S) = \max_a \{R(S) - C(a) + V^*(S')\}$$

or the value of a state S when using an optimal policy is the maximum, over all potential actions a that can be performed, of the immediate net reward and future net rewards. This is usually called "Bellman's equation," and leads to the definition of the optimal policy: if a decision-maker is in a state S, which delivers a reward and imposes a cost as soon as it is entered, the optimal policy for state S, $\pi^*(S)$, will be

$$\pi^*(S) = \arg\max_a \{R(a) - C(a) + V^*(S')\}$$

or the action that will lead to the highest future net rate of reward when taking its cost into account, assuming the agent implements an optimal policy from that point

forward.

The original problem dealt with deterministic transitions; that is, when a decision-maker takes action a in state S, it leads unequivocally to state S'. The problem can be extended to non-deterministic state transitions by introducing the state-transition probability function $\mathcal{T} = \mathcal{P}[S'|S, a]$, which gives the probability of entering state S' when taking action a in state S. The value function therefore becomes,

$$V^*(S) = \max_a \left\{ R(S) - C(a) + \sum_{S'} (\mathcal{T} V^*(S')) \right\}$$

where the expected net reward over future states involves the sum over the random variable S', and the optimal policy becomes

$$\pi^*(S) = \arg\max_a \left\{ R(S) - C(a) + \sum_{S'} (\mathcal{T} V^*(S')) \right\}$$

as a consequence of that expectation.

The shortest path problem is isomorphic to a rat working for electrical stimulation. The normative solution is for the rat to find the shortest sequence of actions that will lead to the greatest net reward. It may engage in lever pressing, which will come at some opportunity and effort cost, but will eventually provide it with a brain stimulation reward; or it may engage in other activities, which will bring about their own intrinsic rewards but will not lead to electrical brain stimulation.

The original problem contains an absorbing state—Z—from which no further action is possible. In this case, calculating V^* is trivial, because there is a finite number of locations that will be visited before reaching this state. If, on the other hand, there is no absorbing state, there will be infinitely many locations visited following the current state, including (possibly) the current state again. As a result, the value of any given state could become infinite, because it sums the values of infinitely many future states. This is what is meant by the infinite-horizon problem. Since a rat

working for electrical rewards may not know the duration of a trial when it begins, and can revisit certain trial states arbitrarily often, the problem of action selection has, for all intents and purposes, an infinite horizon. If the trial state represented by the rat includes a measure of how much trial time is left, there is once again an absorbing state (the end of the trial), and action selection reduces to Bellman's original shortest-path problem.

In the absence of an absorbing state, the rat must discount rewards it expects to receive some time in the distant future. Without this discount factor, the MDP cannot be solved. Commonly, this is accomplished with an exponential discounting function. Rather than maximize the total net rewards, the solution requires maximizing the total net discounted rewards, where rewards in the future are less valuable than those that can be obtained immediately, and costs in the future loom less than those that must be incurred right away. Future rewards (and costs) are exponentially discounted by a factor of γ, where $\gamma = 0$ would indicate a decision-maker for whom future rewards and costs are irrelevant, and $\gamma = 1$ would indicate a decision-maker for whom all future rewards are as valuable as immediate rewards. The recursive expression of Bellman's equation becomes

$$V^*(S) = \max_a \left\{ R(S) - C(a) + \gamma \sum_{S'} (\mathcal{T} V^*(S')) \right\}$$

in this case, and the task is to find the optimal policy according to

$$\pi^*(S) = \arg\max_a \left\{ R(S) - C(a) + \gamma \sum_{S'} (\mathcal{T} V^*(S)) \right\}.$$

A large literature on inter-temporal choice generally contradicts this normative model. Exponential discounting would predict that an outcome is devalued at a constant rate across time: with a discount rate of 10% per day, a dollar tomorrow is worth 90 cents today, and a dollar in two days is worth 81 cents today. If two options

are presented—a small amount sooner (42$ in one day) or a larger amount (52$ in 10 days) later—seemingly impulsive Harvard undergraduates will tend to prefer the smaller/sooner option (Kirby and Herrnstein, 1995). This preference implies that the additional 10$ cannot overcome the temporal discount factor: a dollar in 10 days is worth less than 81 cents tomorrow, so it is preferable to select the sooner option to the later one. The finding itself is perfectly valid and ecologically reasonable. A resource that can be obtained now is preferable to one that can only be reaped later and which may no longer be available. As a result, a reasonable forager will discount future rewards compared to their value when immediately available. It would be foolish indeed for a restaurant owner to accept payment for a meal in six decades, or for a hungry diner to accept a wait of many days before being served. Temporal discounting is made explicit in temporal-difference reinforcement learning algorithms by giving exponentially less weight to future rewards as a function of time.

The important test of exponential discounting is whether displacing the two options by the same lag will reverse the preference. If the rate of discounting is constant, when a dollar tomorrow is only worth 81 cents in 10 days, a dollar in one week and 10 days is worth 81 cents in one week. In other words, Harvard undergraduates should also prefer 42$ in 6 days to 52$ in 16 days, and certainly to 52$ in 27 days. Alas, impulsive undergraduates (Kirby and Herrnstein, 1995), pigeons (Ainslie and Herrnstein, 1981) and rodents (Logan and Spanier, 1970) all appear to discount at a time-varying rate that has often been approximated by a hyperbolic curve. The 19 year-old Harvard undergraduate, for whom a dollar tomorrow was worth under 81 cents in 10 days, prefers an offer of 52$ in 27 days equally to 42$ in 6 days. In other words, although a constant discount rate implies that the smaller-sooner reward should be preferred when both options are displaced in time, the preference reverses (thereby suggesting a time-varying, hyperbolic discount rate) when the options are displaced by a median of 6 days. Some reformulations of reinforcement learning have

incorporated hyperbolic temporal discounting (Kurth-Nelson et al., 2012), but even these models fail to account for other features of the data that will be presented here.

1.3.2 Unsupervised learning

The problem of action selection faced by a rat lever-pressing for brain stimulation rewards, re-cast in terms of a MDP, is to find a policy (π^*) to follow in order to maximize the net rate of reward. The policy gives the action to perform in a particular state, for some known transition function that maps the current state to the next state when an action is taken. The rat receives a particular reward, $R(S)$, when it reaches state S, incurs cost $C(S)$, and transitions according to a state-transition function from state S to state S' when action a is taken with probability $\mathcal{T}(S, S', a) = \mathcal{P}[S'|S, a]$. The rat then receives a different reward R and incurs a different cost C for taking action a' in state S', and so on. The rat does not know, *a priori*, the value of each action in each state. As a result, the problem of action selection is compounded by the problem of unsupervised learning. What is learned will determine how an animal selects actions.

When the value of all state-action pairs is known, the rat only needs to identify the policy that will result in the greatest net rate of reward. When the value of all state-action pairs is not known, the rat must learn the values by trial and error. If there is an absorbing state the rat only needs to look ahead through every possible series of states and actions and identify the sequence that will bring about the greatest net reward at the lowest cost by the end of the trial. Although the solution is straightforward, as we will demonstrate, implementing it in neural circuitry with limited resources is intractable and unlikely.

Instead of an animal working for rewards, we can imagine two chess players competing against each other. In this scenario, there is, indeed, an absorbing state: the game will necessarily end in a win, loss, or draw for the two players. Beginning

with some initial state of the world—the pieces in their arrangement at the start of the game—it would be possible, given sufficient time and resources, to enumerate every single sequence of moves that each player can possibly make. The move a player should make given any arrangement of pieces ought to be the one which leads to more wins than losses or draws down the line. The grandmaster-in-training must keep track of not only the rewards that can be reaped from moving a piece to a location (an action) when the board is in a particular configuration (a state), but also the distribution of possible board configurations that will result when it will be their turn once again. As very few board configurations are one move away from a win, loss or draw, the prediction of board configurations must also be projected arbitrarily into the future. Naturally, no human player actually looks arbitrarily far ahead to every possible absorbing state before evaluating which piece to move. Such a procedure would require searching through what are pragmatically infinitely many decision trees of infinite length.

This description of the solution to the action-selection problem is called "model-based," because it relies on the decision-maker to hold in memory a model of the function that maps previous states (like the position of both sides' pieces on the chess board when it will be their turn it is at time t) to subsequent states (like the distribution of piece positions at time $t + 1$) in terms of an action taken at time t (the identity and location of the moved piece). Action selection is trivially solved by searching through each possible sequence, and selecting the sequence of actions (the policy) that will maximize the net rate of rewards. However, the trivial solution may involve a search through a non-trivial state space requiring a non-trivial amount of time to compute.

Rather than searching through a stored decision tree of all possible sequences, thereby selecting that which is optimal, it is possible to store only the value of taking a particular action in a particular state, without representing how states transition

to each other (the state transition function). A mediocre chess player might simply have learned that moving a pawn when the board is in a particular configuration is more desirable than moving the queen, without necessarily knowing why, or what the resulting configuration will be. Since all that is learned is the value of taking an action in a particular state, without any representation of what future states may be, there is no need to engage in a time- and resource-consuming search through an extensive decision tree. In the case of self-stimulating rats, during training, the rat would learn that lever-pressing leads to an electrical reward of magnitude I_{bsr}, after pressing for P_o seconds.

This type of scheme is called "model-free" because all that is kept in memory is a record of the net amount of reward subject to costs, rather than the rat's model of the world instantiated by a state-transition function. While "model-based" reinforcement learning requires the rat to store state-related information (such as reward identity) along with magnitude information, "model-free" reinforcement learning requires the rat to store only the net reward.

For example, one may train a rat that lever pressing will deliver an amount of food pellets giving rise to reward R, having flavour F. The food pellet with flavour F may subsequently be devalued by pairing it with lithium chloride. A rat employing a model-based learning system will learn not only that lever pressing results in R units of reward, but that lever pressing results in the delivery of food pellets with flavour F. If the model-based system has control over the rat's performance after F is paired with illness, the rat will cease to lever press. In essence, the rat is looking ahead to the world state that will result from lever-pressing, and making its decision based not only the reward that was paired with the action of lever pressing, but also on what lever-pressing does to change the state of the world. In contrast, a rat using a completely model-free learning mechanism will only have stored the value of the state that is produced by lever-pressing, and thus will continue to lever-press as it

did before the food reward was paired with illness. Whereas model-free systems are insensitive to changes in underlying motivational states, model-based systems allow the decision-maker to flexibly alter what policy to implement at any given time, at the cost of increased representation.

Neither model-based nor model-free reinforcement learning models can account for all performance in all settings. When a rat has only been moderately trained, its performance following reinforcer devaluation operates as though it had a world model of lever-pressing: namely, the rat behaves as though it knows that pressing the lever will lead to a reward of not only a particular magnitude but also of a particular identity. When that reward is devalued, the rat will cease to lever-press. In contrast, following extensive training, a rat's performance appears to follow a model-free algorithm: the rat behaves as though lever-pressing is associated only with a reward magnitude. While the rat will not consume the devalued reward, it will continue to press the lever because it has not maintained a state-related representation of the identity of the reward that is produced by lever-pressing.

When the model-free reinforcement system is in control of behaviour, the rat still selects actions that maximize its net rate of reward. The critical difference between model-free and model-based systems is that performance can only depend on cached values of the next state when using model-free systems. If all that is learned is that lever-pressing leads to a state with value $V(bsr)$, then un-cued but predictable changes in $V(bsr)$ will have to be learned on-line, as a function of the discrepancy between the old and new values. If the rat also acquires a model of the world, as is the case with model-based reinforcement learning, then un-cued but predictable changes in $V(bsr)$ will not have to be learned. The change in $V(bsr)$ is presumably already extant in the model of the world the rat has learned. What is learned by the rat—be it both reward and state-transition information or reward information alone—will dictate the actions the rat selects following training.

1.3.2.1 Model-based TDRL

In model-based reinforcement learning, the rat acquires a world model that includes more than merely the value of rewards it can expect to receive. More specifically, a rat acting based on a model of the world will act as though it is thinking ahead about how its actions will alter the world within which it lives, rather than just the amount of reward that has been associated with that action. The rat acquires, in some sense, a state-transition function that maps a state S to a future state S' when it takes action a. That state-transition function essentially allows a feed-forward mechanism to inform behaviour.

In the case of electrical rewards, it may be unclear how model-based systems are involved with driving performance. After all, the task is simple, and the reward may not be identifiable. Nonetheless, there is still a state-transition function that can be learned in principle. In traditional two-dimensional parametric procedures, where the reward is delivered following every lever press, the strength of the stimulation changes predictably (usually in descending fashion) from cued trial to cued trial, in a repeating sequence. In the inter-trial interval that precedes every trial, priming stimulation is delivered of identical strength to the stimulation the rat will receive in the upcoming trial if it fulfils the work criterion, accompanied by a flashing house light. Throughout the trial period, the stimulation strength and cost will remain constant. These may all provide cues about what state of the world the rat is in, and the rat may learn that a trial delivering weaker stimulation will follow one with all but negligible strength, and a trial delivering excessively weak stimulation will be followed by one delivering very strong stimulation. Furthermore, the rat may learn that following the flashing house light, the strength and cost of brain stimulation rewards will remain constant. In both of these cases, the rat does not learn exclusively the value of lever-pressing. Instead, it learns a rule about which states are likely to follow others.

Using exclusively model-based systems, the rat either searches through an extensive tree of which states follow which other states when taking particular actions, or else it "fills in" the key determinants of its decision to press and bases its performance on a rule for identifying which state it is currently in and, therefore, which policy is optimal. In any case, the rat must construct a model of how the next trial in a session is related to the current one, or the next trial state is related to the current trial state, and base its policy on that model. Ultimately, model-based reinforcement learning models require the rat to learn $\mathcal{T} = \mathcal{P}[S'|S, a]$ for an arbitrarily large state space, which may include the total net reward it obtained, the action it just took, the running session time, its current position in the sequence of trials, whether there has been an inter-trial interval, and more. On the basis of this model of $\mathcal{T}$, the rat must then implement the policy that will maximize its total net reward, bearing in mind all the attributes of the state space which may or may not be relevant to the task. With a sufficiently extensive model of how one state leads to the next, the rat could (at least in principle) determine the optimal policy from the time it is placed in the operant chamber. Although the horizon is not infinite—with a sufficiently complete model, the rat can implement an optimal policy from beginning to end—the decision space grows exponentially with the size of the state space and the number of possible actions. As a result, it is unlikely that rats use a completely model-based system. A much simpler, though equally imperfect, algorithm for learning which actions to perform in a particular state is model-free temporal difference reinforcement learning. We shall turn to this scheme now.

1.3.2.2 Model-free TDRL

In model-free TDRL, the rat updates its estimate of the reward it receives with successive trials, but does not maintain any information about how one state of the world transitions to the next. Although this formulation would appear to be most

suited to brain stimulation rewards, it also implies that the rat does not keep track of any information above and beyond total net reward, and therefore does not maintain a representation of how key decision variables change throughout the experiment. In its simplest form, the rat has an expectation for reward, pursuant to costs, in any particular state. When the actual value (net reward) differs from what is expected, a discrepancy signal modifies the value of the state that followed the action.

During training, a rat using model-free reinforcement learning systems will try to minimize the discrepancy between the total net reward earned at a point in time and the total net reward it expected for that point in time, called a reward prediction error. Before any training occurs, the rat has no expectancy for reward. When the rat first holds the lever long enough to fulfil the work requirement, the lever is retracted and an electrical reward is delivered. At this point, there is a discrepancy (a prediction error) between the value of the state that is brought about by lever pressing and its expected value. As a result, the optimum $V^*(S)$ is different, because the total discounted sum of future rewards is larger than expected. At time t, the rat expects an optimal total net discounted value of $V^*(S_t)$, but receives $r_t + \gamma V^*(S_{t+1})$ instead. The reward prediction error can be expressed as

$$\delta_t = [r_t + \gamma V^*(S_{t+1})] - V^*(S_t)$$

or the difference between rewards (immediate and predicted to come) actually delivered and those that were predicted to have been delivered (Dayan and Abbott, 2001). The name temporal-difference reinforcement learning comes from the assumption that learning involves an update to the value of actions according to the degree to which the rewards and costs incurred at each time step violate expectation. When the reward prediction error (δ_t) becomes 0, the predicted value of a trial state corresponds to the total net discounted reward that is delivered in that state.

If the rat increments its estimate of the total net reward of this state in proportion to the discrepancy between that expected and that observed (δ_t), the next time around, the rat will behave with an updated estimate of $V^*(S')$. By induction, the discrepancy will appear earlier and earlier in responding to the first moment the reward can be predicted. Neural activity appearing like such a discrepancy signal has been observed in non-human primates (Apicella et al., 1991) and forms the basis of a major computational model of dopamine signalling (Montague et al., 1996).

The degree to which the discrepancy affects a subject's revised estimate of the trial state's expected total net reward is set by a learning rate parameter. If the learning rate parameter is 1, the current estimate is simply replaced by the last obtained reward. In contrast, if the learning rate parameter is 0, the current estimate is never updated. Since the policy depends on the current estimate of the total discounted reward of all states, in the absence of any model of what rewards may be expected from lever-pressing, the animal will need to re-learn that mapping every time the subjective intensity and opportunity cost of the reward changes. If the current estimate is updated rapidly, temporal-difference reinforcement models require the rat to either (A) learn at a very high rate updating over very few time steps, or else (B) to have a model of how the task is designed. The former is not very likely without some process allowing the learning rate to be tuned to the task structure, since performance would be highly dependent on the immediately previous rate of reinforcement on variable interval schedules, an observation that does not obtain (Gallistel et al., 2001, though see Neiman and Loewenstein, 2013). Instead, we propose that rats develop a model of how the operant task is set up allowing them to tune the rate at which the value of lever-pressing is updated—in other words, a model of their world.

1.3.2.3 Average-reward temporal difference model

One solution to the infinite-horizon problem that does not involve exponential discounting is to penalize actions with a long latency. Niv (2008) developed a model in which animals choose both what to do and the latency with which to perform it. When waiting a long time before lever-pressing, the rat forgoes the opportunity to collect rewards for longer. When waiting a very short time, the rat incurs a non-negligible vigour cost for performing the lever-press with a shorter latency. The task for the rat is to maximize its net rate of reward when choosing actions and latencies with which to perform them. Each state is associated with an optimal total net future reward corresponding to the rewards that may be reaped now (subtracting costs), and those that may be reaped when following an optimal policy from now on (subtracting costs), all expressed as a difference from the average rate at which rewards will be delivered when following an optimal policy. If there are only two punctate trial states (lever is up, lever is down), two punctate actions (lever press, groom), and rewards and costs are delivered and levied at punctate state transitions, two recursively-defined equations specify the value of each state. If we presume that the lever can only be down if the action taken is a press, and up if the action is to groom, the two equations become

$$\begin{aligned} V^*(\text{up}) = {} & \text{reward(up)-cost(groom)} \\ & + \mathcal{P}[\text{groom} \mid \text{up}]\ V^*(\text{up}) \\ & + \mathcal{P}[\text{press} \mid \text{up}]\ V^*(\text{down}) \\ & - \text{average reward}^* \ , \end{aligned}$$

and

$$\begin{aligned} V^*(\text{down}) = {} & \text{reward(down)-cost(press)} \\ & + \mathcal{P}[\text{groom} \mid \text{down}]\ V^*(\text{up}) \\ & + \mathcal{P}[\text{press} \mid \text{down}]\ V^*(\text{down}) \end{aligned}$$

$$-\text{average reward}^* \ ,$$

where V^* represents the optimal total net reward and $\mathcal{P}[action|state]$ is the policy. The system can easily be re-written as

$$V^* = r^* - c^* + \mathcal{T}V^* - \bar{R}^*, \text{ or}$$

$$(1 - \mathcal{T})V^* = r^* - c^* - \bar{R}^*,$$

where $\mathcal{T}$ is a state-transition probability, V^* is the total net reward rate when engaging in an optimal policy, r and c are immediate rewards and costs, respectively, and $\bar{R}^*$ is the average reward rate when the policy is optimal.

Since this definition only holds when the policy is optimal, the solution to the system of equations directly provides the optimal policy. As the rows of $\mathcal{T}$ in the above expression must all sum to 1 (the sum of the probabilities of transitioning from a state to every possible state is necessarily 1), the matrix is rank n-1 (in our example, 1) rather than full rank (in the above example, 2). Subtracting the average rate of reward when the policy is optimal allows the system to be solved up to an additive constant (Niv, 2008), thereby solving the infinite horizon problem.

The average-reward formulation described by Niv (2008) proposes that animals choose both an action as well as a latency with which to perform the action. The rat may choose to begin a press more quickly, thereby incurring a heavy cost related to its vigour of responding, but sacrificing fewer rewards, or to begin a press more slowly, thereby incurring a lower vigour cost but sacrificing many rewards from waiting longer. The model penalizes both fast responding as a function of the vigour cost divided by the latency, as well as slow responding as a linear function of the average rate of reward and the latency. In essence, the total net reward obtained in a state S

by following policy π is

$$V_\pi(S) = U_r(S) - C_u(a) - \frac{C_v(a)}{\tau} - \bar{R}\tau + \sum_{S'} \mathcal{P}[S'|S,a]V_\pi(S')$$

where action a and latency τ are given by the policy, states provide unit reward U_r when they are visited, actions incur unit cost C_u each time they are performed and variable cost C_v related to how fast they are performed, and the average rate of reward $\bar{R}$ decreases as the latency to perform actions is increased.

In her original formulation, the rat makes punctate actions—lever pressing, nose poking, or "other"—and obtains food rewards by lever pressing according to a ratio schedule of reinforcement, which can be collected by nose poking into the feeder. Although attractive, the concept of vigour as described here only makes sense when the various activities are punctate and occur following a latency. It is only in this type of task that "vigour" as it is defined makes any sense. When the rat must choose what to do and how long to do it for, rather than how quickly to do it, the costs the rat incurs will necessarily differ from this proposed hyperbolic relationship. The longer the rat chooses to perform an action, the more effort it must put into that action. The trade-off between doing something more quickly and losing reward opportunities disappears. The longer the animal chooses to spend performing an action, the greater both the vigour and opportunity costs.

As a result, we propose an action selection mechanism that differs from the average reward model in three important respects. First, the animal selects what to do and for how long to do it, rather than how soon to do it. An alternate, but equally valid way to say this is that specific activities, which may or may not be directly observable, are chosen by the animal and the animal chooses when those activities terminate. Second, the animal selects actions according to a policy that depends on the scalar, not additive combination, of subjectively mapped key determinants of

decision, which we call the payoff. Third, and perhaps most importantly, the action to be selected is not chosen on the basis of a slowly updating policy. Instead, the chosen action is the result of a rapidly-updating ("filling-in") process that integrates a model of the task demands with internal and external stimuli that directly sets what to do and the probability per unit time that the action will cease.

1.3.3 The matching law

A very different approach to modelling the action-selection problem derives from Herrnstein's Matching Law. The matching law states that the ratio of response rates for two sources of reward will match the ratio of rates at which those two sources deliver rewards. Rather than providing a normative basis for what the animal ought to do in any given circumstance, the Matching Law provides an empirical basis for what animals actually do. The matching law states that relative response rate over an entire trial will match the relative rate of reinforcement on that trial, mathematically instantiated as

$$\frac{R_{experimenter}}{R_{extraneous}} = \frac{Rf_{experimenter}}{Rf_{extraneous}}$$

where R is the response rate and Rf is the rate of reinforcement (de Villiers and Herrnstein, 1976). Assuming there is a constant "amount" of behaviour k to be partitioned between the two, the expression can be reduced to

$$R_{experimenter} = k\frac{Rf_{experimenter}}{Rf_{experimenter} + Rf_{extraneous}}.$$

Although the Matching Law describes what animals do over long periods of time, it does not describe performance on the molecular level. The explanation that Herrnstein gave (Vaughan, 1981; Herrnstein and Prelec, 1991) is melioration, which proposes that animals respond to local rates of reinforcement in choosing which action to perform. For example, suppose that one manipulandum (A) provides rewards at

a rate of one reward per second and the other (B) provides them at a rate of one every two seconds. If the rat were to press 6 times per second on A, the local rate of reinforcement from that option would be 1/6. If it were then to press 6 times per second on the B, the local rate of reinforcement from that option would be 1/12. Since the local rate of reinforcement from option 1 is much higher than from option 2, the rat may then decide to allocate 10 responses per second to A, and 1 per second to B. In this case, the local rate of reinforcement from A would become 1/10, while that from B would become 1/2. As the local rate of reinforcement from A is much lower than from B, the animal will shift responding to B. When the local rates of reinforcement from both alternatives are equal—that is, 10 responses per second to A producing a local rate of 1/10 and 5 responses per second to B producing a local rate of 1/10—the animal also matches the relative rate of responding (10/5) to the relative rate of reinforcement (1/0.5). Melioration theory is not uncontroversial (Green et al., 1983), as matching in this case is also a maximizing strategy, and neither melioration nor maximization can account for molecular level performance with the same normative framework that is afforded by the temporal-difference reinforcement learning models described above.

As we will show, matching emerges from real-time level interactions between the rat and the lever, and though matching is not built on first principles, the only assumptions made in our modelling of real time performance are that (1) the animal bases its decision on payoff, and (2) the stream of holds and releases can be decomposed into multiple types of actions, (3) some of which have payoff-dependent duration.

1.4 Neural substrates

The implementation of action selection in neural machinery involves multiple interacting regions, which have been ascribed various functions depending on the quantitative approach used to model action selection. Below, we describe four structures that have been extensively studied in the context of action selection, from purely model-free, purely model-based, average-reward, and Matching Law-scaling points of view: the medial forebrain bundle, the ventral striatum, the dorsal striatum, and the orbitofrontal cortex. Their contribution to decision-making likely integrates all approaches, as a single approach to the action-selection problem is unlikely to be exclusively correct in modelling how animals partition their time among competing goals. Nonetheless, a parsimonious, yet, comprehensive description of performance in real time could provide a stronger basis for identifying how a particular structure participates in an animal's ongoing decision to engage in one of a variety of behaviours in a particular context.

1.4.1 Medial forebrain bundle

The region this dissertation will most directly investigate is the medial forebrain bundle, as it passes through the location at which the electrode is implanted. The medial forebrain bundle is a large tract of fibres running ventrally from the olfactory tubercle to the tegmental mesencephalon, and neurons with cell bodies in dozens of locations send projections though the bundle both in ascending and descending directions (Nieuwenhuys et al., 1982; Veening et al., 1982). Although the electrode has the potential to induce activation in any subset of this heterogeneous collection of axons running near the tip, those neurons responsible for the rewarding effect likely represent only a small fraction of those coursing past the electrode. Early studies (Wise, 1982) supposed that the rewarding effect was produced by directly ac-

tivating ascending dopamine fibres. The behaviourally-derived characteristics of the neurons involved in the rewarding effect of medial forebrain bundle stimulation are, however, incompatible with known properties of dopamine neurons: the first-stage neurons have short absolute refractory periods (Yeomans and Davis, 1975; Yeomans, 1979); they are likely fine and myelinated and have fast conduction velocities (Shizgal et al., 1980; Bielajew and Shizgal, 1982); and at least a subset of them project in the anterior-posterior direction (Bielajew and Shizgal, 1986). Although VTA dopamine neurons are activated by the electrical stimulation (Hernandez et al., 2006), it is highly unlikely that dopamine is itself the cable along which reward-producing stimulation has its effect.

Similarly, computational models of the activity of dopamine cells (Montague et al., 1996), whose activity as measured in electrophysiological recordings appears to track the discrepancy between expected and obtained rewards, assume that the electrode is a dopa-trode: electrical stimulation causes dopamine neurons to fire, dopamine neurons adjust the synaptic weights involved in assigning value to stimuli like the lever, and the self-stimulating rat engages in lever-pressing because of the increased value of the lever. Since the discrepancy between expected and obtained rewards is the proximal cause of electrical stimulation, these synaptic weights either grow to infinity or saturation. Thus, the rat would always come to expect that lever-pressing will provide a maximal reward. This view is also incompatible with empirical findings: rats will adjust performance to sub-maximally rewarding brain stimulation (Hodos and Valenstein, 1962) and can respond to a wide range of stimulation strengths at differing rates (Hodos and Valenstein, 1960).

Instead, electrical stimulation induces a volley of action potentials that appears to be spatio-temporally integrated by a network, or system of networks, whose peak activity is encoded somewhere in a memory engram of the subjective intensity of the rewarding effect (Gallistel et al., 1974). The growth of this intensity with stim-

ulation strength is well-described by a power function at low pulse frequencies that saturates at high pulse frequencies (Simmons and Gallistel, 1994), and can thus be reasonably approximated by a logistic function. We propose that the rat maintains this key subjective decision variable, as well as other subjective decision variables like the opportunities foregone, the effort expended, and the risk involved, in memory, and they are subsequently combined multiplicatively, as suggested by Baum (1979) and later by Leon and Gallistel (1998). The scalar combination, termed "payoff" in the remainder of this thesis, is the organizing principle by which actions are selected. Factors that may be deemed more "cognitive," like statistical regularities inherent in the operant task structure, also have representations which may inform and potentially very rapidly update the payoff that can be expected from engaging in a variety of actions. Below, we briefly describe three regions that have been heavily implicated in the process of action selection, and briefly describe how they may fit into this organization.

1.4.2 Orbitofrontal cortex and task modelling

One region that has garnered interest in the study of decision-making is the orbitofrontal cortex (OFC). Activity within the OFC has been correlated with flavour perception of pleasant foods (Kringelbach et al., 2003; Rolls, 2005), delay (Roesch and Olson, 2005), risk (Kepecs et al., 2008) and cost (Kennerley et al., 2009), and responses within the region are highly heterogeneous. Lesions of orbitofrontal cortex impair discriminations between outcomes that differ with respect to their identity but not their value, but abolish the facilitation in learning a cue-response pairing when the outcomes of the two responses differ by their value (McDannald et al., 2011). These results suggest a general mechanism is at work in the orbitofrontal cortex that provides a mapping of task-related outcome identity information (what, when, how likely, how hard, under which circumstances) to other stimuli, responses, and goals.

As we shall discuss later, the process is similar to what is meant by a world model, which establishes how stimuli both internal and external to an agent are related to each other, if indeed they are.

1.4.3 Ventral striatum and value updating

Another region that has received considerable attention is the ventral striatum, a major component of the basal ganglia system. The ventral striatum receives afferent connections from cortical structures, including the orbitofrontal cortex (Eblen and Graybiel, 1995) as well as subcortical structures like the hippocampus (Kelley and Domesick, 1982), amygdala (Kelley et al., 1982) and ventral tegmental area (Ikemoto, 2010). Activity in sub-populations of ventral striatum cells correlates with reward receipt (Apicella et al., 1991), some show anticipatory responses (van der Meer and Redish, 2009), and others fire prior to specific cued actions but not the nature of the cues that preceded them (Taha et al., 2007). These results would imply that the ventral striatum is involved in some way with the mapping between actions and the degree to which these actions ought to be chosen. As will be argued later, it is possible that neurons within the ventral striatum integrate task-specific information regarding the set of world models the rat may have acquired in training with specific and non-specific outcomes so that the mapping between actions and their desirability may be maintained.

1.4.4 Dorsal striatum and response-outcome gating

In contrast to the ventral striatum, in which activity is tied to updating the mapping between actions and their putative payoffs in exploratory phases of a task (van der Meer and Redish, 2009), the dorsal striatum may maintain this mapping for as long as necessary. Ensemble recordings of the dorsal striatum show an increase in the coding efficiency of units during decision-rich portions of a spatial naviga-

tion task as the correct path is learned (van der Meer et al., 2010). Human functional neuroimaging studies have found differential activation in the dorsal striatum in response to high-calorie visual food stimuli when presented to obese individuals (Rothemund et al., 2007). Lesions of the dorsal striatum disrupt the acquisition of a dual-operant task when each response alternative provides a different (sucrose or food pellet) reward, as well as subjects' sensitivity to devaluation of the outcomes (Yin et al., 2005). Post-training lesions also impair performance and render animals' lever-pressing insensitive to devaluation of the outcomes that each response bring about. These results appear to imply that the dorsal striatum is sufficient for an accurate representation of the mapping between actions and outcomes, and necessary for acquiring and maintaining that mapping. It is possible that while ventral striatum is involved in integrating task-specific information with current payoff estimates to update the desirability of an action, the dorsal striatum is involved in maintaining the map.

1.5 This thesis

1.5.1 A computational molar model

In this thesis, we will elaborate on a molar model of performance for brain stimulation rewards based on a modification of Herrnstein's Matching Law, termed the Shizgal Reinforcement Mountain Model. In essence, brain stimulation delivered to the medial forebrain bundle elicits a volley of action potentials that travel caudally to a downstream network that performs a spatio-temporal integration of the signal. Peak activity is then recorded in an engram of subjective reward intensity, the growth of which can be fairly well approximated by a power function at low pulse frequencies that rises to asymptote at high pulse frequencies (Simmons and Gallistel, 1994). The engram provides the information required to direct future behaviour, and this

representation is combined with the subjective impact of other key determinants of decision like the cumulative amount of time the lever must be depressed (the price), the force required to depress the lever, and the probability that a reward is delivered following successful completion of the work criterion. The scalar combination of these quantities we term the "payoff," and it will be a central organizing principle of this thesis. It is on the basis of the payoff that the rat allocates its time between self-stimulation and non-self stimulation activities. At the molar level, we have modelled the dependence of time allocation to self-stimulation activities on payoff as the ratio of suitably transformed payoff from self-stimulation to the sum of suitably transformed payoffs from self-stimulation and non-self stimulation activities. Manipulations that affect reward circuitry prior to the peak detection stage, and thus before an engram can be recorded, will alter the sensitivity of the psychophysical translation of pulse frequency into subjective reward intensity. That is, any interference with information processing before a reward intensity can be recorded will change the ability of stimulation pulses to drive the integration network to a particular relative level of reward intensity. Interference with information processing upstream from or at the output of the spatio-temporal integration network will ultimately alter the absolute scaling of the reward intensity without affecting its sensitivity to inputs. Since the relevant decision variable is presumed to be a scalar combination of the arbitrarily scaled intensity with the subjective impact of the price, and given that the arbitrary scaling can be simply set to 1, changes beyond the output of the peak detection stage will result in changes to the animal's inferred responsiveness to the price. When we define a criterion price at which the payoff from a maximal reward will only produce half-maximal time allocation, manipulations occurring at or beyond the peak detection stage will result in changes in this criterion, whereas manipulations prior to the peak detection stage will result in changes in the pulse frequency that produces a half maximal reward.

A distinction can thus be made between the sensitivity of the reward-growth function, mapping stimulation strength into subjective reward intensity, and its gain. When reward growth is made more sensitive to changes in stimulation strength without altering the gain, weaker stimulation will have a greater subjective impact, but strong stimulation will be unaffected. When the gain of reward growth is increased, all stimulation strengths are scaled. Similar processes occur in sensory-perceptual systems; adaptation to the darkness of a movie theatre alters the visual system's sensitivity to light, such that even a few photons will be perceptible. As one leaves the movie theatre, one is blinded because even moderate light is perceived at maximum. An increase in gain would, in contrast, make all lights, bright and dim–appear brighter. The mountain model can similarly distinguish changes in the sensitivity of the reward substrate, which affect only the relative impact of rewards, from changes in gain, which affect only the absolute impact of rewards.

The model, in its current state, has been previously validated (Breton et al., 2013; Arvanitogiannis and Shizgal, 2008) with respect to its ability to correctly detect the effect of a manipulation affecting the substrate for self-stimulation, and has proved useful in reinterpreting the effects on self-stimulation produced by alterations in dopaminergic transmission (Hernandez et al., 2010). A previous experiment validated the strong positive prediction that manipulation of the directly stimulated neurons (increasing the duration for which they are stimulated) should affect the sensitivity of the psychophysical process translating pulse frequency into subjective reward intensity (Breton et al., 2013). However, in a subset of animals, the weaker negative prediction that this manipulation should not affect the gain or absolute scaling of the psychophysical process was not supported. One proposed mechanism for the observed changes in gain was that stimulation in these animals provoked activity in multiple integrators with different strength-duration trade-off functions. Decreases in train duration would then displace the strength-duration relationship in one of

these integrators sufficiently that it no longer contributes to the stored record of reward. The hypothesis relies on the assumption that, indeed, changes in the subjective intensity of a reward are dissociable and orthogonal from changes in the subjective opportunity cost. Similarly, alterations in dopaminergic neurotransmission, either by cocaine (Hernandez et al., 2010) or GBR-12909 (Hernandez et al., 2012), produced reliable changes in gain with unreliable or trivial changes in sensitivity as assessed by the mountain model. These previous studies have assumed that changes in gain (P_e) are orthogonal to and experimentally dissociable from changes in sensitivity (F_{hm}). This thesis will demonstrate that manipulations of the payoff from self-stimulation that do not alter the post-synaptic impact of the stimulation can indeed be correctly identified using the Shizgal Mountain model.

In the chapters that follow, we shall first validate the molar model's capacity to identify the stage of processing at which risk acts. Then, we shall use the estimates of subjective reward intensity and opportunity cost inferred in fitting this molar model to describe some of the supra-molar strategies that rats may use in our procedures and the molecular processes by which actions are selected in real time.

1.5.2 World models

Although a well-trained rat's performance on any given trial can be modelled in the aggregate by the Shizgal Reinforcement Mountain, it could be even more useful to describe performance on the molecular level. Trial performance reduces the entire stream of holds and releases to a single number, and the Shizgal Reinforcement Mountain uses trial performance to identify the stage of processing at which a manipulation has occurred. Since our experimental protocol follows a definite structure, it is possible that the rat develops a model of its own—a world model—of the statistical regularities it encounters in the course of testing.

World models allow the rat to quickly learn which actions are optimal in a

variety of settings. For example, if it learns that the pulse frequency and price of trains of brain stimulation rewards will remain constant throughout a trial, the rat only needs a single exemplar of price and frequency to tune its optimal policy of pressing and releasing. If it learns that following a trial of very low pulse frequencies it will be presented with a trial with a very high payoff, it can begin working as soon as the high payoff trial has begun to collect rewards at as fast a rate as possible. If it learns that following a trial with intermediate payoff, it will be presented with a trial with very low payoff, then it can forgo pressing entirely.

Chapters 3 and 4 will explore the question of whether rats construct a world model of the triad structure of the session, and whether they construct a world model of the stability of the trials.

1.5.3 A computational molecular model

If rats can be said to have a session model of triad structure and a trial model of trial structure, it should be possible to describe what animals do in real time once the subjective intensity and price of the electrical reward are known. This description, which I call a computational molecular model of performance, provides a means of identifying what actions the animal takes in real time. Such a model makes possible a much more fine-tuned analysis of how various manipulations affect performance in real time. For example, the computational molecular model makes possible the investigation of whether activity in various brain regions is related to the rat's varied behavioural states, some of which may not be directly observable even in principle. It also allows for a detailed description of what lesions and pharmacological interventions do to the propensity of the rat to enter various behavioural states. Indeed, a real-time, molecular model of choice is indispensable for experiments in which millisecond-scale physical measurements—electrophysiological recordings, electrochemical assays, and optogenetic manipulations—are obtained.

Chapter 5 concerns this descriptive model. The computational molecular model of performance assumes that the various observable activities the rat is in—pausing after reinforcement, holding, releasing for a short while, releasing for a long while—are the result of a mixture of underlying behavioural states with characteristic properties. The hidden behavioural states produce stay durations in each activity according to characteristic gamma distributions, whose mean depends only on the payoff for the trial. Following this descriptive model, we compare the proportion of time allocated to holds or short (tapping-related) releases predicted by the computational molecular model to the molar prediction of the Shizgal Reinforcement Mountain. If the two are in good agreement, then the molar performance described in the Shizgal Reinforcement Mountain Model is an emergent property of molecular model of real time performance.

Chapter 2

A computational molar model of performance for brain stimulation reward

2.1 Introduction

For many animals, the longer one forages, the longer one leaves oneself open to predation. Often, if an animal fails to find food, it will starve. The longer it forages, the less opportunity it will have to copulate, find water, or hide. Pulled by the various goals that would be advantageous for them to pursue, successful animals nonetheless efficiently balance competing objectives. Indeed, for any species to be successful, it must successfully trade off costs and benefits inherent to one goal with the opportunities and risks from goals it has foregone.

Instrumental responding is no different from naturalistic action selection in this respect. Experimental animals must arbitrate among competing goals of experimenter-delivered rewards and those not under experimenter control. An animal that is given the opportunity to harvest brain stimulation rewards will have to balance pursuit of electrical pulses with the rewards that accompany grooming, exploring, and resting. Below, we develop a model of performance for these rewards—the Mountain Model—and provide a reasoned basis for evaluating how performance is changed when variables are manipulated. Indeed, operant responding is the result of a series of trade-offs, translations and combinations. As a result, the tools needed to ascertain how a manipulation has altered responding and what stage of processing has been affected need to take these multiple transformations into account.

2.1.1 The Shizgal Mountain Model

Performance for rewarding brain stimulation is not simply a reflection of its subjective impact. Multiple factors affect whether or not an animal will invest time and energy in obtaining any reward, and brain stimulation is no different in this regard. The amount of time an animal will invest in holding down a lever will depend on the intensity of the rewarding effect produced by the stimulation, the opportunities forgone by holding the lever, and the energetic demands required to maintain its depression. So long as the stimulation is strong, the opportunities forgone are few, and the energetic demands are negligible, the rat will devote almost all of its time to lever-pressing. When the stimulation weakens, the opportunity costs grow, or the energetic demands place an increasing burden, the rat will devote less time in pursuit of the reward. This conjecture implies that the central decision variable for the rat in pursuit of brain stimulation rewards, the single criterion to be traded off, is the payoff derived from self-stimulation activities compared to the payoff derived from all other activities available to the rat in the operant chamber.

When a rat harvests a train of brain stimulation reward, the cathodal pulses induce activity in the fibres of passage and local somata surrounding the electrode tip, injecting a volley of action potentials in, among others, the first-stage neurons composing the substrate within the medial forebrain bundle responsible for the rewarding effect. The injected signal is characterized by an aggregate rate code. The rat behaves as if a small number of fibres (low current stimulation) firing at a high rate (high frequency stimulation) produces a reward of equal magnitude to a larger number of fibres (high current stimulation) firing at a low rate (low frequency stimulation). These findings (Gallistel, 1978) imply that a process, occurring downstream from the first-stage neurons, effectively performs an integration over time and space. The peak activity in this integrator network is committed to memory (Gallistel et al.,

1974). In parallel, a process must have committed to memory the average amount of time invested in obtaining the reward, as well as the average amount of effort expended, in order for these variables to affect future behaviour. These stored subjective determinants of performance are combined, presumably in scalar fashion, for the rat to arrive at a subjective estimate of the payoff that self-stimulation will offer him. A similar process presumably occurs in evaluating the payoff from all other activities: the stored record of the rewards that can be derived from grooming and resting, for example, are combined with the effort and opportunity costs of performing them to provide the self-stimulating rat with the payoff it can expect to receive from these extraneous activities. Action selection is simply the process of allocating time between pursuit of competing goals such that the overall payoff derived from all activities is maximal.

Figure 2.1 shows the presumed sequence of events leading up to a decision regarding the allocation of time to work (self-stimulation) and leisure (non-self stimulation) activities. First, the electrode induces a volley of action potentials in the substrate within which it is embedded. This volley of action potentials travels down the fibres to one or more networks that perform a spatio-temporal integration. The peak activity of this volley of action potentials is identified by a peak detector and stored in a memory engram that represents the subjective intensity of the rewarding effect of the brain stimulation. In parallel, the rat revises its estimate of the opportunity cost of acquiring trains of brain stimulation by considering the average amount of time it spent holding the lever to earn the reward (the price, P) and the opportunities thus foregone. In addition, the rat revises its estimate of the energetic requirements (ξ) of acquiring brain stimulation trains and the probability that it will be rewarded (P_r). The stored records of these subjective determinants of performance are combined multiplicatively, updating the rat's estimate of the payoff that lever-pressing will provide. The rat apportions its time to work and leisure activities on the basis

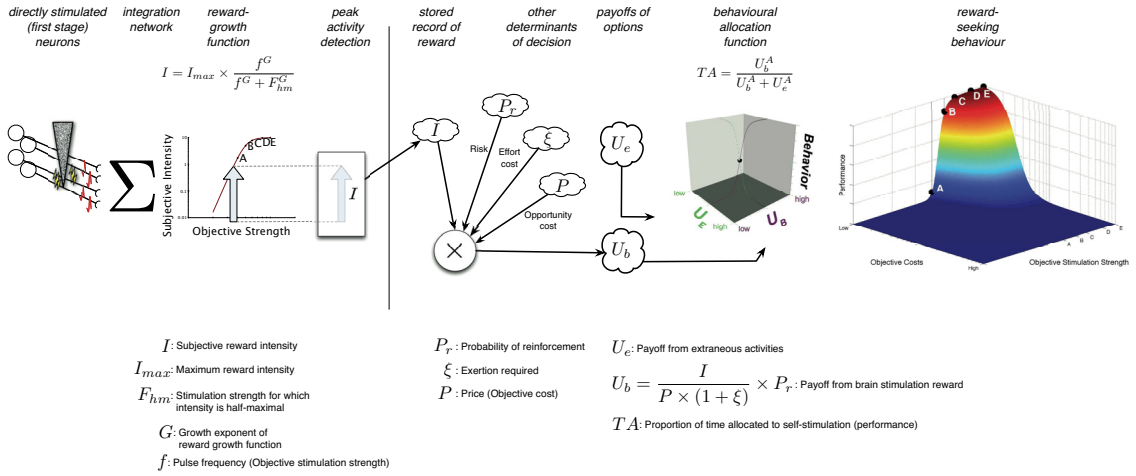

Figure 2.1. Sequence of events in the decision to press. When a rat harvests a brain stimulation reward, the stimulation induces a volley of action potentials that travel to a network that integrates the injected signal over space and time (Σ). This integration results in a subjective reward intensity (I); the peak activity of this subjective intensity signal is committed to memory in an engram. In parallel, the probability of reinforcement (P_r), the amount of time required (P), and the effort invested in acquiring rewards (ξ) is also determined, turned into subjective variables (risk, opportunity cost, and effort cost) and committed to memory. Their scalar combination provides the rat with the payoff it can expect from self-stimulation activities (U_b). A comparison of the payoff from self-stimulation with the payoff the rat expects from all other activities it can perform in the operant box (U_e) provides the rat with the proportion of time it will spend engaged in self stimulation-related activities (TA), which will drive performance for rewards.

of the subjective payoffs it can expect to derive from them.

This model requires that processes implement the psychophysical transformation of objective experimental variables, like stimulation strength, price, reward probability, and exertion, into subjective determinants of choice, like subjective reward intensity, subjective opportunity cost, subjective probability, and subjective effort cost. The psychophysical function that describes the translation of stimulation strength into subjective reward intensity, or the reward-growth function, has been well described (Simmons and Gallistel, 1994; Hernandez et al., 2010). The reward growth function can be reasonably well approximated by a logistic function whose value grows as a power function at low strengths (pulse frequencies) and saturates at high strengths. This function is a critical component of the model, as it prescribes a nonlinearity in a processing stream that otherwise contains multiple scalar combinations which cannot, in principle, be distinguished. The reward-growth function establishes that there is some pulse frequency that will produce a maximal reward (assuming the duration and current remain constant), beyond which higher frequencies will fail to raise the subjective intensity of the rewarding effect.

Figure 2.2 shows the sequence of events in the decision to press, when focusing on the reward growth, payoff-computation, and behavioural allocation functions. Suppose there are five pulse frequencies, A, B, C, D, and E, ordered from lowest to highest. A rat may respond sub-maximally for rewards of pulse frequencies A and B, while responding at a maximal rate for rewards of pulse frequencies C, D and E (extreme right-hand side of figure 2.2). Given a choice between pulse frequency A and B, the rat prefers to respond for B, and when given a choice between pulse frequencies B and C, the rat prefers to respond for C. This is relatively unsurprising, because the rat's single-operant performance demonstrates that the rat responds more to C, D, and E than B, and the rat responds more to B than A. However, even though the rat responds at the same, maximal rate to pulse frequencies C, D, and E, it may have a

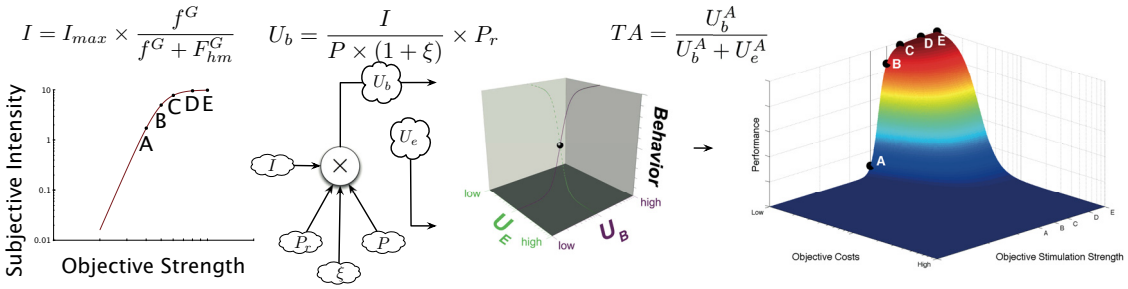

Figure 2.2. Simplified sequence, focusing on reward growth and behavioural allocation functions. The subjective intensity of the stimulation (I) is a logistic function of pulse frequency (f); peak activity in the network that implements this psychophysical transformation is committed to memory. Consequently, while pulse frequencies A through D elicit different remembered reward intensities, pulse frequencies D and E elicit the same, maximal reward intensity. Intensity is then combined with probability (P_r), opportunity cost (P), and effort cost (ξ) multiplicatively to provide the animal with a payoff from self-stimulation activities (U_b). Performance itself is a non-linear function of the payoffs from self-stimulation activities (U_b) and non-self stimulation activities (U_e). As a result of this composition of functions ($TA = b(g(x, y))$, where b and g are the non-linear behavioural-allocation and reward-growth functions, respectively), even though the rat allocates equal proportions of time to acquiring stimulation trains of pulse frequencies C, D, E, their subjective intensities are not equal. In effect, the performance level motivated by any given pulse frequency is not a reflection of its underlying reward intensity alone.

preference, measured in a dual-operant setting, of D to C, and no preference between pulse frequencies D and E. This implies that the subjective intensity of the rewarding effect is both non-linear (since it saturates at high pulse frequencies) and convolved with a process that translates the intensity into performance (since stimulation producing equal response rates are not necessarily equi-preferred). Because the rat's preference (or lack thereof) between two trains of different pulse frequencies is governed by the subjective intensity that they produce, pulse frequencies D and E may produce subjectively equal reward intensities, while pulse frequencies C and D may produce different subjective reward intensities. Furthermore, the rat's performance for pulse frequencies C and D in a single-operant setting is not a direct reflection of their subjective intensity, but rather, a functional composition of the reward growth and behavioural allocation functions, both of which are non-linear.

Unlike linear functions, for which location (position along the abscissa) is confounded with scale (position along the ordinate), a non-linear function allows us to interpret the effect of a manipulation very clearly. In the case of the reward-growth function, a change in the location of the logistic is identical to a change in the relative subjective impact of stimulation strengths, such that weaker stimulation produces a more intense rewarding effect or stronger stimulation produces a less intense reward. A maximum reward, however, remains at a constant absolute value; it is simply the relative impact of each additional stimulation pulse that is altered. A change in the scale of the reward-growth function is identical to a change in the absolute impact of stimulation strengths, such that the intensity of all stimulation trains is multiplied by a constant. The relative impact of a particular train will remain constant, even though its absolute value will be scaled up or down.

At constant train duration and current, the pulse frequency whose subjective impact is half-maximal (F_{hm}) is an index of the relative impact of stimulation trains. When injected action potentials are more effective at producing a level of reward rela-

tive to maximum—that is, the reward system is highly *sensitive*—the pulse frequency that produces a half-maximal reward will be lowered. As such, a manipulation that alters F_{hm} will act prior to the output of the process that identifies the peak activity in the substrate and commits the subjective intensity of the rewarding effect to memory. It is therefore possible to identify a manipulation of the first-stage neurons themselves, like prolonging the duration of the train or increasing the current of the stimulation pulses (Breton et al., 2013; Arvanitogiannis and Shizgal, 2008). In contrast, because the output of the reward growth function, implemented by the peak activity of a spatio-temporal integration network, is committed to memory and combined in scalar fashion with the other determinants of the decision, any alteration of the neural machinery of choice beyond the output of the peak detector will scale the reward growth function along the ordinate.

The payoff from brain stimulation reward is a scalar combination of the reward growth function with psychophysical translations of required work time (P), required effort (ξ), and probability of reinforcement (P_r). The resulting payoff (U_b) is compared to the payoff from everything else the rat could do in the operant chamber (U_e), such as grooming, resting, and exploring. A common means of describing the translation of payoff into performance, and the means we have adopted here, is to adapt Herrnstein's Matching Law (Herrnstein, 1974) to the single operant context (McDowell, 2005; de Villiers and Herrnstein, 1976; Hamilton and Stellar, 1985, see) using a suitable exponent (A) to take into account the partial substitutability of the fruits of work with those of leisure. In this case, the payoff from everything else is equal to the payoff from brain stimulation reward when the rat spends equal time engaged in each; as a result, the proportion of time allocated to self-stimulation will be 0.5 when work and leisure are equi-preferred. If we define a price, P_e, at which a maximal reward (I_{max}) produces half maximal time allocation, the payoff derived from leisure activities will

be

$$TA = \frac{U_b^A}{U_b^A + U_e^A}$$

$$TA = \frac{(\frac{I_{max}}{P(1+\xi)} P_r)^A}{(\frac{I_{max}}{P(1+\xi)} P_r)^A + U_e^A}$$

$$0.5 = \frac{U_b^A}{U_b^A + U_e^A}$$

$$U_b = U_e \text{ at } TA = 0.5$$

$$U_e = \frac{I_{max}}{P_e(1+\xi)} P_r$$

where U_b is the payoff from brain stimulation rewards, U_e is the payoff from extraneous ("leisure") activities, I_{max} is the maximum subjective intensity of the rewarding effect, P is the opportunity cost ("price") of the reward, ξ is the effort cost ("exertion") of the reward, and P_r is the probability of reinforcement.

Re-arranging terms, the price at which a maximal reward produces half-maximal time allocation, P_e, is the following function of the maximal intensity of the reward, the effort cost of the reward, the probability of reward, and the payoff from everything else:

$$P_e = \frac{I_{max}}{U_e(1+\xi)} P_r$$

Substituting the equation of the logistic reward growth function for the intensity of the rewarding effect, and expanding the payoffs,

$$TA = \frac{(I_{max} \frac{f^G}{f^G + F_{hm}^G} \times \frac{P_r}{(1+\xi)P})^A}{(I_{max} \frac{f^G}{f^G + F_{hm}^G} \times \frac{P_r}{(1+\xi)P})^A + (I_{max} \times \frac{P_r}{(1+\xi)P_e})^A}.$$

Simplifying terms, we obtain the following prediction of time allocation as a function

of pulse frequency and price.

$$TA = \frac{(\frac{f^G}{f^G + F_{hm}^G})^A}{(\frac{f^G}{f^G + F_{hm}^G})^A + (\frac{p}{P_e})^A}$$

As stated above, a manipulation that affects the circuitry that underlies decision making at any point between the electrical stimulation and the peak detection process will result in a change in the pulse frequency that produces a half-maximal reward, F_{hm}. A manipulation that affects decision making circuitry beyond the output of the peak detector—one that alters I_{max}, U_e, ξ, or P_r—will result in a change in the price at which a maximal reward drives only half-maximal performance, P_e. Thanks to the non-linearity inherent in the reward growth function, it is possible to identify whether or not a manipulation has occurred beyond the peak detection process. Because of the scalar combination of the subjective determinants of choice, multiple manipulations will similarly and indistinguishably affect the resulting 3D time allocation function (Hernandez et al., 2010).

Previous validation studies of this mountain model have established that a manipulation known to affect the first-stage neurons directly can indeed be correctly identified. A different version of the model, in which the relevant experimental variables were the rate of reinforcement and number of pulses injected, correctly identified alterations in train duration and pulse current as manipulations affecting first-stage neurons directly (Arvanitogiannis and Shizgal, 2008). However, Conover *et al.* (2001a) found that, at least in the case of brain stimulation reward, the assumption that operant tempo was independent of rate of reinforcement did not hold, complicating interpretations of the Arvanitogiannis *et al.* (2008) results that were obtained under an infinite-hold variable-interval schedule of reinforcement. Furthermore, the data in that experiment were obtained by systematically varying a single independent variable (pulse frequency or programmed rate of reinforcement) while holding

the other constant. Breton *et al.* (2009) found that this experimental design distorted performance, by reducing the evaluability of price, which requires multiple exemplars for the rat to gain an accurate estimate of the opportunity cost of brain stimulation rewards In a separate validation experiment, Breton et al. (2013) used a cumulative handling time schedule which provides a tighter control of the rate of reinforcement. Breton et al. (2013) delivered stimulation trains 1.0s and 0.25s in duration, and observed differences in F_{hm} from one duration condition to the next. Additionally, Breton et al. (2013) used the experimental design suggested by Breton et al. (2009), which dramatically increases the evaluability of relevant independent variables by presenting pulse frequency-price pairs in random order alongside high payoff (low price, high pulse frequency) and low payoff (low price, low pulse frequency) trials that provide anchors for evaluating the payoff on randomly selected test trials. These validation studies establish that the mountain model, in its many incarnations and across multiple performance-probing procedures, is capable of correctly identifying a manipulation affecting the reward circuitry prior to the output of the peak detection process.

Fewer studies have been conducted, however, explicitly altering the payoff of brain stimulation reward without affecting the directly stimulated substrate. Arvanitogiannis (1997) altered the payoff derived from leisure activities by delivering sporadic leisure-contingent stimulation; although this produced changes in Re, the rate of reinforcement that drove performance for a maximal reward to 50% time allocation (related indirectly to the reciprocal of P_e), the effects were messy. Providing background stimulation during leisure activities would indeed increase the payoff from leisure-related sources of reward, but also drastically increases the degree to which leisure-derived and work-derived rewards can be substituted for one another. Moreover, such a manipulation did not alter the payoff from brain stimulation reward directly; rather, it alters performance for a given payoff without altering the payoff

itself. To determine whether the mountain model can, indeed, correctly detect an orthogonal change occurring beyond the output of the peak detector that also alters the payoff from brain stimulation reward, we compared performance for certain rewards to that for risky rewards. The probability of reinforcement is expected to alter the rat's estimate of the opportunity cost for rewards, in the sense that it will have to devote more time for every reward delivery when successful completion of the work requirements (the price) does not always lead to a train of electrical brain stimulation.

In addition to validating the ability of the mountain model to correctly identify post-peak detection manipulations, the degree to which a particular probability changes the payoff from self-stimulation can be quantified. If P_{e_1} is the price at which a maximum, certain reward produces a payoff from self-stimulation that is equal to the payoff from everything else, and P_{e_2} is the price at which a maximum, risky reward produces a payoff from self-stimulation that is equal to the payoff from everything else, then it follows that the ratio of the two is the factor by which the rat discounts a risky reward compared to a certain one. For example, if the ratio between a maximal reward delivered with probability 0.75 and a maximal reward delivered with certainty (probability 1.00) is 0.75, then the rat discounts two identical (maximal) rewards by a factor of 0.75. If that ratio is 0.7, then the risky reward is relatively *under*-weighted; if that ratio is 0.8, the risky reward is relatively *over*-weighted, as compared to the normative probability. As a result, in addition to testing whether the mountain model could correctly identify a post-peak detection effect, we set out to quantify the probability discounting function for reward probabilities ranging from 0.5 to 1.0.

2.1.2 Probability discounting

2.1.2.1 Previous studies

The study of how risk impacts the selection of actions that lead to probabilistic outcomes has been of considerable interest to economists and neuroscientists alike. In the classical economic analysis of decision-making under uncertainty, the agent weighs the gains by their probability of occurrence to extract an expected value for either option, and performs the action which will lead to the greatest expected value. Fundamentally, the process that drives the decision is arbitrary and the ultimate outcome is of utmost importance. In contrast, the typical neuroscientist takes for granted the assumption that agents maximize this expected value, and studies correlates of this expected value to identify where economic determinants of decision-making are represented (van Duuren et al., 2008, 2009; Pennartz et al., 2011).

The idea that humans maximize objective, probability-weighted objective outcomes is not without its detractors. Kahneman and Tversky (Kahneman and Tversky, 1979) proposed that losses loom larger than gains (that is, the objective loss $-x$ is much more undesirable than the objective gain of x is desirable). For example, individuals will have a strong aversion to an option that pits a gain with 50% probability against an equally likely loss of the same amount of money, and of two gambles with symmetric outcomes (a gain of x with probability 0.50 or a loss of x with a probability of 0.50), people will choose the one in which the expected gain/loss is smallest. Furthermore, Kahneman and Tversky propose that while outcomes with very low probabilities (0.001 and 0.002) carry more-than-normative weight in a stated option, non-certain (non-zero and non-unitary) probabilities carry less-than-normative weight. Individuals overwhelmingly (73%) prefer a lottery where the probability of winning 6000 Israeli pounds with probability 0.001 (and the complementary probability of nothing) to one where the probability of winning 3000 Israeli pounds is

0.002, implying that the ratio of probabilities 0.001/0.002 is greater than 1/2—the normative value (Kahneman and Tversky, 1979). Moreover, individuals prefer (72%) a 0.001 probability of winning 5000 Israeli pounds to the objective expected value of the ticket (5 pounds) for certain (Kahneman and Tversky, 1979).

Despite overweighting small probabilities, individuals prefer (82%) 2400 pounds with certainty to a gamble where they can win 2500 pounds with probability 0.33, 2400 pounds with probability 0.66, or nothing with probability 0.01. This preference implies that the utility of 2400 pounds is more desirable than the utility of 2500 weighted by 0.33 and summed with the utility of 2400 weighted by 0.66, or

$$u(2400) > 0.33u(2500) + 0.66u(2400).$$

Subtracting the final term (0.66 times the utility of 2400 pounds) from both sides, we obtain a preference of

$$0.34u(2400) > 0.33u(2500).$$

In other words, the preferences imply that 2400 Israeli pounds with probability 0.34 is better than 2500 pounds with probability 0.33.

In a separate question, respondents had to choose between 2500 Israeli pounds with probability 0.33 (or 0 with probability 0.67) and 2400 pounds with probability 0.34. A large majority of individuals (83%) opted for the first choice, which is opposite what would be expected from above. The two preferences imply that the sum of the weight associated with a probability of 0.66 with that associated with a probability of 0.34 is less than the sum of their objective probabilities, 1. As a result, for probabilities that are neither a few tenths of a percent nor trivial (0 and 1), the impact of a probabilistic outcome on the subjective value of the option for which it is a part is under-weighted compared to its normative value. In the decades since Kahneman and

Tversky (1979) proposed their prospect theory, which describes choice as a result of a concave function of gains, a convex function of losses, and a non-linear function of probability that is discontinuous at 0 and 1, many other researchers have probed the degree to which risky options are devalued (Rachlin et al., 1991; Gonzalez and Wu, 1999). Individuals consistently violate the assumptions of expected utility theory, according to which equal outcomes for pairs of gambles cancel each other out.

There is a relative paucity of research on probability discounting in non-human animals. Studies of the degree to which rewards become less interesting when they are made probabilistic do not usually perform the adequate psychophysics. MacDonald *et al.* (1991) tested the proposition that rats would exhibit a similar preference reversal as demonstrated by Allais (Allais, 1953) in humans. Suppose a decision-maker is presented with the following choice: A) Y dollars with probability p or X dollars with probability (1-p), or B) Z dollars with probability q or X dollars with probability (1-q). If the individual prefers A over B, we can assume that

$$pY + (1-p)X > qZ + (1-q)X,$$

so we can also assume, multiplying both sides by a constant, that

$$rpY + (1-rp)X > rqZ + (1-rq)X.$$

If the decision-maker is presented with choice C) Y dollars with probability rp or X dollars with probability (1-rp), or D) Z dollars with probability rq or X dollars with probability (1-rq), the preference is reversed for small values of r.

This effect, called the common ratio effect (because in both gambles, the ratio of probabilities of Y to Z is the same) has been observed in many cases, and was investigated in water-deprived rats making choices between levers that delivered either A) 8 cups of water with probability 1, or B) 13 cups with probability 0.75 or 1 cup with

probability 0.25. Following this preference test, rats were given the choice between two different outcomes that maintain a ratio common to the first: C) 8 cups of water with probability 1/3 or 1 cup with probability 2/3, or D) 13 cups with probability 1/4 or 1 cup with probability 3/4.

MacDonald *et al.* (MacDonald et al., 1991) found that rats, indeed, preferred the certain option (A) when it was available, but their preference switched when the options were multiplied by a common factor. In other words, rats appear susceptible to the certainty effect, and show preference reversals similar to those that led Kahneman and Tversky (1979) to propose prospect theory in explaining choice between uncertain outcomes.

Despite these studies, most investigators assume that delivery of 4 food pellets is 4 times as rewarding as a single food pellet and that a reward delivered with 0.75 probability will be 75% as desirable as the same reward delivered with certainty (van Duuren et al., 2009). In each case, a psychophysical mapping exists between the variable that can be manipulated and its subjective impact. In order to truly assess how much less valuable a reward has become by virtue of its uncertainty, the psychophysics of that reward must be conducted. In the case of brain stimulation, the subjective intensity of the rewarding effect of stimulation is a (non-linear) logistic function of the pulse frequency (Simmons and Gallistel, 1994). The rat's actual performance is a non-linear function of this subjective intensity, the subjective opportunity and effort costs, and the payoffs the rat derives from other activities. As a result, the comparison between riskless and probabilistic rewards at a single strength and response requirement is an inaccurate assay of how much risk has discounted the reward.

As an example, consider rewards delivered with a probability of 0.75. Even when the objective probability and its subjective impact are given a one-to-one mapping, such a risky reward would raise the subjective opportunity cost 4/3-fold, as the

rat must expend (on average) 4/3 as much time in pursuit of the reward when it is only delivered 3/4 of the time compared to when it is always delivered. If a rat is presented with a maximally intense reward at a sufficiently low price, the observed effect of probability would be undetectable, as performance would still be nearly maximal. If the response requirement is raised, the observed effect of probability on performance will be much more pronounced. As a result, some researchers may detect an effect, and some may not. Only by measuring the performance that results from many reward strengths and opportunity costs can the true effect of probability on performance be assessed. Only when the subjective intensity of the reward is accurately estimated can the true degree to which it is degraded by probabilistic delivery be found.

2.1.2.2 The present study

The probability that a reward will be delivered when the response requirement (price) has been fulfilled is not likely to affect the effectiveness of the stimulation to drive a given relative level of subjective reward intensity. It will, however, affect the subjective opportunity cost of obtaining rewards, since the rat will have to spend more time lever-pressing per reward when the reward is not delivered every time the lever has been held for the required amount of time. For example, the rat will have to spend roughly twice as much time depressing the lever when the reward is delivered with a probability of 0.5 as it would when the reward is delivered with certainty.

As a result, although the payoff from working for electrical rewards is altered by the probability that the reward will actually be delivered, the subjective intensity of the reward the rat eventually receives will not be affected by its probability. In effect, the probability of reward delivery affects decision-making beyond the peak detection stage, leaving F_{hm} unaffected while altering the payoff from self-stimulation. A probabilistic reward will therefore produce a decrement in P_e (the benchmark for

manipulations that occur beyond the peak-detection stage) and leave F_{hm} unchanged.

The degree to which probability affects P_e can also be used to quantify the degree to which probabilistic rewards are discounted as compared to their riskless counterparts. Provided the probability of a reward affects all opportunity costs equally, and affects neither effort costs nor the payoff from all other activities, then the decrement in P_e that results from the probabilistic nature of a reward is itself a proxy for the decision weight of that reward. If an animal is willing to work 30 seconds for a maximal reward given with certainty, and it is willing to work only 20 seconds for that same reward delivered with 0.75 probability, then the probabilistic reward is effectively under-weighted in the decision. In this example, a 4/3-fold decrement in probability is subjectively weighted as a larger, 3/2-fold decrement in payoff. In other words, the probabilistic reward carries less weight than would be normatively expected.

The psychophysical mapping between an outcome's objective probability and its subjective impact on choice has been extensively studied in humans (Kahneman and Tversky, 1979) and provides one of the two important bases of prospect theory, which seeks to answer how humans evaluate economic gains and losses under risk. Since the decisions presented to people are word-and-number problems, it would be easy to assume that the fact that human beings underweight non-trivial probabilities and overweight very low probabilities results from being literate and numerate. No "prospect theory" of rats has ever been advanced, despite various risk preference reversals in rats and pigeons (Kalenscher and van Wingerden, 2011; Kacelnik and Bateson, 1996). In many experiments in which rats perform a probabilistic task, the subjective decision-weight of a probabilistic reward is presumed equal to its objective probability (Roitman and Roitman, 2010; Gilbert et al., 2011; Zeeb and Winstanley, 2011; St Onge et al., 2010, 2011; Cardinal and Howes, 2005). It remains to be shown whether prospect theory is universally true, or whether non-human animals

underweight middle probabilities as humans do.

The goal of the following experiment is two-fold. First, we wished to validate the ability of the Reinforcement Mountain Model to correctly identify a manipulation that occurs strictly beyond the output of the peak detector. To do this, we fit the model to two types of trials the rats were exposed to in random order: riskless trials on which reward is always delivered when the response requirement is met, and risky trials on which reward is delivered with non-certain (0.75 or 0.5) probability when the response requirement is met. Second, we wished to quantify the degree to which the two probabilities of reinforcement affected the payoff, thereby deriving the objective-to-subjective psychophysical mapping of probability at 0.75 and 0.50 objective probability values.

2.2 Methods

2.2.1 Surgical procedure

Long-Evans rats (Charles River, St-Constant, QC) weighing at least 350g at the time of surgery, were stereotaxically implanted bilaterally with macro-electrodes aimed at the lateral hypothalamic level of the medial forebrain bundle. Macro-electrodes were fashioned from 00-gauge insect pins insulated to within 0.5mm of the tip with Formvar enamel. Rats received a subcutaneous injection of Atropine (0.02 to 0.05 mg/kg) to reduce mucous secretions, an intra-peritoneal injection of a Ketamine/Xylazine cocktail (87 mg/kg and 13 mg/kg, respectively) to induce anaesthesia, subcutaneous buprenorphine (0.05 mg/kg) as an analgesic, and intramuscular Penicillin (0.3 ml) to reduce infection. Rats were maintained on 0.5% isofluorane at a flow rate of 800 ml/min for the duration of stereotaxic surgery. Stereotaxic coordinates for stimulating electrodes were 2.3mm posterior to bregma, 1.7mm lateral to the midline, and halfway between 9mm from the skull surface and 8.2mm from dura

mater. A return wire was affixed to two of the skull screws anchoring the implant. The head cap, containing Amphenol pins connected to each stimulating electrode and the return wire, was cemented on the skull surface (anchored by a minimum of 5 skull screws) with dental acrylic.

Immediately following surgery, rats were given a second injection of buprenorphine (0.025 mg/kg). Rats were also given mild analgesic injections (Anafen, 5 mg/kg) 24 and 48 hours after surgery. Rats were allowed to recover for at least one week from the day of surgery before screening for self-stimulation began.

2.2.2 Behavioural protocol

Following surgical implantation of stimulation electrodes, rats were screened for self-stimulation behaviour in non-computerized operant chambers in which every depression of the lever triggered a 500ms train of 0.1ms cathodal pulses delivered to one of the hemispheres, on a continuous reinforcement schedule. Only animals who quickly learned to avidly depress the lever without stimulation-induced involuntary movements or evident signs of aversion (vocalizations, withdrawal or escape behaviors) were retained for this study. Currents were tested from 200 to 1000uA, adjusted for each rat and each electrode to provide optimal performance.

After screening, rats underwent operant training in the computer-controlled testing boxes that would eventually be used for the experiment. All tests were conducted in the dark phase of their light/dark cycle. Rats were first presented with a repeating sequence of 10 trials in which the first two trials were identical and each subsequent trial delivered stimulation that decremented in pulse frequency by 0.1 common logarithmic steps. Trials were signalled by a house light that flashed for the duration of a ten-second inter-trial interval; priming stimulation consisting of the highest pulse frequency the animal could tolerate at a train duration of 500msec was delivered two seconds before the end of the trial. Each trial lasted 25 times the

objective price, allowing the rat to obtain a maximum of 25 rewards if it held the lever continuously throughout the trial. The price, pulse frequency, and probability of reinforcement were held constant for the duration of a trial. A cumulative handling time schedule (Breton et al., 2009) was in effect for the remainder of the experiment. In this schedule of reinforcement, a reward is delivered only when the cumulative amount of time the rat has spent holding the lever reaches a set criterion (the "price" of the stimulation). For this first phase of training, the price was set to 1s. Pulse frequencies were adjusted throughout to ensure a range of frequencies that produced high time allocation ratios, a range that produced low time allocation ratios, and a range that produced intermediate time allocation ratios.

When performance on such training "frequency sweeps" was reliably high on high-frequency trials and low on low-frequency trials, as determined by visual inspection, rats were presented with a repeating sequence of 10 trials in which the first two trials were identical and each subsequent trial delivered stimulation that incremented in price by 0.125 common logarithmic steps. The pulse frequency delivered on these trials was as high as the animals would tolerate without involuntary stimulation-induced movements or vocalizations. Training on these "price sweeps" was considered complete when low prices produced reliably high time allocation ratios and high prices produced reliably low time allocation ratios, as determined by visual inspection.

Following "price sweep" training, rats were presented with a repeating sequence of 10 trials in which the first two were identical and each subsequent trial decremented in pulse frequency and incremented in price. The actual prices and frequencies of stimulation were arrayed along a line that passed through a price of 4s and the pulse frequency delivered during price sweeps, and through the price and frequency that produced half-maximal performance on price and frequency sweeps, respectively, in logarithmic space. Training on these "radial sweeps" was considered

complete when high payoff (high frequency, low price) trials produced reliably high time allocation ratios, and low payoff (low frequency, high price) trials produced reliably low time allocation ratios by visual inspection.

When training was complete, animals progressed to the discounting portion of the experiment. Preliminary fits to the frequency, price, and radial sweeps were used to aim three vectors in the sample space of prices and pulse frequencies: a vector of 9 frequencies obtained at a price of 4s (the frequency pseudo-sweep), a vector of 9 prices obtained at the highest frequency the animal would tolerate (the price pseudo-sweep), and a vector of 9 price-frequency pairs that was arrayed along the line that passed through the intersection of the frequency and price pseudo-sweeps and through the anticipated value of F_{hm} and P_e. The vectors thus formed describe the set of price-frequency pairs that would be delivered on certain (P=1.00) trials. These vectors were shifted leftward along the price axis by 0.125 common logarithmic units (decreasing all prices on those trials by roughly 25%) for the list of price-frequency pairs that would be delivered on risky trials where the probability of reinforcement following successful completion of the work requirement was 0.75. The vectors were shifted leftward along the price axis by 0.30 common logarithmic units (decreasing all prices on those trials by roughly 50%) for the list of price-frequency pairs that would be delivered on risky trials where the probability of reinforcement following successful completion of the work requirement was 0.50.

The first probability condition rats encountered was 0.75 (P1vp75). A master list combining the frequency, price, and radial pseudo-sweeps for P=1.00 and P=0.75 conditions was assembled. The central 5 price-frequency pairs of each pseudo-sweep (the 3rd through the 8th elements of each pseudo-sweep when ordered by payoff) were repeated in this master list. As a result, we collected twice as many observations of the time allocation ratio in the dynamic range of the performance curve, reducing our uncertainty about the position of the curve along either the frequency or price axes.

This master list was then randomized in a new list, providing one full "survey," or a full description of performance at each point in the parameter space that was tested.

Rats were presented with repeating triads in which the first trial delivered the highest pulse frequency the animal could tolerate at a price of 1s. The price and pulse frequency of the stimulation delivered on the second trial were drawn without replacement from the randomized list. The third trial of the triad delivered 10Hz stimulation, a pulse train so weak the animals would never work for it, at a price of 1s. This triad sequence was repeated until all trials in the master list had been presented, for a maximum session duration of 9h a day. On certain trials (P=1.00), the reward was always delivered when the cumulative amount of time the rat spent holding the lever reached the price on that trial. On risky (P=0.75) trials, the reward was only delivered with a probability of 0.75 when the cumulative amount of time the lever had been depressed reached the price on that trial. Only one lever was armed on any given trial.

For rat MA5, the same lever served as manipulandum for both certain (P=1.00) and risky (P<1.00). For rats DE1, DE3, DE7, PD8, DE14, DE15, DE19 and DE20, one lever was mapped to all trials in which reward was certain and the other lever was mapped to all trials in which reward was risky. In all cases, a steady cue light mounted above the lever signalled that reward would be delivered with certainty, while a flashing cue light (300ms on, 300ms off) signalled that the reward would not be delivered with certainty.

When performance was judged stable by visual inspection for 8 consecutive "surveys," or, in other words, when the entire master list had been presented 8 times and led to reliably high time allocation ratios on high payoff trials, reliably low time allocation on low payoff trials, and intermediate time allocation ratios on intermediate trials, the probability of reinforcement was changed to 0.50 (P1vp5). A new master list was created by amalgamating the frequency, price, and radial pseudo-sweeps for

the certain (P=1.00) condition with the new risky (P=0.50) condition. As above, the central 5 points of each pseudo-sweep were double-sampled. The list was presented again, in triads for which the 2nd trial was now randomly drawn without replacement from the new master list.

When performance on this new probability condition was judged stable by visual inspection for 8 consecutive "surveys," the location of the certain (P=1.00) and risky (P=0.50) rewards was switched (P1vp5sw). A steady cue light still signalled that the lever would always deliver reward, and a flashing cue light still signalled that the lever would not always deliver reward, but the mapping of levers to those probabilities was inverted. If, for example, the lever delivering certain rewards was on the left side for the previous two probability conditions, the right lever would now fulfil that role, and vice-versa. This switch enabled us to partly control for side-preferences.

After rats completed 8 stable surveys comparing certain and risky rewards, the probability was changed again to 0.75 (P1vp75sw). A master list was constructed again by amalgamating pseudo-sweeps for the P=1.00 condition with those for the P=0.75 condition, double-sampling the central 5 price-frequency pairs as above. The levers maintained their inverse mapping, and the 2nd trial of every trial was drawn at random without replacement from this final master list. Data collection continued until 8 stable surveys were collected under this switched certain (P=1.00) compared to risky (P=0.75) condition.

Rats DE1, DE3, and DE7 began the experiment as MA5, with probabilities mapped to the same lever but signalled with a steady or flashing light. As no difference in performance was observed under either P1vp75 or P1vp5 between certain and risky conditions, mapping of levers to probabilities was instituted, as described above. Then, 8 stable surveys at P1vp75 and 8 (DE1), 5 (DE3) or 6 (DE7) surveys at P1vp5.

In summary, rats were presented with a triad sequence of trials in which the

first delivered strong, inexpensive stimulation, the second delivered a trial of random price and frequency drawn from the P1vp75, P1vp5, P1vp5sw or P1vp75sw lists, and a third trial delivered weak, inexpensive stimulation. The order of the probability conditions was always P1vp75, followed by P1vp5, P1vp5sw, and finally P1vp75sw. Rat MA5 did not undergo the lever-switch conditions, as a single lever was used for both conditions. Due to the substantial duration of the individual conditions, most rats did not complete the entire experiment. Rat DE1 became ill at the start of P1vp5, rat DE7 became ill at the start of P1vp5sw, PD8 became ill midway through P1vp5, DE14 became ill at the end of P1vp5sw, and DE15 became ill at the end of P1vp5.

2.2.3 Statistical analysis

The dependent measure was corrected time allocation, the proportion of trial time the animal spent working for brain stimulation rewards. The correction involved including lever releases lasting less than 1s along with lever holds as our measure of corrected work time (Hernandez et al., 2010). Corrected time allocation was therefore calculated as the total amount of time the lever was depressed (for any period of time) or released for less than 1s, divided by the total trial time. The Reinforcement Mountain Model surface

$$TA = (TA_{max} - TA_{min})\frac{(\frac{f^G}{f^G + F_{hm}^G})^A}{(\frac{f^G}{f^G + F_{hm}^G})^A + (\frac{p}{P_e})^A} + TA_{min}$$

was fit to the corrected time allocation measured at each combination of pulse frequency (f), price (p), and probability condition. The only parameters of the model that were free to vary between probability conditions were F_{hm}, the location of the surface along the frequency axis, and P_e, its location along the price axis. Slope (A, G), ceiling (TA_{max}) and floor (TA_{min}) parameters were not free to vary between

probability conditions. Separate fits were conducted for P1vp75, P1vp5, P1vp5sw and P1vp75sw conditions.

A bootstrapping approach was used to derive the confidence intervals around F_{hm}, P_e, and the probability condition-induced difference in both. The bootstrapping and fitting algorithms were both implemented in MATLAB R2011b (The Mathworks, Natick, MA). Corrected time allocation values were sampled 1000 times from the observed data with replacement. For example, if 8 time allocation values were observed at a particular price, pulse frequency, and reward probability, the bootstrapping procedure would obtain 1000 samples of 8 time allocation values obtained pseudo-randomly from that list of 8 (with replacement). A mountain surface was fit to each of the 1000 re-sampled replications, thereby producing 1000 estimates of F_{hm} at each probability condition and 1000 estimates of P_e at each probability condition. The 95%, bootstrap-derived confidence interval about F_{hm} and P_e was defined as the range within which the central 950 sample F_{hm} and P_e values lay, excluding the lowest and highest 25. Similarly, we computed the difference between estimates of F_{hm} and P_e during riskless and risky trials by obtaining the difference for each replication. In other words, each replication had an estimate of F_{hm} for riskless (P=1.00) trials and one for risky (P=0.75 or P=0.50) trials, and the parameter difference in F_{hm} for the replication was the difference between each. The 95% bootstrap-derived confidence interval about the difference in F_{hm} and P_e was defined as the range within which the central 950 sample differences lay for each parameter, excluding the lowest and highest 25. Our criterion for statistical reliability was non-overlap of the 95% confidence interval about the difference with 0. A probability-induced difference in F_{hm} or P_e was therefore considered statistically reliable if and only if the 95% confidence interval about the difference did not include 0. Any confidence interval around a difference that included 0 was considered statistically unreliable.

To evaluate our parsing of the probability conditions, we fit a series of moun-

tains to the data from trials in the P=1.00 conditions of each phase. Risk-less test trials were extracted from the P1vp75, P1vp5, P1vp5sw, and P1vp75sw phases, and a mountain surface was fit to the data from each rat according to which slope (A, G) and range (TAmax, TAmin) parameters were common to all conditions and location (F_{hm}, P_e) parameters were free to vary among the different phases. The same re-sampling method was conducted as described above. We then obtained the difference, where applicable, in estimated F_{hm} (or P_e) values between P1vp75 and each phase that followed it for the 1000 re-sampled estimates, and identified the central tendency (median) of the 1000 differences between P1vp75 and P1vp5, P1vp75 and P1vp5sw, and P1vp75 and P1vp75sw. These differences provide an indication of the degree to which F_{hm} and P_e fluctuated throughout the experiment. Because of large fluctuations occurring in the course of months-long testing conditions, we chose to consider each presented pair of probabilities separately rather than in the aggregate.

2.3 Results

2.3.1 Stability of F_{hm} and P_e

All phases of the experiment—P1vp75, P1vp5, P1vp5sw and P1vp75sw—have the P=1.00 condition in common. Since the data from P=1.00 conditions differ across phases with respect to time, statistically reliable changes in the estimated F_{hm} and P_e for the P=1.00 across the experimental phases would complicate aggregating all of the data together. In the absence of any drift, the estimated F_{hm} and P_e values for rewards delivered with certainty would be identical in all phases, thereby justifying a single, three-way comparison between P=1.00, P=0.75, and P=0.50 conditions. We therefore sought to test the null hypothesis that mountain estimates when the reward was riskless were identical in all phases of the experiment.

Figure 2.3 shows the difference in F_{hm} (left) and P_e (right) between the first

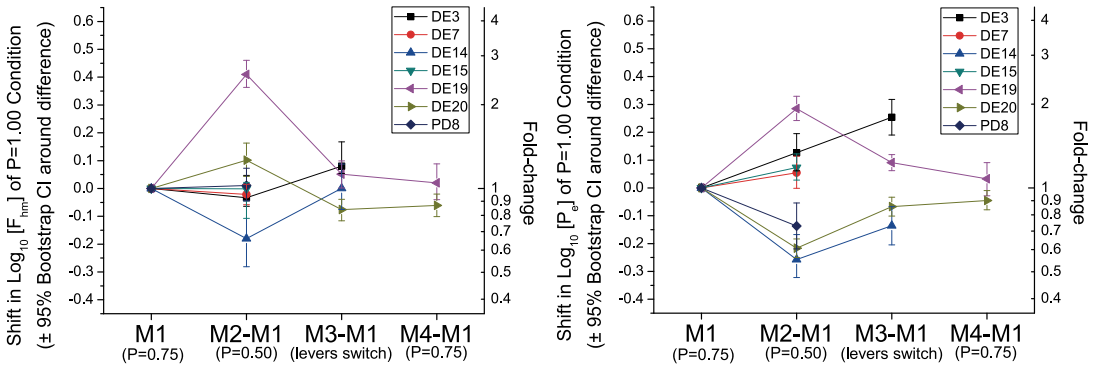

Figure 2.3. Stability of F_{hm} and P_e across phases of riskless conditions. The estimated F_{hm} (left) and P_e (right) of the P=1.00 condition are compared across each phase of the experiment, normalized to the first (P1vp75) condition encountered. Although there is little evidence for a systematic drift in parameters over time, there are large reliable (though non-systematic) changes from one condition to the next in all animals.

phase of the experiment and each subsequent phase, along with the bootstrap-derived confidence interval associated with that difference. Although there is a clear indication of a steady drift in the parameter values from the start to the end of the experiment in only one rat (DE3), all other animals show statistically reliable changes (confidence intervals about the difference that do not include 0) in those parameters at least at one point in time. Since probabilistic and risk-less trials are presented inter-digitated in random fashion, these drifts in F_{hm} and P_e would constitute part of the statistical noise in estimating the probability-induced difference in those parameters. Over the course of the entire experiment, however, these data suggest it is unreasonable to assume the subject and electrode/brain interface that underwent the first phase of the experiment is in every way the exact same subject and electrode/brain interface when it underwent the second phase of the experiment, occurring perhaps many months later. We therefore present here separate comparisons of probability values for each phase of the experiment, for each subject that was tested in that phase.

2.3.2 Effects of probability on F_{hm} and P_e

2.3.2.1 P1vp75

All rats completed the first phase of the experiment, P1vp75. In order to gauge the estimated difference in parameters from one condition to the other, the list of 1000 estimates of F_{hm} and P_e for the riskless condition obtained by re-sampling were subtracted from the 1000 estimates of F_{hm} and P_e for the risky (P=0.75) condition. The median of $riskless\ estimate - risky\ estimate$ was used as a measure of central tendency for the change in parameter estimate produced by risk. The 2.5% and 97.5% percentiles of the differences were used as an estimate of the 95% confidence interval surrounding the difference in F_{hm} and P_e between 1.00 and 0.75 probability conditions.

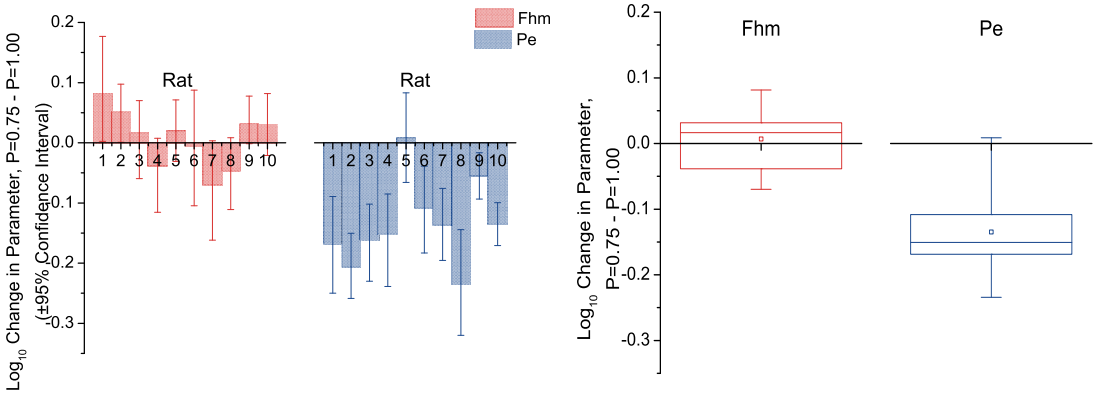

Figure 2.4. Shift in F_{hm} and P_e for P1vp75. Bar graphs (left) provide the magnitude (±95% bootstrap confidence interval) of the difference in F_{hm} (red) and P_e (blue) from riskless (P=1.00) to risky (P=0.75) conditions in the first phase of the experiment. Positive numbers indicate that the risky conditions have greater estimates, while negative numbers indicate that the estimates are lower on risky trials. The box-whisker plot (right) provides the median (middle line), inter-quartile range (box), full range (whiskers), and mean (square) of the estimated differences.

Figure 2.4 shows the estimated difference in F_{hm} and P_e for all animals. In the left-hand panel, the bar graph depicts, for all animals, the estimated difference in F_{hm} (red) and P_e (blue) along with the bootstrap-estimated 95% confidence interval about the difference. In the right-hand panel, a box-whisker plot depicts the estimated difference (red for F_{hm}, blue for P_e) collapsed across all animals. Although some animals showed only modest or unreliable shifts along the price axis, the median difference between risk-less and risky rewards is a 0.14379 decrease, with an interquartile range spanning from 0.10817 to 0.16865. Conversely, although one animal (DE1) showed a statistically reliable shift along the frequency axis, the median difference across all animals is approximately zero (-0.01791, IQR spans -0.0385 to 0.03159).

2.3.2.2 P1vp5

Rats DE3, DE7, PD8, DE14, DE19 and DE20 completed the second phase of the experiment, P1vp5. As above, the list of 1000 estimates of F_{hm} and P_e fit by the bootstrapping procedure was used to derive the estimated difference in F_{hm} and P_e produced by risk. The 1000 estimates of F_{hm} at each probability condition were subtracted from each other, and the 1000 estimates of P_e at each probability condition were subtracted from each other. The median difference was used as a measure of central tendency of the difference produced by risk, and the 2.5% and 97.5% percentiles were used as the bounds of the 95% confidence interval surrounding the difference.

Figure 2.5 shows the difference in parameter estimate produced by delivering rewards with 50% probability. The left-hand panel shows the estimated difference in F_{hm} (red) and P_e (blue) for these 5 animals between risky (probability of 0.50) and riskless (probability of 1.00) rewards. Error bars represent the 95% confidence interval about this difference. The right-hand panel collapses the estimated difference across all rats, depicting a box-whisker plot of the difference in F_{hm} (red) and

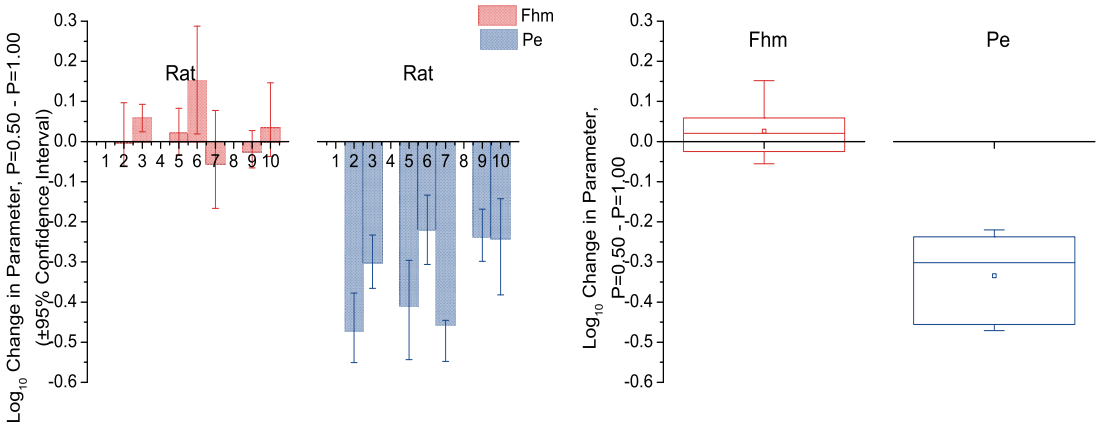

Figure 2.5. Shift in F_{hm} and P_e for P1vp5. Bar graphs (left) provide the magnitude (±95% bootstrap confidence interval) of the difference in F_{hm} (red) and P_e (blue) from riskless (P=1.00) to risky (P=0.5) conditions in the second phase of the experiment. Positive numbers indicate that the risky conditions have greater estimates, while negative numbers indicate that the estimates are lower on risky trials. The box-whisker plot (right) provides the median (middle line), inter-quartile range (box), full range (whiskers), and mean (square) of the estimated differences.

P_e (blue) produced by risk. Although two animals (DE7, DE14) showed a reliable shift in F_{hm}, overall, delivering rewards with a probability of 0.5 did not alter F_{hm} systematically. Instead, as expected, risky rewards produced large (median of 0.30199 common logarithmic units) statistically reliable shifts in P_e. The change in F_{hm} produced by delivering rewards with a probability of 0.5 had a median value of 0.02056, with an interquartile interval spanning from −0.02482 to 0.05908. The change in P_e, in contrast, had a median value of −.30199 with an interquartile interval spanning from −0.45593 to −0.23712.

2.3.2.3 P1vp5sw

Rats DE3, DE14, DE19 and DE20 completed the third phase of the experiment, P1vp5sw. In this condition, the mapping between levers and probabilities switched, such that probabilistic rewards were associated with the right lever and certain rewards were associated with the left lever. The procedure for estimating the difference in F_{hm} and P_e produced by probability was the same as for phases P1vp75 and P1vp5.

Figure 2.6 shows the difference in parameter estimates between probability conditions, with lever-to-probability mappings switched. In the left-hand panel, the difference in F_{hm} (red) and P_e (blue) is shown for each rat, accompanied by the bootstrapped-derived estimates of the 95% confidence intervals. In the right-hand panel, the difference is collapsed across all 4 animals in a box-whisker plot. Although the shift in F_{hm} is not statistically reliable in only two cases (DE19, DE20), the difference produced by a change in probability is inconsistent across the small number of rats that made it to this stage of the experiment, and inconsistent with the changes in F_{hm} that were observed in previous probability conditions. Estimated differences in P_e are similarly accompanied by wide confidence intervals, but their magnitude is more consistent across animals than the difference in F_{hm}. The median difference in

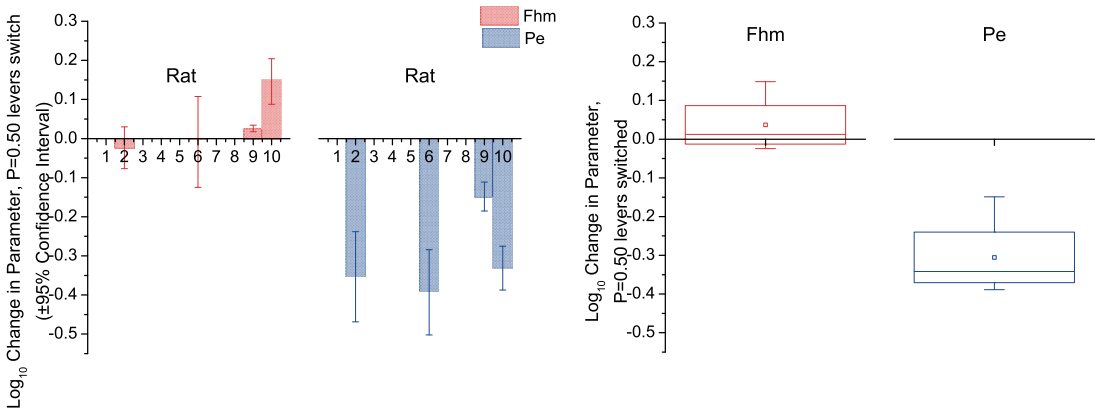

Figure 2.6. Shift in F_{hm} and P_e for P1vp5sw. Bar graphs (left) provide the magnitude (±95% bootstrap confidence interval) of the difference in F_{hm} (red) and P_e (blue) from riskless (P=1.00) to risky (P=0.5) conditions in the third phase of the experiment, when the mapping of lever sides to probability are switched with respect to the second phase. Positive numbers indicate that the risky conditions have greater estimates, while negative numbers indicate that the estimates are lower on risky trials. The box-whisker plot (right) provides the median (middle line), inter-quartile range (box), full range (whiskers), and mean (square) of the estimated differences.

F_{hm} for this phase of the experiment was a 0.01212 decrease, with an interquartile range spanning from −0.01264 to 0.08674, while the median difference in P_e was a 0.3418 decrease, with an interquartile range spanning from −0.24 to −0.37072.

2.3.2.4 P1vp75sw

Rats DE19 and DE20 completed the final phase of the experiment, P1vp75sw. In this phase, the probability of reinforcement on the risky lever was increased back to 0.75, but the lever mapping was retained as in the third phase (P1vp5sw). Estimated differences in parameter estimates and their associated confidence intervals were computed as before, using the list of 1000 estimates in the bootstrapping procedure.

Figure 2.7 shows the effect of probability on the estimated difference in parameter estimates for this final phase of the experiment. The figure depicts the change in F_{hm} (red) and P_e (blue) produced by delivering rewards with a probability of 0.75 rather than 1.00. The difference was not collapsed across animals, as only two rats survived to the end of the experiment. Qualitatively, the shifts observed during this replication of the P1vp75 condition are comparable to those observed in the first phase of the experiment, albeit with much greater variability. Rat DE19 showed only a very modest shift in P_e during phase P1vp75, which went in the opposite direction (and was not reliably different from 0) in phase P1vp75sw, while DE20 showed a slightly larger shift in P_e during phase P1vp75 that was slightly reduced during phase P1vp75sw. The shift in F_{hm} was only reliable for DE20, which is inconsistent with the shift in F_{hm} observed for this animal in the first phase of the experiment.

2.3.3 Magnitude of the difference in F_{hm} and P_e

The estimated difference in F_{hm} and P_e induced by a given probability condition can be collapsed across animals and lever-mapping conditions to provide an estimate of how probability affects each parameter. Figure 2.8 shows the box-whisker

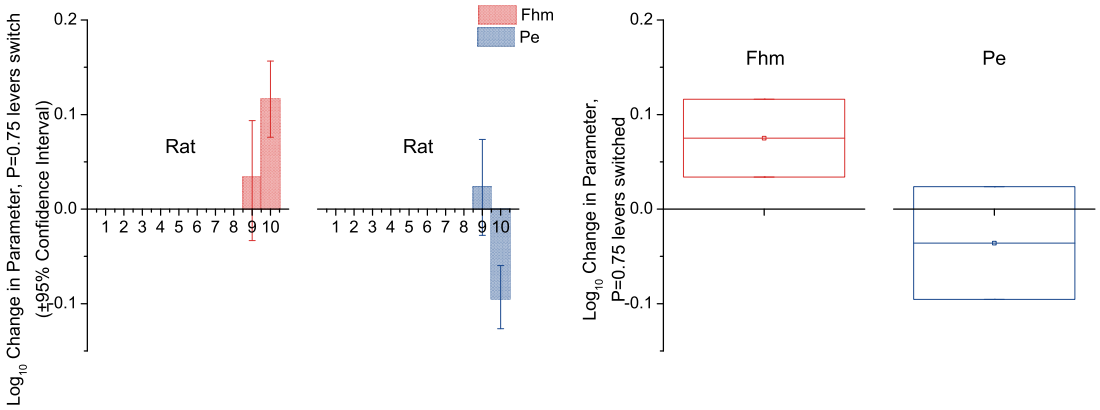

Figure 2.7. Shift in F_{hm} and P_e for P1vp75sw. Bar graphs (left) provide the magnitude (±95% bootstrap confidence interval) of the difference in F_{hm} (red) and P_e (blue) from riskless (P=1.00) to risky (P=0.75) conditions in the final phase of the experiment, when the mapping between levers and probabilities are switched with respect to the first phase. Positive numbers indicate that the risky conditions have greater estimates, while negative numbers indicate that the estimates are lower on risky trials. The box-whisker plot (right) provides the median (middle line), inter-quartile range (box), full range (whiskers), and mean (square) of the estimated differences.

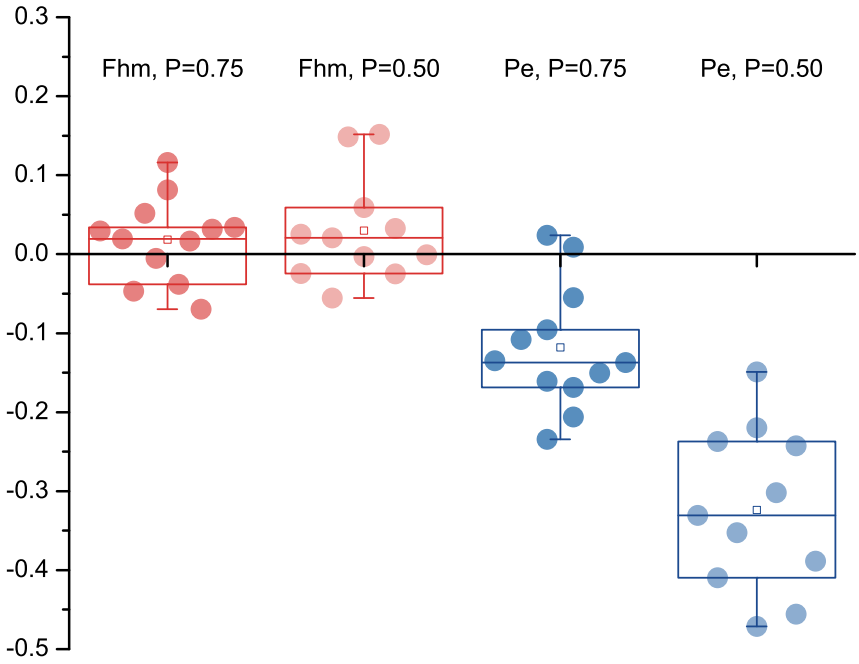

Figure 2.8. Magnitude of the difference in F_{hm} and P_e across all conditions. Box-whisker plots show the change in F_{hm} (red) and P_e (blue) resulting from a decrease in probability to 0.75 (dark symbols) and 0.50 (light symbols). Squares represent means, whiskers represent the full range of differences. Negative numbers indicate that risk decreases the estimate; positive numbers indicate that risk increases it. Filled circles represent differences for each animal in each condition.

plot of all the shifts in F_{hm} and P_e observed between P=1.00 and P=0.75 conditions (collapsing P1vp75 and P1vp75sw together), as well as the bar-whisker plot of all the shifts in F_{hm} and P_e observed between P=1.00 and P=0.50 conditions (collapsing P1vp5 and P1vp5sw together). The shift in F_{hm} is close to 0 for both probabilistic rewards (median increase of 0.02422 for 0.75 and median increase of 0.02056 for 0.50), whereas the shift in P_e is probability dependent. When rewards are delivered with 0.75 probability, the estimate of P_e decreases by 0.13598 (the inter-quartile range spans 0.07531 to 0.16481) compared to those delivered with a probability of 1.00. When rewards are delivered with 0.50 probability, the estimate of P_e decreases by 0.33089 (inter-quartile range spanning spans 0.23712 to 0.4095) compared to those delivered with a probability of 1.00.

2.4 Discussion

2.4.1 Utility of the mountain model

The Reinforcement Mountain Model has been proposed as a means to infer the stage of processing at which a manipulation acts to alter reward seeking. According to the model, manipulations that alter the pulse frequency that produces a half-maximal reward (F_{hm}) occur prior to the output of a circuit that integrates the aggregate activity induced by the stimulation electrode. In other words, manipulations that alter F_{hm} presumably operate on neurons responsible for the rewarding effect of electrical stimulation of the medial forebrain bundle. In contrast, manipulations that alter the price that drives equi-preference between a maximally rewarding train of brain stimulation and all other activities (P_e) occur at or beyond the output of the integration network, modifying the payoff from self-stimulation activities, rescaling the output of the peak detector, or changing the payoff from everything else. As a result, manipulations that alter P_e do not affect the primary neurons responsible

for brain stimulation reward *per se*, but at some later stage instead.

Previous validations of the Reinforcement Mountain Model have focussed on demonstrating its capacity to identify a manipulation known to affect putative reward neurons. For example, Arvanitogiannis *et al.* (2008) validated the Reinforcement Mountain Model's ability to correctly identify that alterations in both train duration and current occur prior to the output of the peak-detection stage. More recently, Breton et al. (2013) validated the model's ability to correctly identify the stage at which alterations in train duration affect brain reward circuitry using same the experimental procedure as used in the present paper. Other studies have used the mountain model to infer the stage of processing at which cocaine (Hernandez et al., 2010), the specific dopamine transporter blocker, GBR-12909 (Hernandez et al., 2012), and cannabinoid antagonists (Trujillo-Pisanty et al., 2011) act. The authors of these studies have generally concluded that the predominant effect of these pharmacological interventions occurs beyond the output of the peak detector, altering P_e with small and inconsistent effects on F_{hm}. However, it has not, until now, been shown that the Reinforcement Mountain Model can correctly identify such an effect.

We report a large, probability-dependent decrease in P_e, as predicted by the Reinforcement Mountain Model, that is accompanied by small, unreliable, and inconsistent shifts in F_{hm}. These results suggest that the model is, indeed, capable of correctly identifying a manipulation known to affect payoff without affecting the effectiveness of the induced rate of impulse flow in driving reward intensity to a given proportion of its maximal level. Along with the findings reported by Breton et al. (2013), the results of this experiment establish that the Reinforcement Mountain Model is a valid means of addressing which stage of processing a manipulation acts.

If the Reinforcement Mountain Model is a valid tool for identifying stages of processing that are affected by a manipulation, it can be used to isolate manipulations of early stages of the stimulated reward circuitry from all other types of manipula-

tions. For example, a lesion to the cell bodies from which primary reward fibres originate would increase F_{hm} without altering P_e. A pharmacological intervention that improves the efficacy of the primary reward neurons' input to the spatio-temporal integration process would decrease F_{hm} without altering P_e. A drug that decreases the subjective effort cost of lever-pressing would increase P_e without altering F_{hm}. A drug that scales down the magnitude of all rewards by a constant factor would decrease P_e without altering F_{hm}. With this new valid psychophysical tool, a world of possibilities is opened for re-interpreting the effect of a multitude of causal manipulations.

Moreover, the method used here is limited neither to electrical rewards nor to lever-pressing. It would be possible to perform a similar analysis using a different reward—such as sucrose—of varying strength (like concentration), or a different manipulandum—such as a nose-poke hole—with varying work requirements (like time spent). A similar logic would apply: the objective reinforcer, time spent, and caloric expenditure involved would each be psychophysically mapped to subjective determinants of choice, the scalar combination of which would be compared to the payoff of all other activities. By fitting a similar "sucrose mountain," varying sucrose concentration and work requirements, one would be able to extract the stage at which a manipulation has occurred.

2.4.2 A rat prospect theory?

The magnitude of the shifts in P_e from riskless to risky trials suggest that the mapping between the objective probability of a prospect and the weight it carries in a decision is linear. Indeed, the degree to which price must compensate for probability in our hands is approximately what would be normatively expected if the probability simply scaled by its reciprocal the opportunity cost of seeking rewarding brain stimulation. Suppose the effect of a reward delivered with 1/2 probability was to double the subjective opportunity cost of the reward. In such a case, a simple halving of

the price would compensate for the lowered probability. Similarly, if the effect of a reward delivered with 3/4 probability was to increase the subjective opportunity cost by 4/3-fold, the probabilistic nature of the reward would be completely compensated for by reducing the price by a factor of 3/4. The expected difference in P_e would be a decrement of 0.125 common logarithmic units (the common log of 0.75) from a probability of 1 to 0.75, and that expected difference would be a decrement of 0.3 common logarithmic units (the common log of 0.50) from a probability of 1 to 0.50. The median shift in P_e from a probability of 1.00 to 0.75 was approximately 0.13598. This suggests that the subjective impact of rewards delivered with 75% probability was to devalue them by 73%, with an inter-quartile range spanning 68.4% to 84.1%. Similarly, the median shift in P_e from a probability of 1.00 to 0.50 was approximately 0.33089. This shift implies that rewards delivered with 50% probability were devalued by 47%, with an inter-quartile range spanning 38.9% to 57.9%.

Unlike what is observed in human participants asked to evaluate probabilities in word-and-number problems (though see Hertwig et al., 2004), there was no clear evidence of a non-linear mapping between objective probability and subjective decision weight in our data. If we attribute a decision weight of 1 to a probability of 1, the subjective decision weight of a reward delivered with 0.75 probability was approximately 0.73, and that of a reward delivered with 0.50 probability was approximately 0.46. It remains to be shown whether, like humans, the subjective decision weight of very low probabilities is overweighted compared to the normative value. Various groups (van Duuren et al., 2009; Burke and Tobler, 2011) have attempted to record from populations of neurons while animals made choices among probabilistic prospects. In all of these cases, the mapping of both magnitude and probability to their subjective equivalents has been assumed linear for the purposes of correlating activity to various determinants of decisions. The present study provides evidence that, at least in the case of rewarding brain stimulation delivered with probabilities of

0.75 and 0.50, the assumption of a one-to-one mapping holds at the level of precision we were able to achieve.

The median shift in F_{hm} induced by a probabilistic reward is, as predicted by the mountain model, within the level of session-to-session noise one would expect. However, the median shift in P_e induced by probabilistic rewards is large and probability-dependent. Assuming the probabilistic nature of the reward alters only the rat's subjective evaluation of the opportunity costs of acquiring it, rather than the payoff from everything else or the subjective effort costs of self-stimulation, the shift in P_e can be used to derive the approximate subjective decision weight of the risky reward. At a probability of 1, the subjective decision weight is necessarily 1: the rat knows it will get rewarded if it invests a sufficient amount of time into lever pressing at the expense of all other activities. At a probability of 0.5, normative accounts of how the rat ought to allocate his time would dictate that, because the rat must (on average) fulfil the work requirement twice in order to trigger an electrical reward, the subjective opportunity cost ought to double. This would mean that if a maximal reward delivered with certainty will require a price P_{e_1} to drive the payoff from self-stimulation to the same level as that of everything else, a maximal reward delivered with a probability of 0.5 will require only half of P_{e_1} to drive its payoff to that of everything else. The subjective opportunity cost of that reward has effectively doubled. The data presented here provide evidence that the payoff from rewards delivered with 0.75 and 0.5 probabilities is scaled by roughly 0.75 and 0.5 compared to rewards delivered with certainty. Figure 2.9 provides a graphical representation of the derived subjective weight of the two probabilities (the anti-log of the shift $riskless - risky$) as a function of the objective probability of reinforcement, along with the standard error of the mean surrounding this estimate. When looking across rats, collapsing across lever mappings, the function is remarkably close to the identity function (i.e., subjective equals objective probability).

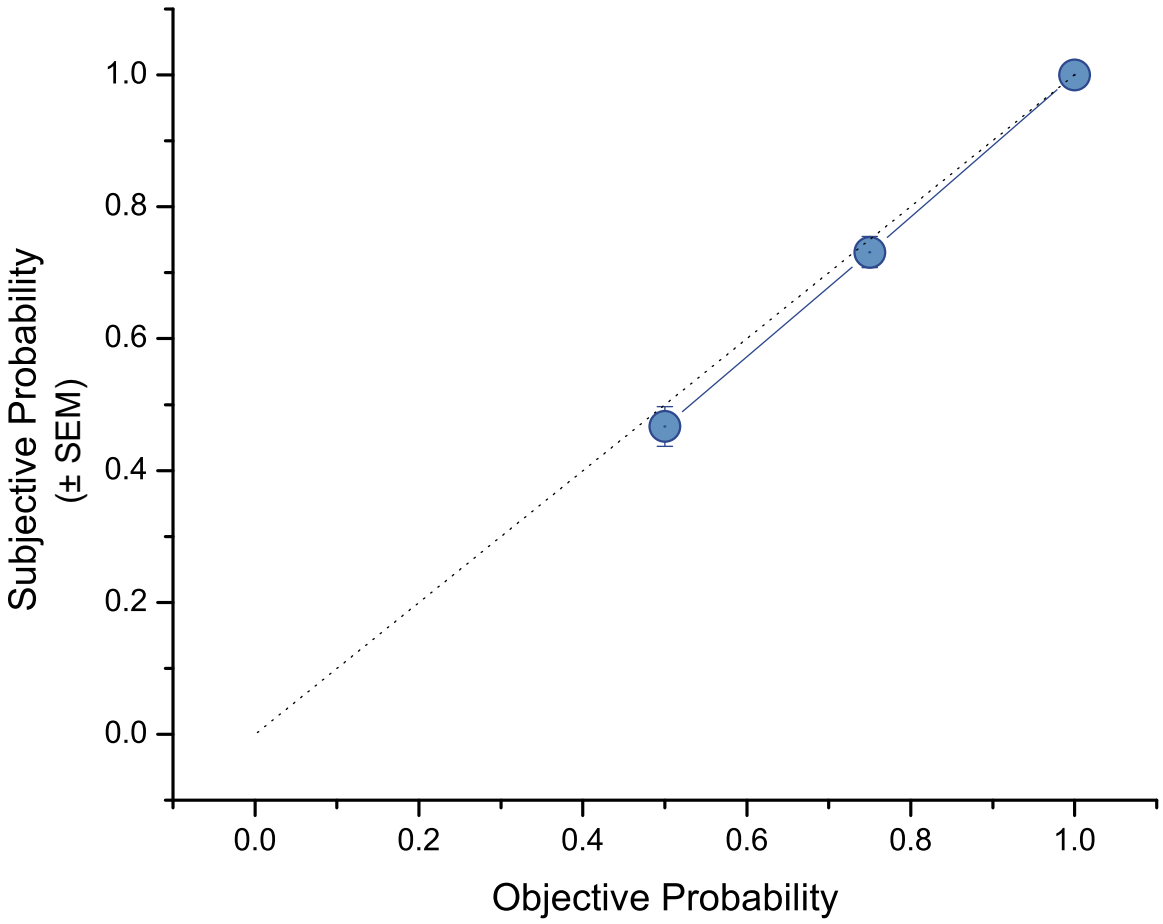

Figure 2.9. Derived subjective-to-objective mapping of probability. The dotted straight line provides a reference for what the mapping of probability would be if subjective risk were equal to objective probability. Assuming that $P = 1.00$ is subjectively interpreted as 1.00, the anti-log of the change in $Log_{10}[P_e]$ from riskless to risky is an index of the subjective probability. Blue circles indicate the mean derived subjective probability of rewards ($\pm SEM$) delivered with $P = 0.75$ and $P = 0.50$ probabilities.

2.4.3 General discussion

The purpose of the current experiment was two-fold: to confirm the Reinforcement Mountain Model's validity in correctly identifying an effect occurring beyond the output of the spatiotemporal integrator, and to quantify the degree to which various probabilities of reinforcement affect the payoff from self-stimulation. When rats were presented with trials in which rewards were delivered probabilistically, we observed no replicable, stationary, consistent differences in F_{hm} compared to trials in which rewards were delivered deterministically. However, we observed large, consistent and reliable changes in P_e that were dependent on the probability of reinforcement. As a result, the evidence we present here supports all the predictions of the model. The Reinforcement Mountain Model could correctly identify that probabilistic reinforcement, which does not affect the translation of stimulation strength into subjective reward intensity, did not affect reward circuitry prior to the output of the peak detection stage, by showing no statistically reliable evidence of a change in F_{hm}. Conversely, the Reinforcement Mountain Model correctly identified that probabilistic reinforcement, which affects the payoff from self-stimulation, affected reward circuitry beyond the output of the peak detection stage, by showing overwhelming evidence of a change in P_e. Furthermore, probabilistic reinforcement has a greater impact on the payoff when the probability is lower, and the difference in P_e between probabilistic and deterministic trials was also probability-dependent.

The present experiment cements the validity of the Reinforcement Mountain Model by providing the obverse set of predictions from Arvanitogiannis and Shizgal (2008) and Breton et al. (2013). In those experiments, the Reinforcement Mountain Model was shown to be a valid means of identifying that a manipulation acting prior to the output of the peak detector, and not after, has occurred. In these experiments, train duration (Breton et al., 2013; Arvanitogiannis and Shizgal, 2008) or pulse current

(Arvanitogiannis and Shizgal, 2008) were altered, and changes in F_{hm} were observed in the absence of changes in P_e (with some notable exceptions). In contrast, the present experiment demonstrates that altering probability of reinforcement led to changes in P_e in the absence of changes in F_{hm}.

With this valid measurement tool, it is now possible to identify the stage of processing at which a large number of manipulations act. For example, the Reinforcement Mountain Model has been used to identify the stage of processing at which cocaine (Hernandez et al., 2010), the dopamine transporter blocker GBR-12909 (Hernandez et al., 2012), the CB1 receptor antagonist AM-251 (Trujillo-Pisanty et al., 2011), and the neuroleptic pimozide (Trujillo-Pisanty et al., 2012) act. The entire catalogue of manipulations affecting brain reward circuitry can be re-examined in the Reinforcement Mountain Model context, providing a more refined means of identifying which manipulations alter the psychophysical mapping of stimulation strength to subjective reward intensity, and which of them alter the payoffs of self-stimulation and other activities, independently of the first-stage neurons responsible for the rewarding effect. It would be very useful, indeed, to incorporate recent advances in optogenetics with the Reinforcement Mountain Model to identify the source of the rewarding effect of electrical stimulation. ‘

Chapter 3

The rat's world model of session structure

3.1 Introduction

As an animal navigates its environment, it will come across patches where food is bountiful, patches where food is scarce, patches where food is delectable and patches where food is barely worth eating. Similarly, a patch where foraging will provide a high payoff (low cost, high quality food) at one moment may provide a much lower payoff (high cost or low quality food) at a later time. In order to select actions advantageously, the animal stands to benefit from a cognitive map of where and when the payoff from foraging will be high and when it will be low. When the payoff from pursuing food rewards is negligible, the advantageous choice is to pursue other goals. If the payoff from pursuing a reward changes predictably, an animal benefits tremendously from accurately developing and quickly updating a cognitive model of how the payoff from that reward changes over time.

In parametric paradigms entailing two- and three-dimensional measurement of performance for brain stimulation reward, there are certainly periods of time in an experimental session when the reward will have a high subjective intensity and the cost of obtaining it will be low, as there are periods of time when it is sufficiently weak that the animal will prefer to engage in other activities in the operant chamber. Indeed, in the randomized-triads design described in Chapter 2, the pattern of changes in payoff is largely predictable. Every three trials, a high-intensity, low-cost reward will be delivered, every three trials, a reward of variable intensity and cost will be delivered, and every three trials, a low-intensity reward will be delivered. These trials are presented in sequential fashion for months at a time, leading to the question: do rats working for brain stimulation reward develop a cognitive "map" of when the

payoff from self-stimulation will be particularly high and when it will be particularly low? It would certainly be advantageous for the rat to be able to predict these changes in strength and contingency, allowing it to select actions far more efficiently than if it had to first obtain a reward of potentially negligible intensity before implementing a behavioural policy.

The following chapter concerns the existence of such a map, which we call a "world model" of session structure. After establishing that rats working for BSRs behave as if they had a model of the triad structure of an experimental session, we attempt to uncover the rule they might use to infer the current trial's payoff.

3.1.1 World Model

In the commonly used curve-shift paradigm for inferring the effects of manipulations on self-stimulation behaviour (Miliaressis et al., 1986), stimulation strength is systematically decreased, in logarithmic steps, from pulse currents, frequencies or train durations that produce asymptotically high responding to those that produce asymptotically low responding. The sequence of trials on these "sweeps" is repeated for the duration of the session, systematically decreasing from very strong to very weak and returning to strong. On average, the animal allocates all of its time responding to strong stimulation, and none of its time responding to weak stimulation. However, researchers will sometimes find anecdotally an animal that responds vigorously, on a few select trials, to stimulation that should not be particularly motivating. The duration of each trial is usually kept fixed in these protocols, which means the animal can, in principle, know when the stimulation will be highly rewarding again. It is quite possible, in this light, that on some of those very weak stimulation trials, the rat begins to expect that working at the lever will have a high payoff soon.

Indeed, this account of self-stimulation behaviour implies that the rat has a somewhat noisy internal "model" of the structure of the session. Following the trial

on which the stimulation is weakest will be a trial on which the stimulation strength has been reset to its maximum value, and following a trial on which the stimulation strength is strong will be a trial on which the stimulation is slightly weaker. It would behove the rat working for brain stimulation rewards, or indeed any reward, to exploit the statistical regularities in the world within which it lives and to infer the "kind" of trial that it is about to encounter, either with regard to the payoff, the strength, or the opportunity cost of the reward to come. The complex session structure of the randomized design used in Chapter 2, which entails both predictable and unpredictable features, provides a promising opportunity to probe the rat's ability to exploit the regularities inherent in our triad design and infer a payoff for trials that it would not be able to ascertain without an internal representation of the triad structure of experimental sessions.

3.1.2 Introduction to the world model of triad structure

3.1.2.1 Common assumptions

Model-free reinforcement learning accounts of performance for rewards imply that the rat keeps track only of the magnitude of the reward it receives. During training, a rat only learns that certain trial states (like the blackout delay during which the reward is delivered and all lever pressing activities that preceded it) have a particular value. The rat can therefore base its decision to press on the cached value of the current trial state, and if an action can be taken that will lead it to a better total net reward when considering all the costs it will incur, the rat will take such an action. When acting optimally—that is, when implementing an optimal policy—the rat takes actions for which the total net reward it expects is maximal.

According to Bellman's equation (Bellman, 1952), the total net reward at a point in time can be recursively defined as the sum of the current total net reward ob-

tained, $R(t)$, with the total discounted sum of future rewards expected when following an optimal policy. The value of a state at time step t is the sum of the immediate net reward at time step t with the discounted value of the next state, which is itself the sum of the immediate net reward at time step $t+1$ with the discounted value of the state at time step $t+2$, and so on. In other words, this formulation defines the value of a state visited at time step t as

$$V(S_t) = R(t) + \gamma \sum_k \mathcal{T}(S_t = j, S_{t+1} = k) V(S_{t+1})$$

where γ is an exponential temporal discount factor and $\mathcal{T}$ is a function that returns the probability of transitioning from state $S_t = j$ to state $S_{t+1} = k$.

In order to find the total net reward from lever-pressing at time step t, then, one would add to the (possibly zero) reward delivered at time step t the discounted sum of the total net rewards from all trial states that can be reached from lever-pressing at time step t, weighted by the probability that these states can be reached. In the simple scenario where the states can be "lever up" and "lever down", the total net reward from lever-pressing will be the (possibly zero) reward delivered when pressing summed with the temporally-discounted rewards obtained from continuing to press at time $t+1$ and the temporally-discounted rewards obtained from engaging in other activities at time step $t+1$. Just before the end of the required response interval (the price), the rat may choose to press, which will then lead to a reward, or to engage in other activities which will not. If the delivery of a reward is surprising—that is, the rat obtains a reward it may not have been expecting—then the rat must alter its estimate of the value of all lever-pressing (and non-pressing) states that led to this surprising reward. The value $\hat{V}(S_t)$ is updated for the time step just preceding reward delivery. The updated values can then be used to shape a policy that is closer to optimal. The next time the animal is in state S_{t-1}, just before the state that led

to reward delivery, there will again be a discrepancy, which will prompt the animal to update $\hat{V}(S_{t-1})$. Eventually, the estimated value of states cease to change, and the rat necessarily acts in a way that is optimal in the sense that it will pursue a policy that will provide it with the greatest net reward at the highest rate.

This model-free learning account implies that the value of taking an action in a state will be altered as a function of the discrepancy between the cached, or expected, value of a state and the reward that is actually delivered. The rat has an expectation for reward when the trial state is S at time t ($\hat{V}(S_t)$), and an expectation for reward from future trial states S' at time $t+1$ ($\hat{V}(S_{t+1})$). As a result, at time t, the rat expects that the state it is in will have value $\hat{V}(t)$ and that the next trial state it visits will have value $\hat{V}(t+1)$. If the net reward the rat obtains at time t ($R(t)$) differs from expected, then the discrepancy is

$$\delta(t) = R(t) + \gamma \sum_{k} \mathcal{T}(S_t = j, S_{t+1} = k)\hat{V}(S_{t+1}) - \hat{V}(S_t)$$

where $R(t)$ is the net reward obtained at discrete time step t, γ is a temporal discount factor, and $T(S_t, S'_{t+1})$ is the probability of transitioning from state S at time step t to a state S' at time $t+1$, and $\hat{V}(S_{t+1})$ is the total net reward expected from future states.

This delta describes the difference between the current estimate of the value expected by being in state S at time step t, $\hat{V}(t)$, and the actual total net reward by being in state S at time step t, $R(t) + \gamma\mathcal{T}(S_t, S'_{t+1})\hat{V}(t+1)$. The value of state S at time t can then be updated based on this discrepancy by the following learning rule (Dayan and Abbott, 2001)

$$\hat{V}(S_t) \leftarrow \hat{V}(S_t) + \alpha \cdot \delta(t)$$

where α is a parameter that sets the learning rate. The value of a state will be

updated at rate α to a unit step change in reward. With higher values of alpha, a given discrepancy will drastically alter the total net reward expected the next time the animal encounters state S; with lower values, a given discrepancy will have a very small effect on the expected total net reward.

From this formulation, it is clear that the animal does not learn the mapping between current states and future states as a function of the actions it can take in the current state. All that is learned by this scheme is the total discounted net reward obtained in the current state. The rat learns throughout training the total discounted net reward it can expect from lever-pressing states, and adjusts its estimate of this value as the rewards and response requirements change. The rat's decision to press—its policy—will depend on the currently cached expected value of pressing.

Suppose, for example, that the rat has acquired during training an expectation that lever-pressing for 1s will lead to a reward of maximum intensity. When the conditions change, say the response requirement is increased to 4s and the reward delivered is sub-maximal, the rat must update its evaluation of the total net reward it can expect after it has held the lever for 1s and has not yet received a reward. As it holds the lever and updates the total net reward it can expect from various trial states, the rat's estimate of those values converges on the "true" total net reward it can expect by following an optimal policy, and learning ceases until conditions change again. The rat must re-learn, as soon as brain stimulation is no longer at the expected intensity or delivered at the expected time, the value of lever-pressing and non-lever pressing states that may now be encountered.

3.1.2.2 Model-based reinforcement learning

In contrast, model-based formulations imply that the rat learns, in as simple a sense as suffices, how states (and stimuli) are related to each other: taking a particular action will lead to various trial states, each of which will bring about a net reward

that may depend on motivational variables internal to the rat. For example, the rat may learn that pressing the lever for a period of time when it is extended will bring about a blackout period accompanied by a subjective reward. In the case of a rat working for food rewards, it may learn both that it receives a food reward of intensity I and the flavour F of the food reward that will be delivered. If the food is devalued, either by selective satiety or by explicitly pairing the food with illness, a rat using a model-based learning system will cease to respond on the lever.

The underlying mapping implies that rats behave as if they investigate a simple decision tree. In the context of the experimental procedure I will describe below, the mappings underlying a model-based account of session structure are: if the current trial is a leading bracket, the payoff will be high; if the current trial is a test trial, the payoff will be intermediate on average and selected at random from a finite set; if the current trial is a trailing bracket, the payoff will be low. Once that very simple tree has been investigated, the rat need only implement an optimal policy based on the payoff it can expect from lever pressing in the state predicted to come. The cached value of trial states for trials with a completely known payoff need not be updated, because they have in essence already been learned. Only during the test trial, when there will be real variance to the payoff that can be expected, must the rat update its estimate. Without a model of how states of the world transition to each other, the rat would have to re-learn the net reward from pressing on every trial, regardless of whether the rewards and costs changed predictably.

Consider the case in which a rat that has learned a simple, yet, effective world model: Leading trials will deliver high-intensity rewards at a price of one second; they are followed by test trials delivering variable reward intensities and prices that, on average, provide an intermediate payoff; they are followed by trailing trials delivering low-intensity rewards at a price of one second and are followed by leading trials. As the rat is about to begin a trial, the rat can look ahead in the decision tree to the

correct state it will find itself in for the trial to come, identify the series of decisions it should take in order to maximize the total discounted net reward it will receive by taking that series of actions, and simply implement this optimal policy. The total discounted net reward it can expect to receive in any trial state for the trial to come has been represented somewhere which requires no further updating. The simplest state transition function, $\mathcal{T}_{S \to S'}$, is a 3x3 permutation matrix with rule "if the last state was trailing, the next state is leading; if the last state was leading, the next state will be test; if the last state was test, the net state will be trailing; no other state transitions are allowed in the world." The expected value of lever pressing in each of those states, $\hat{V}(S')$, is also an easy function to describe: "if the state is a leading bracket trial, the payoff will be high; if the state is a test trial, the payoff will be variable and intermediate on average; if the state is a trailing bracket trial, the payoff will be low." More complex state transition ($\mathcal{T}$) and value ($\hat{V}(S)$) functions are, obviously, also possible. The rat may count the number of trials, expecting a low payoff every three trials, a high payoff every three trials, and a variable payoff every three trials. An exhaustive search is also possible: if the last trial was a leading bracket trial with high payoff, the next trial could be any of the different test trials with varying probability; if the last trial was one of those test trials, the next trial is a trailing bracket trial with low payoff; if the last trial was a trailing bracket trial, the next trial will be a leading bracket trial.

We propose that instead of searching an exhaustive tree, the rat forms a state-transition function for the sequence of trials, thereby efficiently using a simple world-model of the trial sequence to retrieve a cached mean value of the payoff it expects from lever pressing as well as the variance in payoff (or lack thereof), from the start of the trial. The overall payoff it expects can then set a policy that governs action selection before the intensity and response requirements of the electrical reward it can receive are known.

3.1.2.3 Characteristics of the post-priming pause

In the randomized-trials design, high-payoff trials are always followed by trials with varying payoff, which are always followed by trials with low payoff, which are subsequently followed by high-payoff trials. Each trial is demarcated by the occurrence of a ten-second inter-trial interval, during which the overhead house light flashes and priming stimulation of constant, high pulse frequency is delivered 2s before the trial begins. At this point in the trial, a rat with no world model whatsoever would have no idea what the trial's payoff will be. Since there is no relevant state information for re-evaluating the total net reward from lever-pressing, the rat's best estimate of the payoff from lever-pressing will be its long-run expected value.

In contrast, a rat with even an elementary world model would have no problem quickly adjusting the reward it expects from lever pressing. Supposing the rat had built a world model whereby the state of the world depends on the state of the world before the flashing house light, the flashing house light signals valuable information about the total net reward the rat can expect to receive from lever-pressing. There is, of course, also the possibility that the rat learns a much more complex state-transition function which requires keeping track of what type of trial the previous trials were, and inferring what type of trial it is in based on the number of trial types in the cycle and the current phase. We suggest, instead, a simpler world model: the rat categorizes the trial types in a way similar to the way we have categorized them, and acts on its best guess of the current trial type based on a best guess of the trial type that has just come to pass. Using such a world model allows the rat to rapidly update the payoff from lever-pressing, thanks to a richer representation of what is meant by a state, incorporating both observable variables (lever-up, lever-down) and extrapolated variables (leading, test, trailing). The decision to press or not can thus be made quickly and without requiring the rat to slowly update the net rewards and costs inherent to acquiring brain stimulation rewards.

Once the rat begins pressing, it accumulates information about the objective price as it holds down the lever. The rat may interrupt its pressing, returning to the lever after some time, accumulating more information about the objective price until the work requirement is satisfied and the lever is retracted. A reward is delivered, with some probability that may or may not be 1, a blackout delay elapses, and the lever extends back into the cage.

The critical period to test the existence of a world model of triad trial structure is the first pause made as the trial begins. It occurs following the priming stimulation delivered in the inter-trial interval; as a result, we call it the post-priming pause (PPP). Throughout this period of time, the rat has no information to guide its behaviour except for the world model it may have developed as a result of being presented with triads repeatedly for weeks or months.

When the payoff is high, it would behoove the rat to begin pressing as quickly as possible, because the longer it waits, the fewer rewards it will be able to collect. When the payoff is low, the rat will benefit much more from doing all the other things it is possible for the rat to do in the operant chamber, such as exploring, grooming, and resting. When the payoff is unknown, but intermediate on average, the PPP will be intermediate.

It follows that the PPP is an ideal period of time to analyze if we are to infer whether the rat has developed a world model of the triad structure of the session. Not only can we identify whether the rat behaves as if it has a world model, but we also have a normative account of how long the duration of the PPP ought to be: short on leading bracket trials, longer on test trials, and longest (possibly censored by the end of the trial) on trailing bracket trials. A purely model-free account would make a very different prediction: on leading bracket trials following trailing trials, a PPP based on the last payoff would be longest; on test trials following leading bracket trials, the PPP would be shortest; and on trailing trials, the PPP would be inversely related to

the payoff of the test trial that preceded it.

3.1.2.4 A description of the rat's world model

On the null hypothesis that the rat maintains no model of the triad structure of the session, the very first pause it makes following the priming stimulation (the post-priming pause, or PPP) will be independent of trial type. In this case, the duration of the PPP will be, on average, the same for all trial types. If a model-free reinforcement learning scheme governed their decision to wait before a press, the PPP will depend on the payoff of the last trial: long on leading trials (following trailing trials) and short on test trials (following leading trials). On trailing trials, they will be inversely related to the payoff of the test trials that preceded them. This is because the rat using a model-free reinforcement learning scheme will begin the next trial expecting the payoff it had received on the preceding trial. If, on the other hand, the rat has a rule upon which to base the payoff on the current trial, the PPP will depend on trial type in a predictable manner: trial types with a high payoff (leading bracket trials) will have a short PPP, while trial types with a low payoff (trailing bracket trials) will have a long PPP. Whenever the payoff of the trial cannot be known beforehand, as is the case for test trials, the PPP will be independent of the payoff to come—barring some extraordinary form of rat prescience.

We propose that, rather than slowly updating the reward expected from pressing, the rat has a simple internal model of session structure. If there is a simple rule to be found, it could either involve counting (lead-test-trail-repeat) or comparing (last trial leading-like/test-like/trailing-like). Errors in either of these processes are certainly possible. It is not unreasonable to assume that a rat counting trials may lose track of the trial type it is in while working for brain stimulation rewards. Similarly, a rat comparing the last trial to its best-fitting category may mis-classify a test trial. We turn to these potential errors in world-model inference below.

Assuming the rat is counting trials, errors in the count, and thus uncharacteristically long or short PPP durations, will be spread over all trials. On the other hand, if the rat uses the reward contingency of the previous trial to gauge the current trial's type, the PPP will be uncharacteristically short on trailing bracket trials that follow test trials that resemble the leading or trailing bracket trial. In essence, test trials that are sufficiently close in subjective intensity and opportunity cost to the trailing bracket trial will lead a rat using this comparative rule to infer that the current trial is a leading rather than a trailing bracket trial, resulting in an uncharacteristically short PPP. Similarly, test trials that are sufficiently close in subjective intensity and opportunity cost to the leading bracket trial will lead a rat using a single trial look-back rule to infer that the current trial is a test rather than a trailing bracket trial, resulting in a PPP of equally uncharacteristic duration.

Figure 3.1 shows a diagram of our proposed simplified world model when working in a randomized-triads experiment. The rat maintains a simple set of rules, encapsulated in the estimated state-transition function $\hat{\mathcal{T}}_{S \to S'}$, that provide the rat with an estimate of the subjective opportunity cost ($\hat{P}_{s_{t+1}}$) and reward intensity ($\hat{I}_{bsr_{t+1}}$) on the trial to come based on the cached subjective opportunity cost ($\hat{P}_{s_t}$) and reward intensity ($\hat{I}_{bsr_t}$) on the trial that has just elapsed. The expected payoff from the upcoming trial ($\mathbb{E}[U_t]$) is simply a scalar combination of the estimates provided by the state-transition function. This expected payoff can then set the duration of the PPP. When the PPP terminates uncensored, it necessarily terminates on a hold, which allows the rat to continuously update the subjective opportunity cost of stimulation for the current trial. When the price is paid and the lever retracts, the objective price (and therefore, the subjective opportunity cost) is transparently known, at least in principle. If the reward is delivered with certainty, lever retraction also coincides with delivery of the BSR, which allows the rat to update its estimate of the intensity of the reward on the current trial as well as the payoff. The rat maintains a representa-

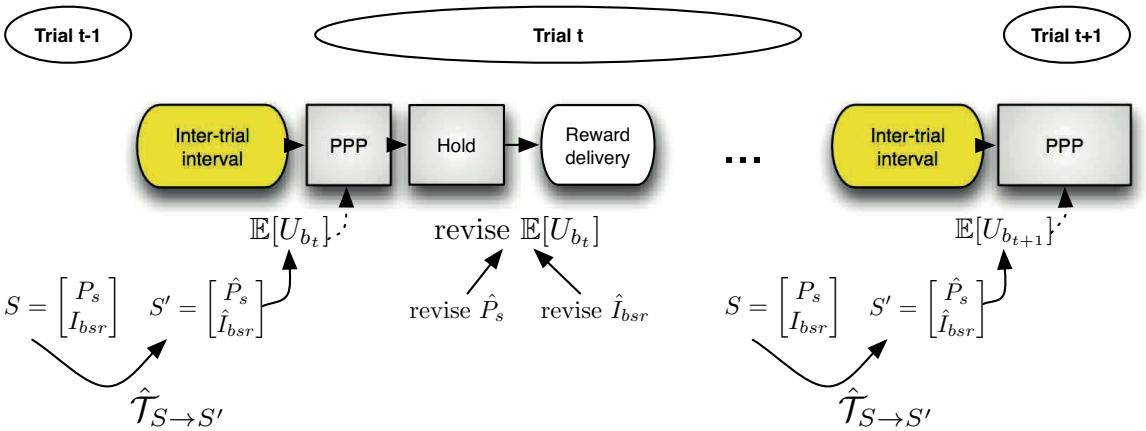

Figure 3.1. Diagram of the proposed world model of session triad structure. Over the course of the trial, the rat obtains an estimate of the subjective opportunity cost (P_s) and reward intensity (I_{bsr}) in effect, which together define the state of a trial in the session. The rat has learned a state-transition function ($\hat{\mathcal{T}}_{S \to S'}$) that describes how the state of the last trial (S) transitions to the state of a new trial (S'). This state-transition function allows the rat to make a prediction about the opportunity cost ($\hat{P}_s$) and reward intensity ($\hat{I}_{bsr}$) that can be expected on the trial to come, cued by the inter-trial interval. The elements of the state also allow the rat to make a prediction of the payoff from self-stimulation than can be expected on the trial to come ($\mathbb{E}[U_{b_t}]$). Lever-pressing allows the rat to revise its estimate of the opportunity cost, and reward delivery allows the rat to revise its estimate of the reward intensity, which could, in principle, be updated continuously throughout the entire trial. By the end of the trial, the rat uses the estimate of P_s and I_{bsr} along with the state transition function to predict conditions on the next trial. Assuming performance depends on an estimate of the payoff, the period of time between delivery of the priming stimulation during the inter-trial interval and the first lever-press produced by the rat (post-priming pause, PPP) is an estimate of the expectation the rat has of the trial to come.

tion of the previous trial's subjective opportunity cost and reward intensity (or those it had inferred if those values were not updated) as well as a representation of the current trial's determinants of decision. If the rat begins lever-pressing, these values are updated as the rat earns a reward. If the rat earns no rewards, the current trial's expected opportunity cost and reward intensity are not updated. When the flashing house light signals a new trial, the most recent estimates of subjective opportunity cost and reward intensity are used to infer what the opportunity cost and intensity will be.

To test whether there is any evidence of a world model, we assessed the duration of this PPP as a function of trial type. Moreover, we tested which of the two rules—counting or comparison—accounted best for the pattern of PPP durations if this duration was reliably related to trial type. Finally, we gauged the degree to which the test trial needed to be similar to the leading bracket trial to induce a mistake if rats indeed used a rule based on comparison. Taken together, the findings from each of these analyses provide support for the working hypothesis depicted in figure 3.1.

3.2 Methods

3.2.1 Behavioural protocol

The data analysed are a subset of the results presented in Chapter 2. Ten rats were the same as in Chapter 2, and thus underwent the protocol described above. Since for all of these rats, except MA5, the active lever on risk-based test trials was different than on leading and trailing bracket trials, we focus here on only those triads for which the probability of reinforcement was one. It would be easy for an animal to learn the repeating pattern of trial types in a triad when the active lever during test trials is different from the active lever during bracket trials; the results reported here concern only those trials which could potentially pose confusion to the rats, when

the active lever is the same for all trial types. All leading trials preceding a test trial for which the reward was not delivered with certainty, all trailing trials following a test trial for which the reward was not delivered with certainty, and all test trials for which the reward was not delivered with certainty were excluded from the analyses reported here.

The data from the ten subjects of the probability discounting experiment were supplemented by data from six subjects of a subjective-price study carried out by Rebecca Solomon. These six rats received implants similar to those in the probability discounting experiment, using the same surgical and screening protocols as described in Chapter 2. For animals F3, F9, F12, F16, F17 and F18, a train of electrical stimulation pulses was delivered when the lever had been held for a cumulative amount of time defined as the objective price. Throughout a trial, signalled in the same way as the animals that underwent the probability discounting experiment, the objective price and pulse frequency delivered were held constant. The duration of the trial was the larger of 25 times the objective price and 25 seconds.

3.2.1.1 Screening and training

For animals F3, F9, F12, F16, F17 and F18, following surgical implantation and screening for self stimulation, as described in Chapter 2, animals were presented with trials during which the price was kept constant at 4s, and the pulse frequency of the electrical stimulation decreased in constant common logarithmic steps. This frequency sweep training entailed presenting the same ten trials repeatedly for a maximum of 4 hours, where the first and second trials had identical price and pulse frequency, and trials 3 through 10 of the series delivered decreasing pulse frequencies when the lever was held for a cumulative 4 seconds. The current (as high as the rat could tolerate), train duration (500 ms), and pulse duration (0.1ms) were kept constant throughout all phases of the experiment. Pulse frequencies were adjusted such

that they spanned the dynamic range of performance, from high frequencies resulting in asymptotically high performance to low frequencies resulting in asymptotically low performance. Frequency sweep training was conducted until there was little variability in performance from ones series presentation to the next, as determined by visual inspection.

After frequency sweep training, rats were presented with trials during which the pulse frequency was kept constant at the highest pulse frequency the animal could tolerate. The objective price of the electrical stimulation increased in constant logarithmic steps from one trial to the next. This price sweep training entailed presenting the same ten trials repeatedly for a maximum of 4 hours, where the first and second trials had identical price and pulse frequency, and trials 3 through 10 of the series required increasing cumulative amounts of time to be spent holding the lever (the objective price) in order for an electrical reward to be delivered. Objective prices were adjusted such that they spanned the dynamic range of performance, from low prices resulting in asymptotically high performance to high prices resulting in asymptotically low performance, as above. Price sweep training was conducted until there was little variability in performance from one series of ten trials to the next, determined by visual inspection.

3.2.1.2 Randomized triads training

Following training, rats were presented with the randomized sweep procedure described in Chapter 2, drawing test trials without replacement from the following aggregated list of price-frequency pairs:
the 9 pulse frequencies at an objective price of 4s that were presented at the end of frequency sweep training (frequency pseudo-sweep),
the 9 prices at the highest pulse frequency the animal could tolerate that were presented at the end of price sweep training (price pseudo-sweep), and

9 prices and frequencies arrayed along a line extending from the intersection of the frequency and price pseudo-sweeps to the presumed coordinates of F_{hm} and P_e, in log-log space (radial pseudo-sweep).

The presumed coordinates of F_{hm} and P_e were estimated at first by using the pulse frequency that produced half-maximal performance during frequency sweep training (for F_{hm}) and the price that produced half-maximal performance during price sweep training (for P_e). The list of price frequencies and pulses was adjusted throughout this phase of the experiment to ensure each pseudo-sweep drove performance from maximum to minimum. Triads (leading bracket-test-trailing bracket) were presented for a maximum of 8 hours per session in this phase; in most cases, presentation of all the trials in the randomized list (a survey) required only one experimental session to complete.

Before beginning experimental sessions described below, rats were given a final set of frequency sweep training sessions, conducted as above, but with the objective price set to 0.125 seconds. This ensured that any competing motor effects did not interfere with the animal's performance at this very low price; if the maximum proportion of time an animal could allocate to self-stimulation activities in this condition was below 0.6, the duration of the blackout delay (during which the reward was delivered) was increased from 2s to 4s for only those animals in the 0.125s frequency pseudo-sweep described below.

3.2.1.3 Randomized triads procedure

In the case of rats DE1, DE3, MA5, DE7, PD8, DE14, DE15, DE19 and DE20, experimental sessions were arranged in the same triad structure as the preliminary, three-pseudo sweep randomized design: a leading bracket trial preceded every test trial, and a trailing bracket trial followed every test trial.

In the case of rats F3, F9, F12, F16, F17 and F18, test trials were chosen to

maximize a sampling of the parameter space (log-frequency and log-price) in very low price regions, in order to assess the degree to which decreases in price fail to require compensatory decreases in frequency to maintain a given level of performance. The parameters of test trials (price-frequency pairs) for rats F12, F16, F17 and F18 were sampled without replacement from a list that contained:

14 pulse frequencies at an objective price of 8s (8s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 4s (4s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 2s (2s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 1s (1s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 0.5s (0.5s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 0.25s (0.25s frequency pseudo-sweep),

14 pulse frequencies at an objective price of 0.125s (0.125s frequency pseudo-sweep),

14 objective prices at the highest pulse frequency the animal could tolerate (price pseudo-sweep), and

14 pulse frequencies and prices arrayed along a ray extending from the intersection of the 4s frequency pseudo-sweep and the price pseudo-sweep to the presumed coordinates of F_{hm} and P_e, in log-log space.

For rats F3 and F9, only 9 pulse frequencies were presented, with the central 5 presented twice as often, as described in chapter 2.

An example of the nine pseudo-sweeps from which test trials were sampled is provided in Figure 3.2. Points along any given pseudo-sweep for rats F12, F16, F17 and F18 were spaced such that the most extreme two values at each end of the sweep were spaced twice as far apart as the central ten. While the highest and lowest two frequencies might be spaced 0.1 common logarithmic units apart, the central ten would be spaced 0.05 common logarithmic units apart.

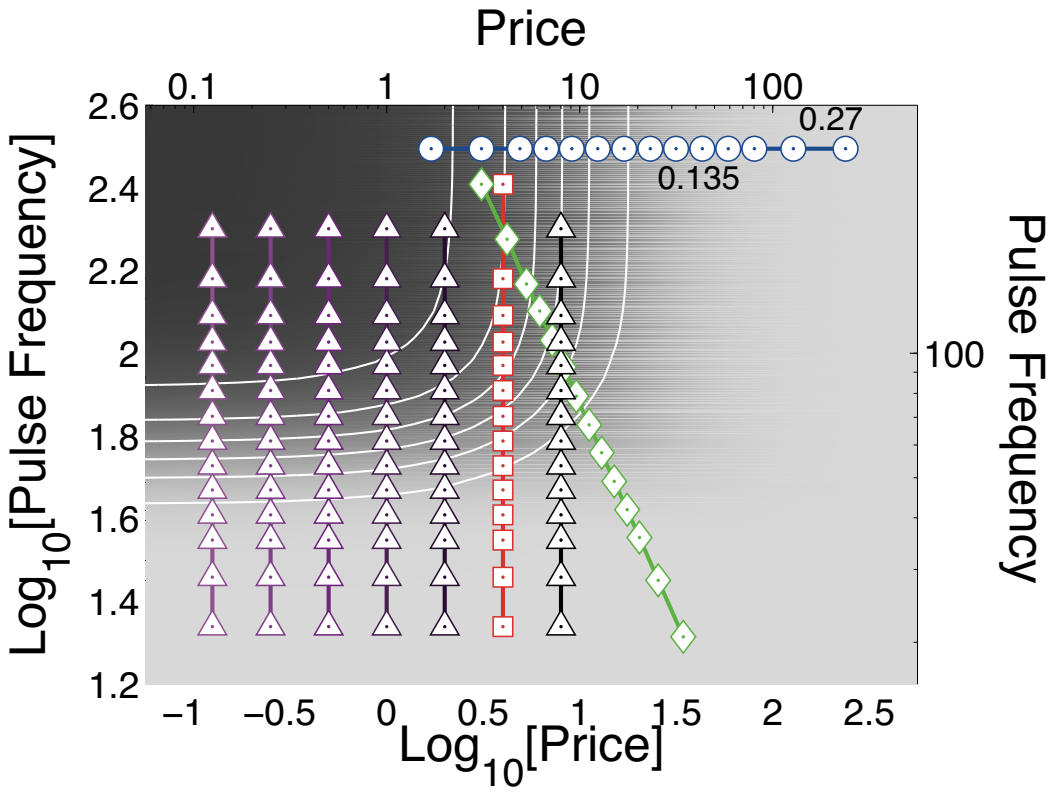

Figure 3.2. Arrangement of pseudo-sweeps in subjective-price study. For rats F12, F16, F17 and F18, price-frequency pairs were sampled at random from the vectors depicted here. Red squares indicate the 4s frequency pseudo-sweep, blue circles indicate the price pseudo-sweep, and green diamonds indicate the radial sweep. These pseudo-sweeps are analogous to the three pseudo-sweeps collected from rats in Chapter 2. Magenta triangles (from bright to black) indicate the extra frequency pseudo-sweeps collected at low (bright magenta, 0.125s) to high (black, 8s) prices. In the case of rats F3 and F9, the price-frequency pairs were obtained as in Chapter 2: nine different pairs per pseudo-sweep were obtained, with the middle 5 sampled twice as often as the extreme 4. In the case of F12, F16, F17 and F18 depicted here, 14 different price-frequency pairs were obtained from each pseudo-sweep, where the extreme 4 were twice as spread as the centre 10. In the example here, the difference between the highest and second-highest price of the price pseudo-sweep is 0.27 common log units, while the difference between the third- and fourth-highest is 0.135 common log units.

3.2.2 Mountain fit

Rats F3, F9, F12, F16, F17 and F18 encountered a large number of very low objective prices. As a result, a modified version of the mountain model was fit to the data for these animals. Rather than assume that decreases in price in the low range required compensatory increases in pulse frequency to maintain a constant level of performance, the model that was fit assumed that the relevant decision variable was the subjective opportunity cost of self-stimulation. It would indeed be unreasonable to assume that a change in objective price from 0.001 to 0.002 seconds requires the same proportional change in pulse frequency as a change from 4 to 8 seconds in order to motivate the rat to work. As a result, we assumed that the psychophysical mapping between objective prices and subjective opportunity costs is scalar at high prices and is constant at very low prices: the subjective opportunity cost of holding for 0.125 seconds would be equivalent to that of holding for 0.25 seconds, but the subjective opportunity cost of holding for 8 seconds would be twice that of holding for 4 seconds. Solomon et al. (2007) have already established that such a mapping can account for choice between two levers that deliver stimulation of differing price. The Reinforcement Mountain Model that was fit to the time allocation data generated by these animals therefore included two extra parameters: the minimum subjective opportunity cost (SPmin), below which further decreases in objective price do not change the opportunity cost, and the sharpness of the transition between this region and the scalar region (SPbend). Estimates of the parameters in the set of equations

$$P_s = SP_{min} + SP_{bend} \times ln\left(1 + e^{\frac{SP_{min} - p_o}{SP_{bend}}})\right)$$

$$I_{rel} = \frac{f^G}{f^G + F_{hm}^G}$$

$$TA = \frac{I_{rel}^A}{I_{rel}^A + \left(\frac{P_s}{P_e}\right)^A}$$

were inferred using the bootstrapping approach described in Chapter 2. Details of the subjective opportunity cost model have been previously described in Solomon et al. (2007).

The mapping between objective price and subjective opportunity cost was then inferred from parameters SP_{min} and SP_{bend} using the first equation above in rats F3, F9, F12, F16, F17 and F18. The psychophysical mapping of pulse frequency to subjective reward intensity was inferred from the parameters F_{hm} and G that were fit from the Reinforcement Mountain Model using the second equation above. Payoff was defined as the ratio of the relative subjective reward intensity (I_{rel}) to the subjective opportunity cost (P_s).

3.2.2.1 Robust Analysis of Variance

To determine whether the rats could detect the type of trial that had just begun, before any information about the stimulation strength and opportunity cost of the reward had been revealed, we extracted the duration of the first pause at the start of the trial, immediately following priming stimulation delivered during the inter-trial interval. This post-priming pause (PPP) reflects a period of time before the animal even begins to work, thereby giving a preliminary indication of the payoff the rat expects to receive on the trial to come. If the trial began with the rat holding down the lever, the post-priming pause was assigned 0 duration. Systematic changes in PPP related to the triad trial types (leading, test, and trailing) would indicate systematic differences from trial type to trial type in the animal's expectancy for the trial to come. This systematic expectancy forms the basis for what we term the rat's world model, in the sense that the rat behaves as if he has a model of the world upon which to base expectations for the payoff to come.

In some cases, there are many instances of PPPs that were censored by the end of the trial; the rat simply did not press for the duration of the trial in these

cases. As a result, we used robust estimates of central tendency and variability, and performed a one-way robust ANOVA on the PPP durations for each trial type. To calculate the robust ANOVA, PPP durations were assigned weights according to the one-shot version of Tukey's bisquare estimator (Hoaglin et al., 1983)

$$w_i = \begin{cases} 0 & \text{if } |u_i| > 1 \\ 1 - (u_i^2/c^2)^2 & \text{if } |u_i| \leq 1 \end{cases}$$

where

$$u_i = \frac{y_i - \tilde{y}}{c \times MAD}$$

using a bisquare tuning constant (c) of 4.4685, and the median PPP duration ($\tilde{y}$) of the trial type, where MAD is the median absolute deviation from the median. The resulting robust estimate of the mean PPP duration for a trial type was the weighted sum of the durations divided by the sum of the weights. The grand mean PPP duration across all trial types was the weighted sum of all durations divided by the sum of all weights.

The (robust) total sum of squared deviations of PPP durations from the grand mean PPP duration (SS_T) was, similarly, the sum of the weighted squared deviation of PPP durations from the grand mean duration. The error term is the (robust) sum of squared deviations of individual PPP durations from their mean trial type PPP duration ($SS_{S/A}$), calculated similarly to SS_T. As total variability (SS_T) can be partitioned into variability due to different trial types (SS_A) and variability due to noise in the process that generates PPP durations ($SS_{S/A}$), the difference $SS_T - SS_{S/A}$ is a measure of the variability that can be attributed to differences in trial type.

The (robust) degrees of freedom for the error term $SS_{S/A}$ is given by

$$df_{S/A} = \sum_j ((\sum_i (w_{ij}))^2 - \sum_i (w_{ij}^2) / \sum_i (w_{ij})$$

which smoothly varies the degrees of freedom in error for k groups between 0 (when all weights are 0) to $n_1-1+n_2-1+\ldots+n_k-1$ (when all weights are 1). As a result, the robust error variance ($MS_{S/A}$) is the robust $SS_{S/A}$ divided by the robust $df_{S/A}$, where $df_{S/A}$ is a real (rather than integer) number. The variance in PPP duration due to trial type is the robust SS_A divided by 2, as there were only 3 trial types. The F test then proceeded as in the non-robust case, taking the ratio $MS_A/MS_{S/A}$, using a real-valued rather than integer-valued $df_{S/A}$.

If the F ratio was found to be statistically significant at the 0.05 level, three Bonferroni-corrected *post-hoc* tests were conducted: leading trial vs test trial, test trial vs trailing trial, and trailing trial vs leading trial PPP durations. Each comparison used the robust methods described above, calculating the grand mean and the variance in PPP duration within trial type conditions only across the two groups being compared.

3.3 Results

3.3.1 Duration of the post-priming pause on each trial type

Figure 3.3 depicts a bar graph, for each rat and in each condition they encountered, of the relationship between trial type and PPP duration. On leading bracket trials, the duration of the PPP is often very short, while on trailing bracket trials, the PPP is generally censored at 25 seconds by the end of the trial. On test trials, the duration of the PPP is greater than the leading bracket trial (ranging from 0.24 to 12.25 seconds longer). The duration of the PPP on test trials is also always much shorter than on trailing trials (ranging from 6.2 to 23.7 seconds shorter). Although there is variability between animals in the mean duration of this pause, there is little variability in their pattern: on leading bracket trials, the PPP is nearly 0, on trailing bracket trials, the PPP is usually censored by the end of the trial at 25 seconds, and

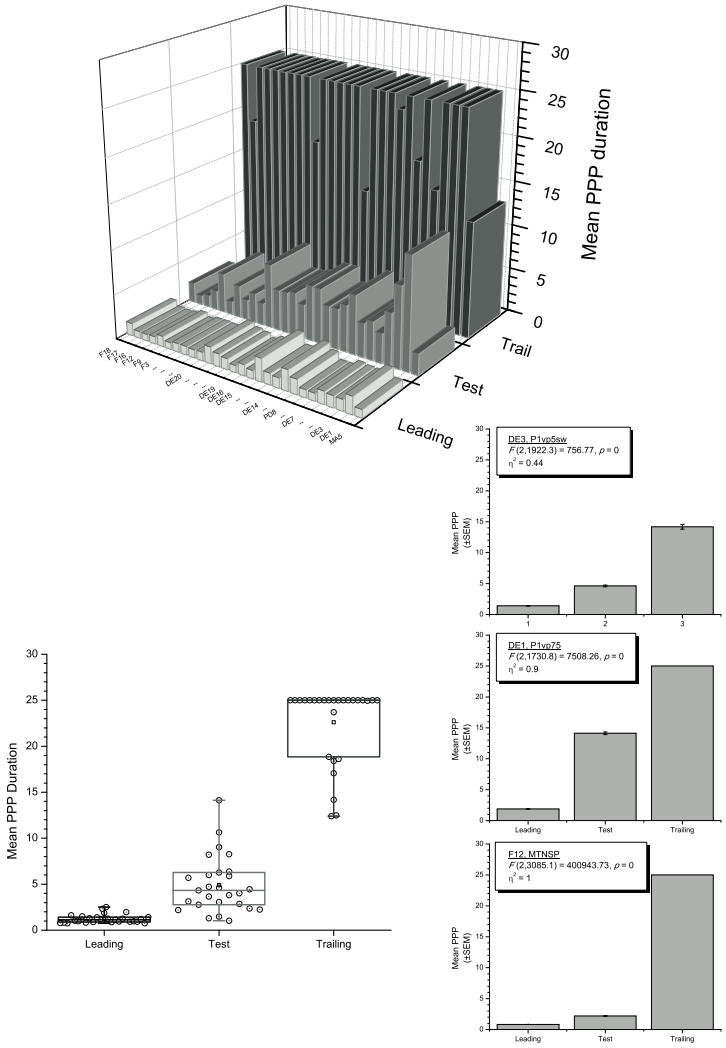

Figure 3.3. Post-priming pauses depend on triad trial type. Top panel shows, for each rat in each condition, the mean PPP on leading bracket, test, and trailing bracket trials. In every case, there is a lawful (short, medium, long) relationship between the payoff on the trial to come and the duration of the first pause taken. Lower left panel shows a box-whisker plot of the durations, with the means indicated with squares. Lower right panels show examples of the best (bottom, F12, $\eta^2 = 1$), typical (middle, DE1, $\eta^2 = 0.9$) and worst (top, DE3, $\eta^2 = 0.44$) cases in which trial type predicts the duration of the PPP in terms of their η^2 value. Bar graphs on far right show robust mean PPP durations with their associated robust standard errors for each trial type.

the PPP on test trials is somewhere between that of the two bracket trials, with a median 3 seconds longer than on the leading bracket and 19.11 seconds shorter than on the trailing bracket. Since there is no way for the rat to know the current trial type without having developed a world model of how the current trial type is related to previously seen trials, these results provide very strong evidence for the existence of a world model of the randomized trials design.

The robust ANOVA revealed a statistically significant effect of trial type on PPP duration in all 15 animals tested, and in all conditions the rats encountered ($p<0.0001$ in all cases). The magnitude of this effect, as calculated by η^2, ranged from 0.44 to nearly 1.0, indicating that the vast majority of the variance in PPP duration throughout the experiment could be accounted for by differences in trial type, for all rats. Bonferroni-corrected *post-hoc* comparisons of each test trial showed that each trial type was characterized by a distinct PPP duration, in the expected direction: PPPs on trailing bracket trials were significantly longer than leading bracket ($p<0.0001$ in all cases) or test trials ($p<0.0001$ in all cases), and PPPs on leading bracket trials were significantly shorter than those on test trials ($p<0.0001$ in all cases).

The bottom left panel of Figure 3.3 is a box-whisker plot of the mean PPP on leading bracket, test, and trailing trials. Overlain on the plot is the mean PPP for each animal in each condition encountered at the start of each trial type. The pattern seen in the aggregate is equivalent to that seen on a rat-by-rat basis: the PPP is reliably shorter on leading bracket trials than on trailing bracket trials, and though mean PPP durations vary greatly between animals on test trials, they tend to have a duration that is intermediate between leading and trailing bracket trials.

The bottom right panel of Figure 3.3 shows three examples of the mean and standard errors surrounding the robust estimates of the PPP duration on leading bracket, test, and trailing bracket trials. The topmost bar graph shows animal DE3 in

the second phase of the P1vp5 condition (during which the levers mapped to rewards delivered with certainty and those delivered with a probability of 0.5 are switched compared to the previous condition). This case represents the series of post-priming pauses for which trial type accounted for the least variance in PPP duration (η^2 was 0.44). For comparison, we also show a typical result (DE1, P1vp75 condition, centre bar graph; η^2 was 0.90) and the data set for which almost all the variance in PPP duration could be accounted for by trial type (F12, bottom bar graph; η^2 was nearly 1.00). The aggregate pattern seen in the upper and lower left panels results from the same pattern seen in each animal, with variation between subjects only in the absolute values of the different pause durations, but not their overall ordering.

If this is true, there must be some type of rule by which rats use either the price and pulse frequency in effect on the previous trial or the current position in the triad to infer the current trial's payoff. The following section will attempt to answer what kind of rule they may be using by considering the PPP on trailing bracket trials, which follows a test trial of variable price and pulse frequency.

3.3.2 Heuristic for trial position

3.3.2.1 Analysis

Two complementary strategies may be used to infer the payoff to be expected on a given trial in the triad. The first and simplest strategy would be to consider the characteristics of the trial that has just come to pass. Predicting the leading bracket trial, with its fixed high payoff, is trivial, because it always follows a trailing bracket trial, with fixed low payoff. However, avoiding this prediction following a low-payoff test trial would be difficult. Predicting the occurrence of the test trial (though not its payoff) is similarly easy, because it, too, follows a trial with a fixed payoff. With enough exemplars of test trial payoffs, an expectation can be formed

for the trial that follows a trial with exceptionally high payoff. However, avoiding this prediction following a high-payoff test trial would be difficult. Finally, although predicting the payoff of the trailing bracket trial is easy because it is fixed, the trailing bracket occurs after a trial with pseudo-randomly selected payoffs. If the rat were to use this one-trial look-back strategy, it could be misled on the trailing trial by the payoff presented on the test trial. A particularly high-payoff test trial would mislead the rat into making a post-priming pause on the next (trailing bracket) trial more characteristic of a test trial, whereas a particularly low-payoff test trial would mislead the rat into making a post-priming pause on the next (trailing bracket) trial more characteristic of a leading bracket trial.

Assuming the rat uses a single-trial look-back rule to infer the payoff on the current trial, there is one set of test trials that will lead the rat to confuse a test trial for a leading bracket trial: when the price is 1 second and the stimulation delivers a maximal reward. As a result of confusing the test trial with a leading bracket trial, the PPP that the rat takes on the trailing bracket trial should be shorter than usual. The rat will be led to believe that the current trial is a test trial when, in fact, it is a trailing bracket trial.

Similarly, the rat will confuse a test trial for a trailing trial when the price for stimulation during the test trial is 1 second and the stimulation is weak. As a result of confusing the test trial with a trailing bracket trial, the PPP that the rat takes on the trailing bracket trial should also be shorter than usual. The rat will be led to believe that the current trial is a leading bracket trial when, in fact, it is a trailing bracket trial.

A second, more cognitively demanding solution is to maintain a representation of both the trial type sequence (as above) and the animal's current position in the sequence. A rat exclusively using this second strategy may be inaccurate in its count, resulting in uncharacteristically short pauses on trailing bracket trials that are inde-

pendent of the payoff on the test trials that preceded them. Whereas using a one-trial look-back rule for retrieving stored transitions would produce systematic differences in PPP duration on trailing bracket trials as a function of test trial opportunity costs and reward intensities, a counting model would produce no such systematic differences.

Test trials that resemble the leading or bracket trials will induce an error—that is, uncharacteristically short PPPs—if the animal is using a single-trial look-back strategy. To determine whether the payoff on the previous test trial had an impact on the post-priming pause at the start of the following trailing trial, we conducted a one-way robust ANOVA on the PPP durations during trailing trials that followed test trials on which the price was 1s (the price in effect during leading and trailing bracket trials) for each particular reward intensity for each rat. One-way robust ANOVAs were also conducted on the PPP durations that followed test trials delivering the highest pulse frequency of each frequency pseudo-sweep (thereby making them similar to leading bracket trials delivering high-frequency stimulation at a 1s price), as well as all those delivering the lowest pulse frequency of each frequency pseudo-sweep (thereby making them similar to trailing bracket trials delivering low-frequency stimulation at a 1s price). Bonferroni-corrected *post-hoc* tests were conducted following each ANOVA to identify PPP durations that were significantly different from the others based on the price and frequency encountered on the previous trial.

3.3.2.2 Confusing trials

Figure 3.4 shows the mean (and standard error) estimated PPP on trailing bracket trials as a function of the subjective reward intensity of the test trial that preceded it, when that test trial delivered stimulation at a 1 second price (the price of the leading and trailing bracket trials). In all cases, the robust ANOVA is statistically significant at the 0.0001 level and accounts for a large proportion of variance in PPP

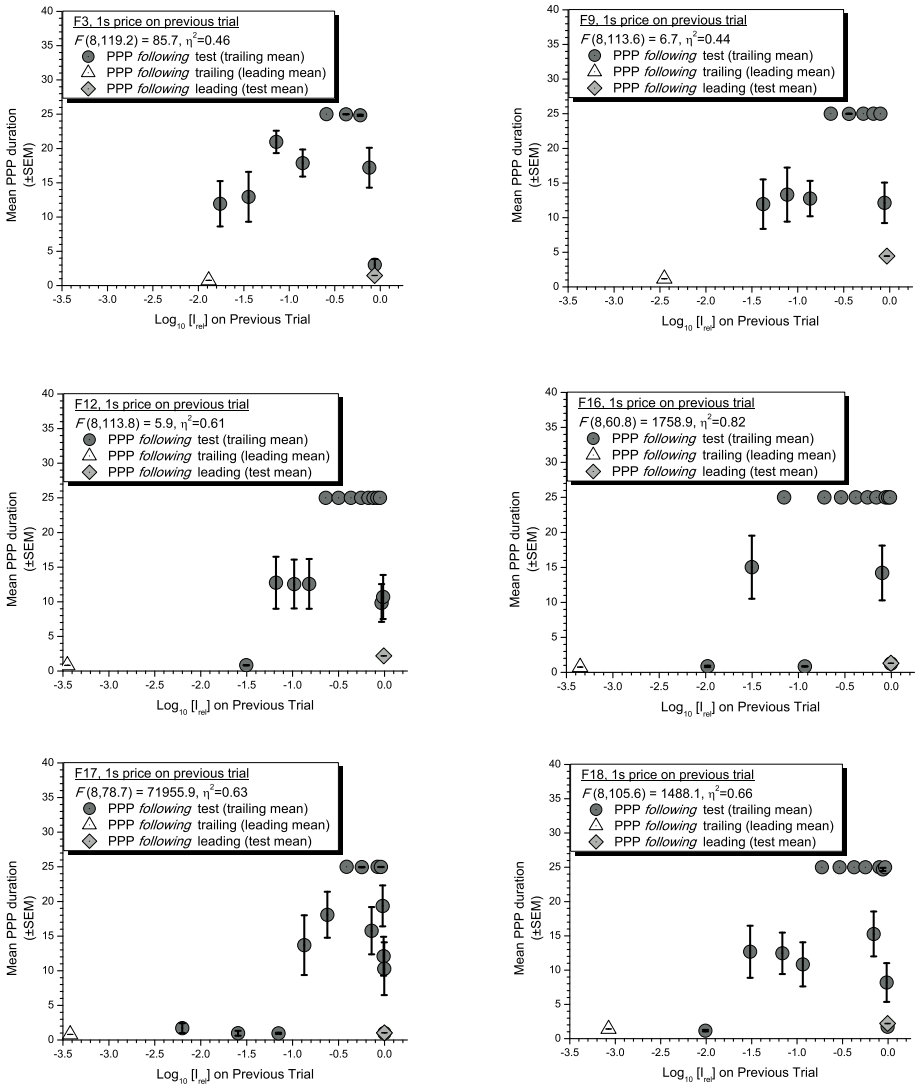

Figure 3.4. "Leading"-like and "trailing"-like test trials induce an error on subsequent true trailing trials. The mean PPP duration ($\pm SEM$) on the trailing bracket trial is plotted as a function of the subjective reward intensity in effect on the preceding test trial, when the price in effect on that test trial was 1s (circles). When the last trial had a similar price and reward intensity as a leading trial (high reward intensity, 1s price), rats make uncharacteristically short post-priming pauses on the subsequent trailing trial that are similar to those made following true leading trials (diamonds). When the last trial had a similar price and reward intensity as a trailing trial (low reward intensity, 1s price), rats also make uncharacteristically short post-priming pauses that are similar to those made following true trailing trials (triangles).

duration (η^2 ranged from 0.44 to 0.82). Most notably, when the test trial presented rewards of either negligible or nearly maximal magnitude at a 1s price, rats make reliably shorter PPPs, while at intermediate magnitudes, rats make PPPs typical of the trailing trial overall. In many cases, the PPPs produced following test trials in which reward intensity was highest ("leading"-like test trials) and lowest ("trailing"-like test trials) are comparable to those produced following true leading and trailing bracket trials. Bonferroni-corrected *post-hoc* tests on the PPP durations revealed significant differences ($p<0.0001$ in all cases) between these PPP durations following test trials delivering either near-minimal or near-maximal reward intensities at a 1 second price, and the PPP duration following the test trial which delivered an intermediate (the 5th highest in the case of rats F3 and F9, or the 8th highest in the case of rats F12 through F18) subjective reward intensity at a 1 second price. In effect, rats presented with "leading"-like test trials are misled to believe the trailing bracket is a test trial and make shorter pauses before beginning to work for stimulation they would otherwise ignore. Similarly, rats presented with "trailing"-like test trials are misled to believe the trailing bracket is a leading bracket trial and make very short pauses before working for stimulation they would otherwise ignore. In contrast, when rats are presented with test trials that are neither "leading"- nor "trailing"-like are not misled and ignore the weak stimulation they can expect to receive on trailing bracket trials.

Figure 3.5 is a depiction of the duration of the PPP on the trailing bracket trial when that trial follows a potentially misleading test trial. In the top panel, the mean PPP on trailing trials that follow a "leading"-like test trial is plotted in dark grey for each rat, along with the overall mean PPP on trailing trials (black bars) and the mean PPP on trials that follow true leading bracket trials (light grey bars). In the bottom panel, the mean PPP on trailing trials that follow a "trailing"-like test trial is plotted in dark grey for each rat, along with the overall mean PPP on trailing

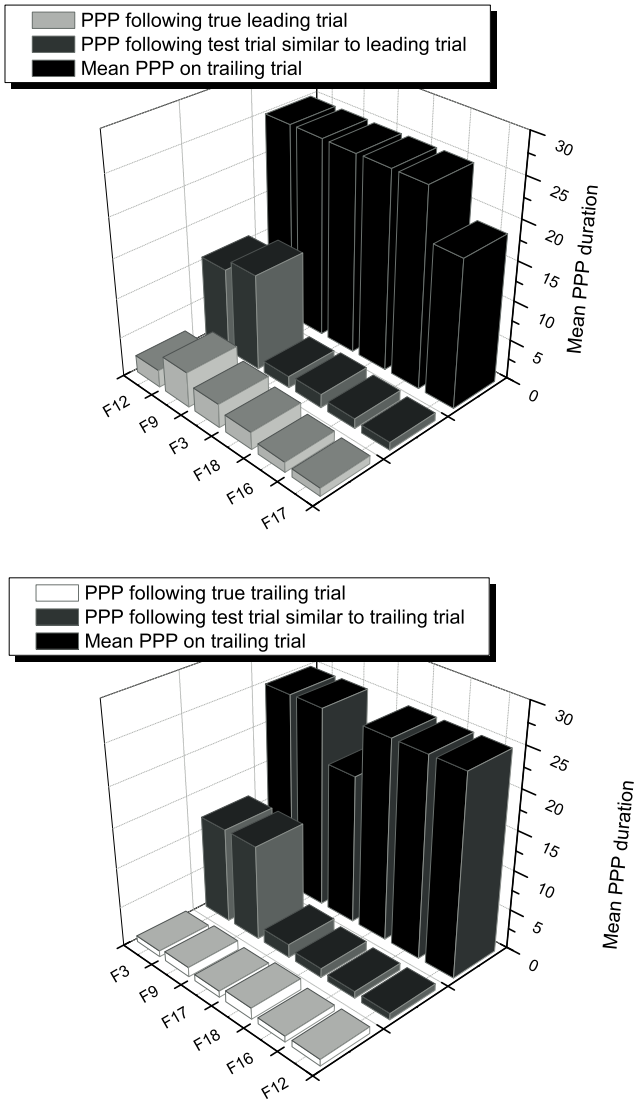

Figure 3.5. Misleading test trials induce unusually short post-priming pauses on subsequent trailing trials. In the upper panel, mean PPP duration for each animal is plotted as a function of whether it follows a true leading bracket (light), a test trial that is similar to a leading bracket (dark) or a test trial that is not similar to a leading bracket (black). In the lower panel, mean PPP duration for each animal is plotted as a function of whether it follows a true trailing bracket (light), a test trial that is similar to a trailing bracket (dark) or a test trial that is not similar to a trailing bracket (black).

trials (black bars) and the mean PPP on trials that follow true trailing bracket trials (light grey bars). In all cases, the mean PPP is uncharacteristically short, indicating some degree of confusion. In 4 of 6 cases, the mean PPP on trailing trials that follow a "leading"-like test trial is comparable to the mean PPP on test trials that follow a "true" leading bracket trial. Similarly, in 4 of 6 cases, the mean PPP on trailing trials that follow a "trailing"-like test trial is comparable to the mean PPP on leading trials that follow a "true" trailing trial.

These results are consistent with the hypothesis that rats rely a great deal on the subjective reward intensity and price of the previous trial to infer an as-yet unknown payoff. Although in 2 of 6 cases the mean PPP is substantially greater than expected when using only a single trial look-back rule, the mean PPP is still substantially lower than expected when using only a counting rule. Since these two rules are not mutually exclusive, a rat could potentially rely on both at any given time. Alternately, it is possible that the rat can discriminate between the subjective reward intensity delivered on maximally confusing test trials and the bracket trials for which they would be confused. If the rat can easily discriminate between a reward with a subjective reward intensity of 0.01 and one with a subjective reward intensity of 0.02, the test trial will not sufficiently resemble the trailing trial to mislead the rat.

If test trials that resemble in price and pulse frequency the leading (or trailing) bracket trial produce uncharacteristically short post-priming pauses, the question becomes: to what degree must the price of stimulation on the test trial differ from the leading or trailing bracket trials to induce this confusion? It is to this question that we now turn.

3.3.2.3 A Gradient

Given that rats are susceptible to confusing strong-stimulation/low-price test trials with the leading trial, it would be interesting to determine which prices induce

this confusion and which ones do not. The generalization gradient—that is, the relationship between the objective price on a test trial and its ability to confuse the rat into behaving as if the following trial is a test trial—will reflect the type of heuristic used to infer the next trial. As the rat does not have access to the pulse frequency of the stimulation, any rule for inferring the next trial type will involve the subjective intensity of the rewarding effect. However, the rat could use the objective price and subjective intensity in its table lookup strategy, the subjective opportunity cost and subjective intensity, or the payoff (the scalar combination of subjective opportunity cost and intensity) of the previous trial to infer the identity of the next trial. In essence, if the basis for determining whether the last trial was a leading bracket trial involves a comparison of the last trial's objective price with the leading bracket trial's objective price, then the gradient for generalization will be steep: if the last trial presented rewards at an objective price that is not reasonably close to one second, the rat's hypothesis that the last trial was a leading bracket trial will be accurately rejected and the rat will not confuse the following trailing bracket trial for a test trial. However, if the basis for determining the trial type of the last trial is the subjective opportunity cost of acquiring BSRs, then the gradient for generalization will be considerably shallower: if the last trial presented rewards at any of a number of objective prices that leads to the same subjective opportunity cost, the rat will fail to reject the hypothesis that the last trial was a leading trial and will therefore confuse the following trailing bracket trial for a test trial.

To investigate this generalization gradient and determine how the objective price on the preceding test trial relates to the PPP on the trailing bracket trial that follows it, we performed the same robust ANOVA on the duration of trailing trials as a function of the price of the test trial that preceded them for test trials that delivered the highest stimulation frequency available. Bonferroni-corrected *post-hoc* tests were then conducted on these data to assess which test trials were followed by

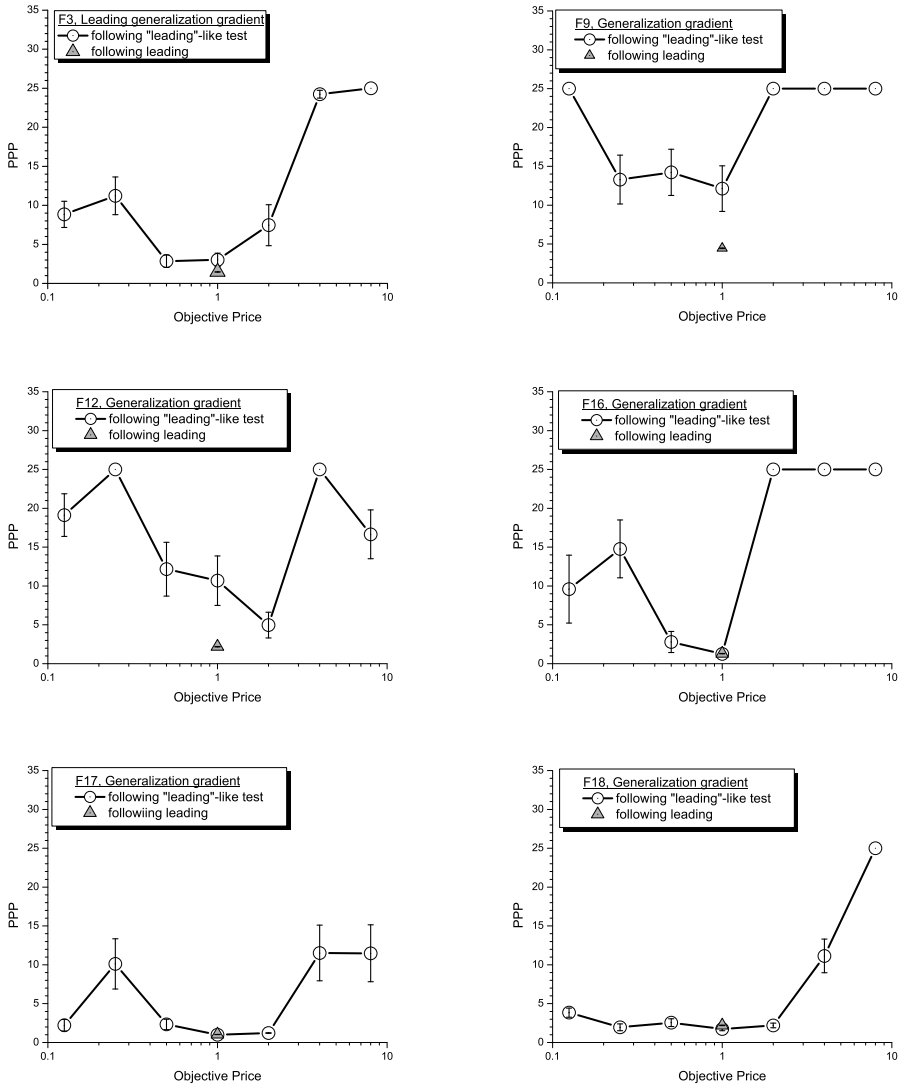

Figure 3.6. A generalization gradient of the similarity of "leading"-like test trials to true leading trials. The duration of the PPP (± SEM) taken on trailing bracket trials is plotted as a function of the objective price in effect on the preceding test trial (circles), when that test trial delivered very strong stimulation that is similar to that delivered on a true leading bracket trial (triangles).

uncharacteristically low PPPs on the trailing trial. In every case, the robust ANOVA was statistically significant at the 0.0001 level, and accounted for 44% to almost 100% of the variance in PPP duration. These results are summarized in figure 3.6.

Figure 3.6 depicts this generalization gradient for each rat by plotting the duration of the trailing trial PPP as a function of the price in effect during the "leading"-like test trial that preceded it. Figure 3.7 depicts the generalization gradient by plotting the duration of the trailing trial PPP as a function of the price in effect during the "trailing"-like test trial that preceded it. In most cases, there is a threshold-like relationship between PPP and the objective price of the preceding test trial, with uncharacteristically low PPPs below 4s and the typical, censored PPP following test trials delivering rewards at a four- or eight-second price.

3.3.2.4 Is this a stimulus or a decision variable discrimination?

It is clear from these data that the objective price is an important determinant of whether an animal that uses a single trial look-back rule will confuse the test trial with either of the bracket trials. Nonetheless, it is possible that the animal bases its heuristic not on a stimulus-discrimination rule ("was the last trial's price sufficiently close to the leading bracket's price"), but rather, on a decision variable rule ("was the last trial's subjective opportunity cost sufficiently close to that of the leading bracket's subjective opportunity cost"). Other groups have shown that rats can discriminate a 400ms tone duration from a 250ms tone duration (Kelly et al., 2006), and even random noise bursts in the 10ms to 50ms range (Pai et al., 2011). These results certainly imply that a 0.25s price can be timed fairly accurately, at least in principle. A rat that could not distinguish 2s, 1s, 0.5s, 0.25s and 0.125s intervals would certainly treat these objective prices equivalently, and would certainly be misled by a test trial delivering nearly-maximal or minimal rewards at any of those prices. Given that animals can distinguish stimuli of durations much shorter than the prices used, it is

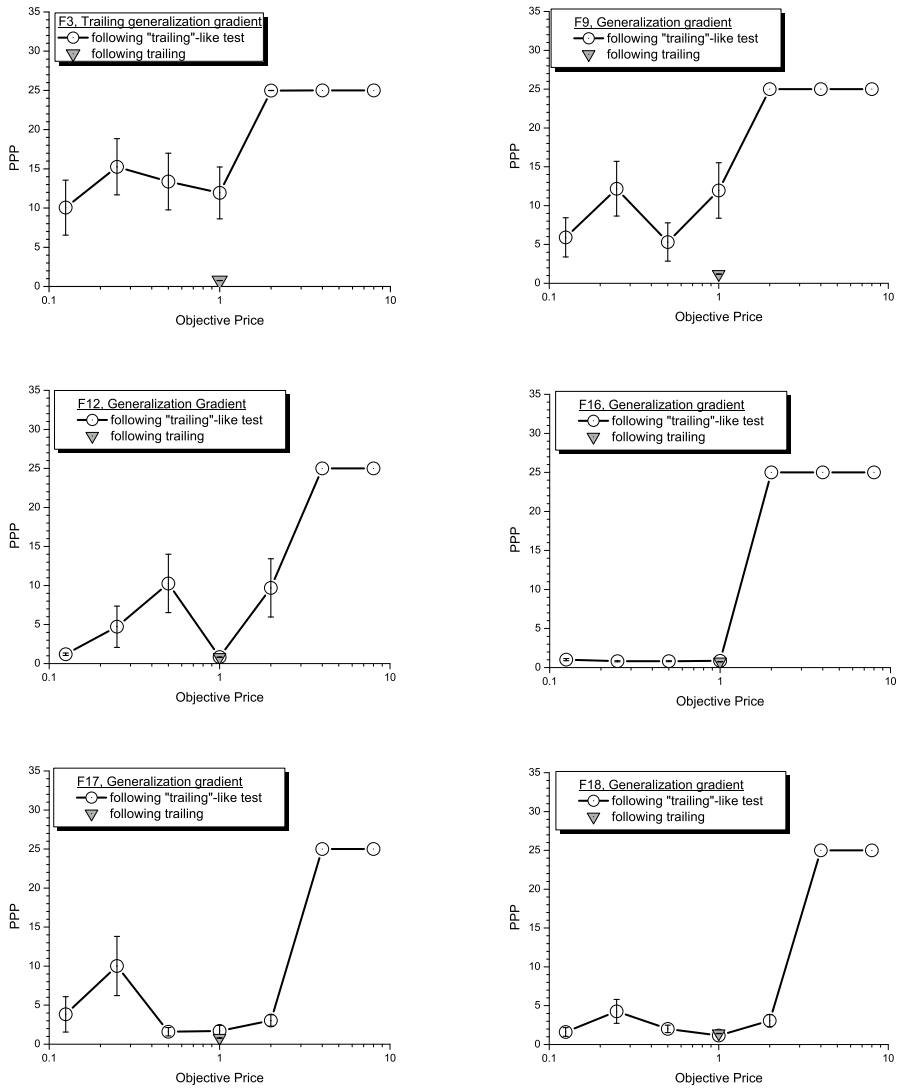

Figure 3.7. A generalization gradient of the similarity of "trailing"-like test trials to true trailing trials. The duration of the PPP (± SEM) taken on trailing bracket trials is plotted as a function of the objective price in effect on the preceding test trial (circles), when that test trial delivered very weak stimulation that is similar to that delivered on a true trailing bracket trial (triangles).

unlikely that the prices used are not discriminable. Instead, we suggest that when animals treat test trials presenting a 0.125s price as though they had just encountered a bracket trial, rats do so on the basis of equivalent subjective opportunity cost rather than the rats' inability to time short latencies.

Figure 3.8 presents the generalization gradient as a function of the deviation of the test trial's subjective opportunity cost from that of the leading (upper left) or trailing (upper right) bracket trial, rather than the objective price (lower panels). In all animals showing a clear confusion effect, the generalization gradient is consistent with a decision variable rule. Whereas there is little confusion at subjective opportunity costs much greater than that of the leading (or trailing) bracket, when the subjective opportunity costs are equivalent, rats apt to be misled by the test trial show considerable confusion regarding whether the trial to come will be a trailing bracket trial.

3.3.2.5 Is last payoff the relevant heuristic

Given that the generalization gradient is consistent with a decision variable rule ("the last subjective intensity and opportunity cost were trailing-like"), it would be tempting to conclude that rats base their decision about whether the test trial that has just occurred is a leading or trailing bracket on the payoff from the trial. The payoff is a scalar combination of subjective intensity and opportunity cost, and an even simpler rule would require a comparison of the payoff on the last trial rather than each of its individual determinants (subjective intensity and opportunity cost). If the basis for inferring that the last trial was a leading bracket depends on having a very high payoff—delivering highly rewarding stimulation at a very low cost—then a test trial delivering equivalently low-cost, high-intensity stimulation would indeed confuse the rat. However, because there is only one way for the payoff to be nearly maximal, it would be impossible to distinguish a rule based only on the payoff ("was the last

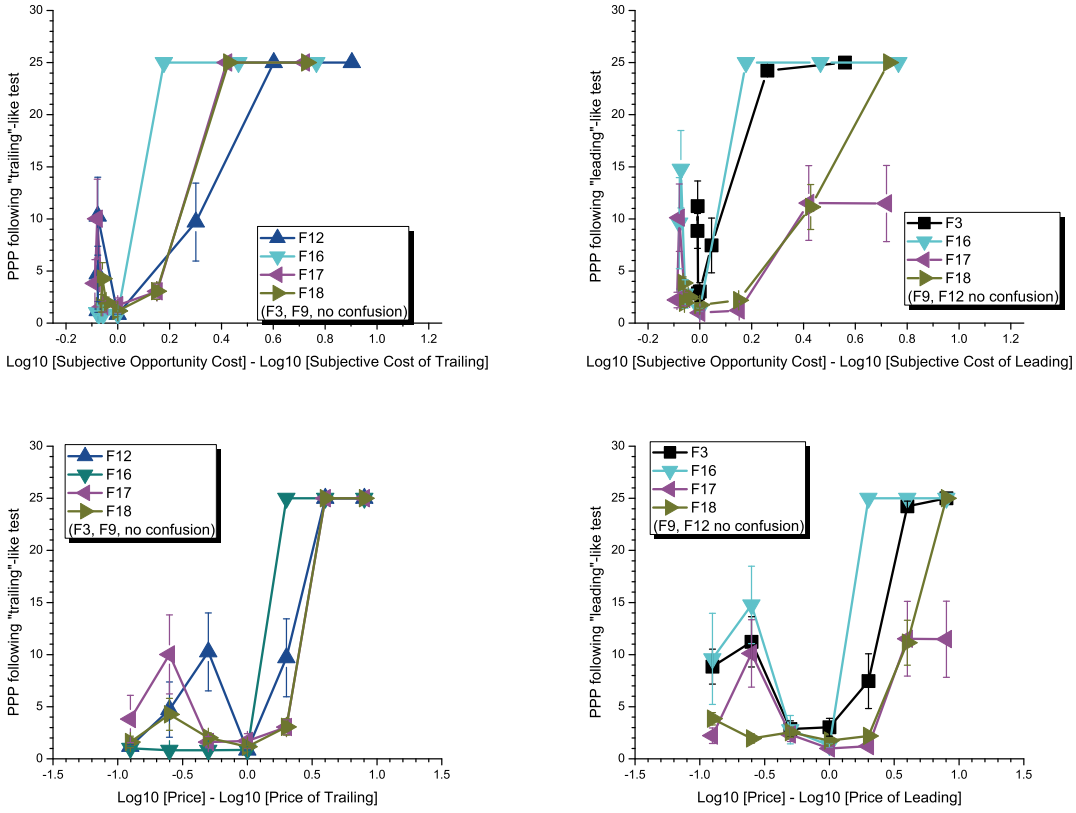

Figure 3.8. The generalization gradient for "leading"- and "trailing"-like test trials is consistent with a subjective opportunity cost discrimination. The mean PPP during trailing bracket trials is plotted as a function of the deviation of the variable on the test trial that preceded it from the variable on true bracket trials. In the top row, the variable is the subjective opportunity cost, while in the bottom row, the variable is the objective price. The left-hand column plots the PPP on trailing bracket trials that followed test trials delivering weak stimulation, as a function of the deviation of test trial variables from a true trailing bracket. The right-hand column plots the PPP on leading bracket trials that followed test trials delivering strong stimulation, as a function of the deviation of test trial variables from a true leading bracket. Different symbols and colours represent different animals.

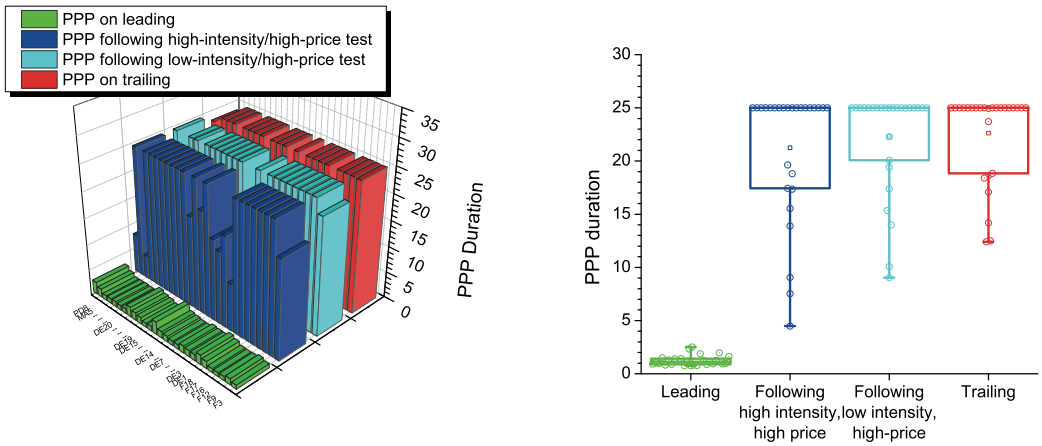

Figure 3.9. Low-payoff, but differently priced test trials do not induce confusion. The left-hand panel shows the duration of the PPP for each rat when the trial follows a true trailing trial (on leading trials, green), when the trailing trial follows a low-payoff test trial where the stimulation and price are unlike trailing trials (high intensity/high price, blue), when the trailing trial follows a low-payoff test trial where the price is unlike trailing trials (low intensity/high price, cyan), or on trailing bracket trials in general (red). The right-hand panel shows a box-whisker plot of the PPP durations on each of these kinds of trials. In all cases, the PPPs on low-payoff trials are unlike what they would be if the rats were mistaken (at left, in green) and similar to what they would be if the rats were not mistaken (at right, in red).

trial's payoff sufficiently close to that of the leading bracket trial") from one based on the appropriate combination of subjective opportunity cost and reward intensity ("were both the subjective opportunity cost and intensity of the last trial sufficiently close to those of the leading bracket trial").

Luckily, there are many combinations of intensity and price that will produce nearly minimal payoffs. One could provide inexpensive stimulation that is sufficiently weak (as is the case on trailing bracket trials), strong stimulation that is sufficiently expensive (as is the case on the highest-priced test trial of the price pseudo-sweep), or sufficiently weak stimulation that is also sufficiently expensive (as is the case on the highest-priced, lowest-intensity test trial of the radial pseudo-sweep). If the rat used a purely payoff-based rule to identify the trial it has just encountered, PPP durations on trailing trials following these low-payoff test trials would be uncharacteristically low and similar to those on leading trials. If, on the other hand, the rat used a rule based on the appropriate combination of key decision variables, these low-payoff trials will fail to confuse the rat, and the rat will produce a PPP that is typical of trailing bracket trials. To identify whether these types of test trials could confuse the rats, we plot in the left-hand panel of figure 3.9 the duration of the PPP on trailing trials that follow high-intensity but high-priced test trials (the highest-priced test trial of the price pseudo-sweep) and trailing trials that follow low-intensity and high-priced test trials (the highest-priced test trial of the radial pseudo-sweep). For comparison, we include the mean PPP duration on leading trials that follow low-intensity and low-priced trailing trials, and the mean PPP duration on trailing trials following all test trials. In all but two animals (in one condition each, for only one of the two trial types), low payoff trials with subjective opportunity costs and reward intensities that are different from those on the trailing bracket trial induce no confusion about whether the trial to come is a trailing trial.

The right-hand panel of figure 3.9 shows a box-whisker plot of the mean PPP

duration on leading bracket trials, trailing bracket trials that follow a test trial with highly rewarding but expensive stimulation, trailing bracket trials that follow a test trial with low intensity and high price stimulation, and trailing bracket trials across all test trial types. Overlain on the box-whisker plot are individual mean PPPs observed on each trial type for each rat and condition they encountered. The range of mean PPP durations on trailing trials following these two kinds of low-payoff trials is no different from those on trailing trials in general, and differs considerably from those on leading trials, suggesting little or no confusion about whether the trial to come will be a trailing bracket trial. Therefore, the rat uses the subjective intensity and opportunity cost individually—not their combination into the payoff—to infer the identity of the trial to come.

3.4 Discussion

3.4.1 A world model of session structure

Model-free temporal difference reinforcement learning models imply that rats learn only the value of a state rather than the full state-transition function. In other words, the rat using only a model-free learning scheme knows at best that conditions have changed when the inter-trial interval begins, but because such a rat does not maintain a model of how the session progresses, it cannot know in what way they have changed. Consequently, the rat that relies on a model-free learning scheme will not act on an expectation for the trial to come. When no representation of the identity of states is maintained, the rat's performance must be based only on feed-back mechanisms that map states to their total net discounted reward. In this case, the rat will expect that conditions on a new trial will be the same as those on the previous trial; as a result, the duration of the PPP would be longest on leading trials (because they follow a trial of low payoff), shortest on test trials (because they follow a trial

of high payoff) and intermediate on average on trailing trials (because they follow a variable trial). When a world model is maintained, the rat's performance can be based on what states it can expect to follow from actions taken in the current state, rather than just their value. In other words, a world model of the triad sequence allows performance to operate on a feed-forward mechanism: the rat has an expectation for the trial to come before it obtains evidence of the consequences of its actions. For example, a representation of "trailing bracket follows test trial" allows the rat to flexibly alter the value of lever-pressing without having to uncover that value through trial and error, and without having to rely on what the value of lever-pressing has been so far. Indeed, the predictions of a model-free learning scheme are opposite to the results reported here: the rat takes pauses, before the payoff can be known, that are indicative of the payoff to come rather than the payoff on the trial that has come to pass.

Our results imply that well-trained rats behave as though they had a model of how triads progress based on a look-back rule to the conditions (subjective intensity and opportunity cost) of the preceding trial. We have presented evidence that rats indeed form this world model of the triad structure, and a potential heuristic rule that rats use to infer the current trial type. Post-priming pause durations vary systematically depending on the trial type to come. Since the price and pulse frequency on a new trial are not signalled at trial onset, the duration of the post-priming pause must reflect some world model of how the trial types of a triad progress. That the duration of the post-priming pause on trailing trials, though usually censored by the end of the trial, can further be related to the price and intensity of the preceding test trial provides some evidence about the nature of this world model. A rat that simply counted would be expected to show uncharacteristically short PPPs on trailing trials regardless of the trial type—it is equally likely to make errors in counting no matter what the test trial's objective price and subjective reward intensity happen to

be. Instead, rats produce uncharacteristically short PPPs following test trials that closely resemble the leading bracket trial in terms of subjective reward intensity and subjective price. Moreover, the generalization gradient for how closely a test trial must resemble a leading bracket trial in order to "confuse" the rat suggests that it is on the basis of the subjective opportunity cost and reward intensity that the rat infers which trial type has just occurred and which trial type will follow. Finally, the discriminative stimulus used to infer the payoff on the trial to come is largely a vector of the subjective opportunity cost and reward intensity encountered on the previous trial, rather than their scalar combination (the trial's payoff). Animals do not confuse test trials delivering very low payoffs, resulting from subjective opportunity costs and reward intensities that differ from the trailing bracket trial, with a trailing bracket trial, though they do confuse test trials on which the subjective opportunity costs and reward intensities are similar to trailing bracket trials with trailing bracket trials.

These results are inconsistent with a purely model-free description of performance on this task. One would have to assume one of two results in the purely model-free approach. The rat could maintain a running average of the net expected value from pressing across all trials, in which case the rat reaches the end of the trial without an accurate estimate of the net expected value of lever-pressing. Results from chapter 2 allow us to rule out this interpretation, since performance is highly payoff-dependent. Otherwise, if the rat manages to update the value of lever-pressing to the "correct" level by the end of a trial, a purely model-free approach would assume that the payoff from self-stimulation expected on trial t is the payoff from self-stimulation encountered on the previous trial, such that when the rat begins a trailing bracket trial, it expects to obtain the same reward it had received on the previous test trial. When the rat begins a leading bracket trial, it expects to obtain the same reward it had received on the previous trailing trial. Clearly, the pattern of PPP durations is inconsistent with this: rats behave as though they expect to obtain a reward they

have not yet seen. Instead of the behavioural inertia predicted by a purely model-free approach, we observe a striking prescience in the rats' PPP that can be attributed to the deterministic progression of leading bracket, test, and trailing bracket trials. The payoff they expect to receive is a reflection of a one-trial look-back rule in a world model that encapsulates a simple set of syllogisms: if the last trial was sufficiently similar (in subjective opportunity cost and reward intensity) to a leading bracket trial, the next trial is likely a test trial; if the last trial was sufficiently similar to a trailing bracket trial, the next trial is likely a leading bracket trial; if the last trial was different from a leading or trailing bracket trial, the next trial is likely a leading bracket trial. Although basing one's decision to begin lever-pressing on this set of syllogisms may sometimes lead to an error, it is considerably less demanding to implement than a counting rule that requires maintaining an abstract representation for the counting index while performing an unrelated task. Even following months of exposure to the randomized-triad design, correct detection of the trailing trial's occurrence when it follows a test trial that is sufficiently similar to either bracket trial is a very rare occurrence. In a majority of cases, we observe PPP durations that are consistent with a simple, one trial look-up rule.

Rats appear to behave as though they have the world model depicted in figure 3.1. On trial $t-1$, rats maintain a representation (S) of the subjective opportunity cost (P_s) and reward intensity (I_{bsr}) for the trial. When the inter-trial interval begins, rats infer the subjective opportunity cost ($\hat{P}_s$) and reward intensity (I_{bsr}) of the trial to come (S'). We make this assumption because rats will readily identify the leading bracket trial following the trailing bracket trial even though the rat rarely ever obtains a sample of the subjective opportunity cost and reward intensity on trailing bracket trials. The inference is made on the basis of a simple state-transition function:

if $S_t = [\hat{P}_{s_t}; \hat{I}_{bsr_t}]$ was sufficiently similar to the vector

$$\begin{bmatrix} 1s \\ I_{max} \end{bmatrix}$$

then $S'_{t+1} = [\hat{P}_{s_{t+1}}; \hat{I}_{bsr_{t+1}}]$ will be

$$\begin{bmatrix} \bar{P}_s \\ \bar{I}_{bsr} \end{bmatrix}$$

(where $\bar{P}_s$ and $\bar{I}_{bsr}$ represent the mean opportunity cost and reward intensity on test trials); if $S_t = [\hat{P}_{s_t}; \hat{I}_{bsr_t}]$ was sufficiently similar to the vector

$$\begin{bmatrix} 1s \\ I_{min} \end{bmatrix}$$

then $S'_{t+1} = [\hat{P}_{s_{t+1}}; \hat{I}_{bsr_{t+1}}]$ will be

$$\begin{bmatrix} 1s \\ I_{max} \end{bmatrix};$$

if $S_t = [\hat{P}_{s_t}; \hat{I}_{bsr_t}]$ was neither $[1s; I_{max}]$ nor $[1s; I_{min}]$, then $S'_{t+1} = [\hat{P}_{s_{t+1}}; \hat{I}_{bsr_{t+1}}]$ will be

$$\begin{bmatrix} 1s \\ I_{min} \end{bmatrix}.$$

The expected payoff ($\mathbb{E}[U_t]$) on trial t—which involves a scalar combination of the relevant key decision variables—can concisely inform the rat of how long its PPP should last. The PPP is terminated (except on non-ambiguous trailing bracket trials) with a lever-press, at which point the rat can update the expected subjec-

tive opportunity cost (revising $\hat{P}_{s_{t+1}}$), which itself allows the rat to also update the expected payoff from self-stimulation for the trial. A reward may then be delivered, which allows the rat to update the subjective reward intensity it can expect to receive for lever-pressing (revising $\hat{I}_{bsr_{t+1}}$), and simultaneously allows the rat to update the expected payoff from self-stimulation once again. On unambiguous trailing bracket trials, the rat usually never presses, so the subjective opportunity costs and reward intensities have not been updated since their expectation ($[\hat{P}_{s_{t+1}}, \hat{I}_{bsr_{t+1}}] = [1s, I_{min}]$) at the beginning of the trial $(t + 1)$, and the rat can still infer that the next trial, S'_{t+2}, will be a leading bracket trial. For example, suppose you were asked to identify whether a dim light appeared red when it is briefly flashed. Following many trials, you recognize that after a red light, you will be presented with a violet light of varying red and blue hues; after a violet light, you will be presented with a blue light; and after a blue light, you will be presented with a red light. If you had blinked and missed the dim blue light, you would still be able to infer that the next trial would be red. We propose that the state transition function, formalized by the above if-then statements, provides the rat with a cached expectation of subjective opportunity cost and reward intensity which can be revised when the animal begins to lever-press and obtains an electrical reward. If no revision is made, the rat simply uses this cached vector to infer the next trial type on the basis of the state transition function, which then updates the subjective opportunity cost and reward intensity to those predicted for the next trial when the current trial is over. This differs from counting in the sense that it is not necessary to maintain a representation of the number of trials in a cycle and the current phase; instead, the rat only needs to maintain a representation of the opportunity cost and reward intensity of the current trial type and their relationship to those that can be expected when the house light flashes again.

3.4.2 How does the rat learn this world model?

Given that when it first encounters the randomized-triad design, the rat cannot have a world model of how leading bracket trials lead to test trials, test trials lead to trailing bracket trials, and trailing bracket trials lead to leading bracket trials, the question becomes: how does this world model arise throughout training?

One possibility is that rats use some variation of the hidden Markov model (HMM) in which the trial types are latent (or "hidden") states which present the rat with subjective opportunity costs and reward intensities according to some underlying emission probability. The rat's task is to then infer the hidden state that emitted the observable decision variables on every given trial. One problem with such an approach is that it is computationally intensive, requiring infinite memory of the entire sequence of subjective opportunity costs and reward intensities encountered on each trial thus far, and requiring the rat to identify across a large set of possible hidden state sequences that which is most likely. Furthermore, because the number of hidden trial types, their state-transition function, and the probability that they "emit" subjective opportunity costs and reward intensities is not known, the animal must estimate these model parameters on-line. To our knowledge, no group has provided a truly on-line description of how such a model could be learned.

This description of how the world model is learned must take into account three ideas, all of which must operate simultaneously. First, the rat must estimate how many trial types there are. If there are k latent states which present opportunity costs and reward intensities with some emission probability, the number of underlying latent states k needs to be estimated. Second, the rat must estimate the mean and variance of the key decision variables that identify a given trial type, which set the emission probabilities with which trial types produce subjective opportunity costs and reward intensities. Finally, the rat must estimate the state-transition function,

which sets the probability that one trial type proceeds to another. Each of these estimates is fundamentally interconnected; without some estimate of how many trial types there are, there is no way to know what types of subjective opportunity costs and reward intensities they will present or their progression. Regardless of how the learning process is modelled, it must reflect the animal's remarkable capacity to infer a fairly complex world model based only on the sequence of subjective opportunity costs and reward intensities, using limited mnemonic resources.

3.4.3 Representing the world model

Representation of this world model is an equally non-trivial question. Our results suggest that the rat maintains a representation of the subjective opportunity cost and reward intensity, which form the basis of the compound "stimulus" that the rat can use to infer the payoff on the next trial. In the case of the leading and trailing bracket trial, there is a single subjective opportunity cost and a single subjective reward intensity for which a representation is needed. However, in the case of the test trial, the animal will encounter any of a range of opportunity costs and reward intensities sampled pseudo-randomly from a finite set. In a strict sense, a model-based reinforcement learning model would assign each combination of subjective reward intensity and opportunity cost a state. The rat must then maintain a large state space and a complex state-transition function whereby the trial state corresponding to the leading trial leads to any of the possible trial states with equal probability, the trailing trial state leads to the leading trial state with certainty, and all possible trial states lead to a trailing trial with certainty.

Instead of this potentially cumbersome and computationally intensive scheme, we propose that the rat maintains three representations and a simple state-transition function. One trial state is the leading bracket trial, for which the rat maintains a representation of high-intensity, low-cost stimulation with minimal variability. One

represents the trailing bracket trial, for which the rat maintains a representation of low-intensity, low-cost stimulation with minimal variability. Finally, one represents the test trial, for which the rat maintains a representation of the central tendency of subjective reward intensity and opportunity cost along with an estimate of the variability in intensity and cost. The transition function here is a simple permutation matrix of the three trial types represented: a leading bracket is followed by a test trial, a test is followed by a trailing bracket trial, and a trailing is followed by a leading bracket trial.

Some modellers (Ma et al., 2006; Pouget et al., 2000; Deneve et al., 2007) have argued that populations of neurons represent a probability distribution over stimulus values, and that in neurons with Poisson noise, Bayesian inference reduces to a simple linear combination of population activities. In our procedure, the rat must identify the upcoming trial type stimulus (s) given various cues (c), such as the subjective opportunity cost and reward intensity of the trial the rat has just encountered. Assuming a flat prior and a Gaussian likelihood function for the probability of the cues given a trial type stimulus, $\mathcal{P}[c_1, c_2|s]$, the reciprocal of the variance (in other words, the precision) of the posterior distribution of the trial type stimulus given the cues , $\mathcal{P}[s|c_1, c_2]$, is the sum of the precisions of each likelihood function:

$$\frac{1}{\sigma^2_{s|c_1,c_2}} = \frac{1}{\sigma^2_{c_1|s}} + \frac{1}{\sigma^2_{c_2|s}}$$

If two populations of neurons now encode the cues as firing rates r_1 and r_2 using gains g_1 and g_2 with Poisson-like noise, the sum of the population responses, $\mathcal{P}[r_1+r_2|s]$ will also have a Poisson-like distribution with variance proportional to the sum of the gains. As a result, the precision of the sum of the population responses, $\mathcal{P}[r_1 + r_2|s]$, will equal the sum of the precisions of each population $\mathcal{P}[r_1|s]$ and $\mathcal{P}[r_2|s]$, implying that the sum of the population responses can indeed encode the

posterior probability of the trial type stimulus given the sum of the response rates. Ma et al. (2006) have established that under a wide variety of conditions, such as non-Gaussian and non-translation invariant tuning curves, a linear combination of various populations produces a distribution with properties identical to the posterior distribution. With such an encoding scheme, representing the relevant statistical properties of leading bracket, test, and trailing trials in the cortex is inherent in how the "product of experts" (the linear combinations of multiple contributing sub-populations) naturally represents the posterior probability of trial type given cues.

3.4.4 The first reward encounter

The findings here paint an interesting picture of the period of time before the payoff from self-stimulation can be known with certainty. The rat appears to engage in a leisure bout that is dependent on the payoff expected to come—short if the payoff is exceptionally high, long if the payoff is exceptionally low, and intermediate (though short) if the payoff is, on average, intermediate. Since the rat could not know what the payoff from self-stimulation will be without some model of how the different trial types progress, our results imply that, from the very start of the trial, the rat has, either directly or indirectly, a representation of the expected payoff.

If the rat begins the trial with an expectation of the payoff from self-stimulation, then as soon as the rat begins holding the lever, it ought to revise that payoff. On true leading bracket trials, no revision will be necessary, as there is no variance in the decision variables that contribute to the payoff from self-stimulation. On true trailing bracket trials, the rat virtually never presses, so no update can be performed. On test trials, and trailing trials following test trials resembling either bracket, the rat ought to continuously update its estimate of the subjective opportunity cost as it holds the lever for longer periods of time. As a result, the estimate of the payoff ought to improve as the rat continues to lever-press: after 6s of pressing, the rat

knows the subjective opportunity cost must be at least that of 6s, which narrows the set of possible prices. As soon as the rat has completed the response requirement, a reward is delivered (on non-probabilistic trials), which then makes the payoff entirely deterministic.

Two issues are of importance here. First, the rat may want to significantly reduce its uncertainty about the expected payoff on a given trial. This is not a concern on leading and trailing bracket trials, where uncertainty surrounding the payoff would be virtually zero (as suggested by the exceptionally short and low-variance PPP on leading bracket, and typically long censored and low-variance PPP on trailing bracket trials). On test trials, however, where the rat can expect to encounter any one of a fairly large range of objective prices (roughly 2.5 common logarithmic units in the case of F rats) and subjective reward intensities (roughly 1 common logarithmic unit), the rat could indeed be driven to reduce this uncertainty so that it may better exploit the alternative providing the greatest net return on its time investment. If there is such a principle at work, the usually competing interests of exploration—seeking out the payoffs from the different sources of reward that may be available—and exploitation—pursuing the goal that will provide the greatest payoff—are aligned. On bracket trials, where exploration can be thought of as negligible, the rat clearly exploits the most attractive alternative: BSR in the case of leading, and "other" activities in the case of trailing bracket trials. On test trials before the payoff is known with any certainty, exploration is synonymous with pursuing the goal of acquiring BSRs, or exploitation. This principle may explain why PPPs on test trials are much closer in duration to those emitted on leading bracket trials than simply the halfway point between leading and trailing bracket trials.

Second, the rat may, at least in principle, have a virtually error-free representation of the payoff for the trial as soon as it has earned a reward. If this is true, the rat does not need to update its estimate of the payoff in the slow, incremental

manner that is described in model-free reinforcement learning models. It may—as we shall demonstrate in Chapter 4—have yet another world model of how rewards progress within a single trial: from the time an inter-trial interval ends to the time a new inter-trial interval begins, the payoff from self-stimulation will be constant. With such a model, the rat need not slowly update the values of lever-pressing states leading up to reward delivery in a model-free reinforcement-learning process. The rat only needs to keep track of the payoff from self-stimulation it can transparently obtain from the subjective opportunity cost and reward intensity of the first reward it receives in a trial. As a result, the period of time in the trial before the delivery of the first reward is, in a sense, distinct from the period of time in the trial following the first reward delivery. It is to this question that we shall return in the next chapter.

3.4.5 Final remarks

Our results pose certain constraints on the neural machinery that underlies action selection. First, we would expect to find neurons involved in the representation of not only the key decision variables controlling performance, such as subjective opportunity cost and reward intensity, but also in the representation of trial type. Second, the populations of neurons involved in representing trial type would need to encode trial type as particular combinations of key decision variables. Finally, there must exist a mechanism by which the previous trial type provides a signal for the trial type to come, and that signal may require computation of the expected payoff from lever-pressing in order to choose a PPP of appropriate duration. In essence, the rat needs machinery that can implement the world model by which they appear to base their decision to begin pressing following a cued inter-trial interval.

It has been argued (McDannald et al., 2012) that neurons in orbito-frontal cortex participate in model-based learning mechanisms. If so, a rat devoid of its orbito-frontal cortex would presumably be incapable of forming this world model,

and would thus produce PPPs that do not systematically vary with the identity of the upcoming trial in the randomized-triad design. However, even if this were the case, there are multiple means by which to interfere with the apparently model-based performance we describe here. Neurons of the orbito-frontal cortex may indeed be involved in implementing the state-transition function or representing the various trial states. Additionally, they may be involved in representing the payoff expected on the trial (FitzGerald et al., 2009), or they may implement the mechanism by which the key decision variables are updated throughout the trial (Noonan et al., 2010). Lesions to rat orbito-frontal cortex would not provide sufficient evidence that the region is involved in the implementation of model-based learning in this paradigm, since interference with any of these functions would result in altered PPP durations. Similarly, ventral striatum has been implicated in a wide range of action-selection tasks (reviewed in van der Meer and Redish, 2011; McDannald et al., 2011; Yacubian et al., 2007; Prévost et al., 2010; Beyene et al., 2010), but it would be impossible to tell, on the basis of lesion studies alone, what its role would be in the model-based decision-making apparent in selecting PPP durations.

Electrophysiological recordings may provide complementary evidence about what role, if any, the orbito-frontal cortex and ventral striatum play in selecting a PPP duration. Neural correlates of the subjective opportunity cost to come, regardless of subjective reward intensity, or of the subjective reward intensity, regardless of subjective opportunity cost, active before the rat has had a chance to update these values, would provide evidence of how and where the expected trial parameters are encoded. Furthermore, if activity in these neurons changes as a function of our presumed updating process—that is, those encoding subjective opportunity cost changed in activity with ongoing lever-pressing and those encoding subjective reward intensity changed in activity with ongoing reward delivery—then the same populations that encode the most recent estimates would provide the appropriate basis for inferring

the next trial type, as predicted by the model laid out in figure 3.1. Importantly, if there are populations of neurons in charge of implementing the rule as we have described it, their activity should accurately anticipate the subjective opportunity cost and reward intensity that would be predicted to follow the previous trial during the inter-trial interval.

New opto-genetic methods (Boyden et al., 2005; Han, 2012) have emerged that permit causal manipulations to be made at time scales as fine as discussed here. Neurons involved in the formation of a world model of the session's triad structure would be prevented from doing so when they are specifically silenced in the training phase of the experiment. In contrast, if one were to silence the neurons involved in switching between maps of trial types (that is, those that implement the syllogism), the rat would be expected to emit a PPP that is more typical of the last trial's post-reinforcement pause than the new trial type. In other words, if the rat is prevented from updating its world model following an unambiguous test trial, the first pause it makes when the inter-trial interval ends and a trailing bracket trial begins will be typical of the first pause it takes following lever extension on the previous trial rather than of the first pause it takes at the start of a trailing bracket trial.

Irrespective of which structures are involved in the model-based elements of the randomized-triad design and how they may contribute, our results provide evidence that well-trained rats behave as though they have developed a non-trivial model of how trials in the triad proceed. In light of these findings, it is quite possible that under slightly different conditions, when there more than 3 different trial types, rats are capable of inferring at least the payoff for the trial to come. It is quite possible, indeed, that when there are many trial types in simple ascending-pattern (as is the case on progressive ratio schedules of reinforcement) or descending-pattern (as is common in the curve-shift paradigm) sequences, rats easily learn a world model in which there are only two rules. If the last trial delivered rewards at a sufficiently high subjective

opportunity cost or with sufficiently negligible intensity, the next trial will deliver rewards appropriately minimal opportunity cost or maximal intensity. If not, the next trial will deliver rewards at a slightly greater cost or lower intensity. In both this systematic-sequence and the randomized-triads designs, the state-transition function can be thought of as a simple permutation matrix for which the trial encountered at time t is a deterministic function of the trial encountered at time $t-1$. Assuming rats have a world model during these systematic sequences that closely resembles (at least in spirit) the world model depicted in figure 3.1, our model makes two empirically-testable predictions. First, the subjective opportunity cost and reward intensity trials in the sequence are cached and updated only when the rat lever-presses. When the rat does not press, it does not update these values, and they remain unchanged from the cached value predicted from the state transition function. The sequence of trials can still progress according to the deterministic world model in well-trained rats, thereby allowing them to predict with reasonable accuracy when the sequence will return to its highest-payoff value even when they have not pressed in multiple trials. Second, assuming a single trial look-back rule for the randomized-triads design, inserting a probe trial with particularly low payoff anywhere in the sequence will "trick" the rat into believing the next trial will have a high payoff, regardless of their current position in the repeating sequence. It remains to be seen whether the principle of a world-model with a single trial look-back rule operates under a wider variety of testing procedures, or whether it is an artefact of the randomized-triads design. Nonetheless, we believe the principles at work can be generalized to multiple contexts. Further studies are needed to establish the conditions under which rats develop a world model such as that observed here. For example, rigorous simulation studies of our proposed world model would have to be conducted to deeply understand the implications, sufficiency, and predictions of our account.

In conclusion, our results provide evidence of the existence of a trial-type

based world model in the randomized-triad design, as well as a potential heuristic by which animals would implement this world model. The apparent behaviourally-derived rules used to implement this world model provide some constraints on how the brain represents its various pieces. We propose that the key decision variables in effect on a particular trial are inferred on the basis of a single trial look-back rule, and that these key decision variables are updated throughout the trial if necessary, a process which may or may not involve model-free reinforcement learning systems.

It is clear that the well-trained rat begins a trial with an expectation of what the payoff will be in the randomized-triads experimental procedure, and at least during test trials, will have to revise its estimate as it acquires more information about the subjective opportunity cost and reward intensity. Although the rat may have a good estimate of the type of trial it can reasonably expect based on the trial that just elapsed, there will be variability in the payoff that is actually delivered on test trials. The rat can update its estimate of the subjective opportunity cost of the reward as it holds down the lever; as it harvests rewards, it can similarly update its estimate of the subjective reward intensity. As the payoff is the scalar combination of reward intensity, opportunity cost, probability, delay, and effort cost, the payoff can also be revised as the rat obtains more information about the determinants of decision-making. We shall now turn to the rapidity and time scale over which this update occurs in the next chapter.

Chapter 4

The rat's world model of trial structure

4.1 Introduction

Matching refers to the general observation that the relative rate of responding for one option compared to its alternatives is equal to the relative rate at which that alternative provides reinforcement (Herrnstein, 1961). On concurrent variable-interval schedules, for example, pigeons will typically allocate their pecks to a given pecking key in proportion to the relative rate at which that manipulandum delivers grain pellets. This is a near-optimal strategy in the sense that other strategies will not provide substantially higher overall rates of reinforcement (see Heyman and Luce, 1979, for a description of the true maximizing strategy); it is nearly optimal because the variable-interval holds the reward for the animal indefinitely as soon as it is armed. If one pecking key provides rewards at an average rate of 1 per second, and the alternative pecking key provides rewards at an average rate of 1 per 3 seconds, on average, after pecking at the first key for 3 seconds, the second key will be armed and ready to deliver a reward. As the probability that a reward will be waiting for the animal increases as it pursues other goals, the optimal strategy is to alternate between the two experimentally available activities, allocating one's time to each goal in direct relation to the rate at which that goal delivers rewards. In other words, the animal ought to match its responding to the relative rate of reinforcement.

The process by which this matching behaviour occurs has long been believed to require a feed-back mechanism. Reward receipt, the consequence of instrumental responding in these procedures, would feed back onto its perceived cause, thereby strengthening the association between response and outcome. Thorndike (1898) was first to describe this Law of Effect, whereby instrumental conditioning occurred be-

cause of a gradual strengthening of the association between the experimenter-desired response and a subject-desired stimulus. This principle was further formalized by Sutton and Barto (1981), using the temporal-difference reward prediction error (RPE) as the critical signal driving instrumental learning. The RPE represents the discrepancy between the total reward at time $t + 1$ that is expected at time t and the actual reward delivered at time $t + 1$; a large (positive or negative) RPE results from a large discrepancy, which consequently adjusts the total reward expected at that time in proportion to the magnitude of the RPE. With sufficient training, the RPE disappears, as the expected total reward begins to approach the true total reward. This computational framework has been useful in designing intelligent machines and has been used to describe the activity of phasic dopamine activity in the ventral tegmental area (Montague et al., 1996).

Purely model-free descriptions of instrumental responding rely heavily on this temporal-difference reinforcement learning model. Performance is the result of implementing a behavioural policy that uses the total future expected reward at every time instant, whereby the value of taking a particular action in a particular trial state is the maximum temporally-discounted reward that can be expected by taking that action in that trial state and assuming one pursues an optimal strategy from there on. The total future expected reward of any given action-trial state pair is gradually updated as the animal engages with its environment; when the RPE associated with the total future expected reward of action-trial state pairs is nil, the animal is said to have "learned" the task. In this description of the task of a rat responding for electrical brain stimulation in the randomized-triads design, the rat begins a new trial expecting the value of lever-down states to be the same as those it was expecting at the end of the last trial. When those expectations are violated because the price-frequency pair in effect have changed, a purely model-free description predicts that the set of expectations—and the behavioural policy they collectively set—will

change gradually as the rat obtains more and more rewards.

In contrast, model-based descriptions imply a sort of feed-forward mechanism. Just as the rat has a world-model of how trials within a triad progress throughout the session (Chapter 3), the rat may have a world-model of how reward deliveries progress throughout the trial. On this view, an internal representation of the constancy of the price and pulse frequency from the time the house lights cease to flash to the time they begin flashing again would not require gradual changes to the expected total future rewards within a trial. Instead, the rat would know that conditions will be stable, and will be able to set these expectations in a single, step-wise change following a sufficient number of exemplars, as soon as it has information about the subjective opportunity cost and reward intensity of the electrical stimulation on offer. As a result, the rat's behavioural policy, both in terms of the duration of the pause it decides to make following each reward delivery and in terms of the overall proportion of time it decides to allocate to harvesting those rewards, will cease to change as soon as the payoff from self-stimulation is known.

The idea of a feed-forward model is not new. Mark and Gallistel (1994) found that rats adjusted their performance quickly (within one or two reward deliveries) to signalled changes in the rate of reinforcement on concurrent-interval schedules of reinforcement. Mazur (1995b) found that pigeons adjusted much more slowly, on the order of tens of minutes, to an unsignalled change, but the time scale over which the rates of reinforcement on each pecking key were stable was also on the order of many days. Gallistel et al. (2001) resolved this apparent discrepancy by showing that rats adjusted to unsignalled changes as quickly as would be expected by an ideal detector. Rats took a long time to adjust to an unsignalled change in average rates on a concurrent variable-interval schedule delivering moderately high pulse frequencies when the average rate of reinforcement on each of the two levers was stable for long periods of time. When the average rates were only stable over short periods of time,

rats adjusted to unsignalled changes within one or two reward deliveries. This implies that rats adjust the duration of time spent working for an alternative as soon as a change in the rate at which the alternative delivers rewards can be detected.

The following chapter provides evidence that in the case of rats working for electrical brain stimulation, for which the reward is delivered after the lever has been held for a fixed cumulative period of time and whose subjective opportunity cost and reward intensity is constant throughout a signalled trial, rats behave as though there has been a single, step-like change to their behavioural policy. Furthermore, we provide evidence that, consistent with the ideal-observer description, this step-like change predominantly occurs as soon as theoretically possible: following delivery of the first reward.

4.1.1 A world model of the reward encounter

The defining feature of a trial in the curve-shift (Miliaressis et al., 1986) method is that the response requirement and reward magnitude are fixed. The trials are usually cued in some way, and the animal may harvest as many rewards as trial and schedule constraints allow in exchange for fulfilling the response requirement. Although we have previously shown in Chapter 3 that the rat readily forms a world model of how the subjective opportunity cost and reward intensity change from trial to trial, the rat may have an additional model of how the subjective opportunity cost and reward intensity remain constant throughout a single trial. We define the reward encounter as the period within a trial, from the time the lever extends into the operant chamber to the time the lever retracts from successful completion of the response requirement or the end of the trial. On trials when the payoff—a scalar combination of the key decision variables—is high, there may be many such reward encounters, as the rat earns many rewards. On trials when the payoff is low, there may be very few reward encounters, as the rat earns very few rewards. Just as the rat

has formed a representation of how trials lead to each other in a predictable way, as demonstrated in Chapter 3, the rat may also form a representation of how successive reward encounters are predictably identical with respect to the subjective opportunity cost, reward intensity, and probability of reinforcement.

We describe this world model—one regarding the stability of key determinants of decision within the trial—as a world model of the reward encounter. Without such a world model, the rat would need to slowly extract and update an estimate of the payoff. In the standard model-free reinforcement learning approach, the rat updates its estimate of the total net reward from lever-pressing according to a delta rule (Dayan and Abbott, 2001) that specifies the degree to which the current total net reward differs from the current estimate.

In the randomized-triads design, the cumulative amount of time the lever must be held in order to harvest a reward (the price) and the pulse frequency of the stimulation in effect during a test trial are drawn pseudo-randomly from a finite set of price-frequency pairs. It has already been established, in Chapter 3, that the rat behaves as if it had an expectation regarding the subjective opportunity cost and subjective reward intensity for the trial to come. This expectation, on test trials, is usually different from the subjective opportunity cost and reward intensity of the BSR to come because of the random sampling method in force. A rat using purely model-free reinforcement learning mechanisms would have to gradually update its estimate of the subjective opportunity cost and reward intensity associated with self-stimulation over multiple reward encounters before it obtained an accurate estimate of the payoff it can expect from self-stimulation.

In contrast, a rat using a model that states that the subjective opportunity cost and intensity of the reward is fixed for the duration of a trial would only need a small number of exemplars—possibly a single reward delivery—before it updated its estimate of the payoff from self-stimulation in a single step. Since there is no

variability in subjective opportunity cost or reward intensity throughout the trial, the rat can simply "fill-in" the appropriate value for the payoff and immediately implement a single policy.

It is clear that at the onset of any test trial in the randomized-triads design, the rat cannot know how long the lever will need to be depressed in order for the reward to be delivered or how strong the stimulation will be when the reward is delivered. As established in Chapter 3, the duration of the pause that begins the trial (the post-priming pause, PPP) is, in part, a function of the payoff from self-stimulation that can be expected for that trial: shortest on leading bracket trials, longer on test trials, and longest (usually censored) on trailing bracket trials. Given that the payoff from self-stimulation on test trials is variable, we focus here on the duration of pauses within this trial type. On leading and trailing bracket trials, the rat need not update its estimate of the payoff. The rat begins the test trial with an estimate of the payoff that will often be inaccurate, and produces a PPP that partly reflects the rat's estimate of the average payoff to come. Following delivery of the first reward of the test trial, the payoff from self-stimulation will usually require revision. Once the blackout delay, a period of time when the lever is retracted from the chamber and the reward is delivered, elapses, the rat takes a (possibly zero-duration) pause. As this pause follows reinforcement, we call it the post-reinforcement pause (PRP).

Any gradient descent scheme, such as model-free reinforcement learning, will require a gradual change in the PRP. In its classical formulation (Montague et al., 1996), model-free operant conditioning is similar to classical conditioning: the manipulandum represents a stimulus for which an action (like lever-pressing) will lead to a total discounted net reward. At time step t, the value of the lever-down state ($\hat{V}_t$) is the reward delivered at time step t (R_t) and the value of the next time step ($\hat{V}_{t+1}$) discounted by a factor γ: $\hat{V}_t = R_t + \gamma\hat{V}_{t+1}$. When a reward is delivered that violates expectation, there is a reward prediction error (δ_t), formalized as the dif-

ference between the current value and the rewards predicted to come. The current estimate of the value of a state at a point in time is updated in proportion to the magnitude of this reward prediction error: when δ_t is a large positive number, the animal has obtained a much larger reward at time t than expected, and when δ_t is a large negative number, the animal has obtained a much lower reward at time t than expected.

The crux of model-free reinforcement learning schemes is in this reward prediction error signal. Over multiple reward deliveries, the value of a lever-down state is modified according to an update rule whereby the old value is increased (when δ_t is positive) or decreased (when δ_t is negative) by a factor α:

$$\hat{V}_{t_{new}} \leftarrow \hat{V}_{t_{old}} + \alpha\delta_t$$

where $\delta_t = R_t + \gamma\hat{V}_{t+1_{old}} - \hat{V}_{t_{old}}$. For example, suppose a reward R delivered at $t = 10$ time steps is 1 arbitrary unit, when no reward was expected ($\hat{V}_{t=20} = 0$). The first time such a reward is delivered, it induces a positive reward prediction error, since

$$\delta_{t=10} = R_{t=10} + \gamma\hat{V}_{t=11} - \hat{V}_{t=10} = 20 + 0 - 0 = 1,$$

so the value of the lever-down state becomes

$$\hat{V}_{t=10} \leftarrow \hat{V}_{t=10} + \alpha\delta_{t=10} = 0 + \alpha \times 1 = \alpha.$$

The next time the manipulandum extends into the chamber, all values up to 9 steps have not changed, but the value of a lever press at 9 time steps violates expectation, since in this second time around,

$$\delta_{t=9} = R_{t=9} + \gamma\hat{V}_{t=10} - \hat{V}_{t=9} = 0 + \gamma\alpha - 0 = \gamma\alpha.$$

The third time the manipulandum extends into the chamber, the value of a lever press at 8 time steps violates expectation, and is similarly updated. Every time the manipulandum extends into the chamber, the value of lever-pressing is updated for one time step earlier, until the first time point at which the reward can be predicted. In the classical conditioning case, the conditional stimulus that predicts reward is presented at random intervals, and thus the first time a reward can be predicted is the onset of this conditional stimulus. Operant conditioning accounts of performance presume that the conditional stimulus is the manipulandum, and the value of action-state pairs is learned as in the classical conditioning case.

Even when the learning rate (α) is one—that is, the value of lever-pressing is updated immediately to its new value—gradient descent models like model-free reinforcement learning imply that performance will change gradually. In the above example, the first time a surprising reward of 1 arbitrary unit is delivered at $t = 10$ time steps,

$$\delta_{t=10} = 1 + 0 - 0 = 1, \text{ and}$$

$$\hat{V}_{t=10} \leftarrow 0 + 1 = 1.$$

The next time the manipulandum extends into the chamber, although $R_t + \gamma \hat{V}_{t+1} - \hat{V}_t$ does not change from time steps 1 through 8, at time step 9,

$$\delta_{t=9} = 0 + \gamma \times 1 = \gamma, \text{ and}$$

$$\hat{V}_{t=9} \leftarrow 0 + \gamma = \gamma.$$

Thus, even when the value of a state is updated to the last reward delivered, this value must back-propagate, one time step at a time, to the earliest time a reward can be predicted: presentation of the conditional stimulus, in the case of classical conditioning, or extension of the manipulandum into the operant chamber, in the

case of operant conditioning.

From this description, it is clear that changes in performance must be gradual when feed-back mechanisms are at work. The back-propagation model works in the classical conditioning and simple, punctate lever-pressing situations because the stimulus onset is itself unpredictable. On repeated stimulus (or manipulandum) presentations, the discrepancy back-propagates to the earliest time the reward can be predicted. Indeed, for the animal to solve the assignment of credit problem—that is, identifying the state-action pair that led to reward—classical model-free formulations (Montague et al., 1996) require back-propagation, which gradually updates the value of all state-action pairs that led to a temporally distant reward. Until this back-propagation is complete, the rat's decision to press must be based on a mechanism that updates the value of a lever-press one time step at a time. Once the value of all state-action pairs no longer needs an update—that is, the back-propagation mechanism has collided to the first possible time step at which reward can be predicted over possibly many reward deliveries—the rat can pursue an action selection policy based on a stable estimate of the discounted total net reward to come.

However, if the rat has a world model of the reward encounter, the rat will not need to update its estimate of the payoff from self-stimulation gradually. After it has obtained a sufficiently large number of exemplars, the payoff can be updated as a step function, producing a step-like change in both the duration of the PRP and the proportion of time it allocates to self-stimulation activities. Whereas a rat using model-free reinforcement learning principles can only represent the value of lever-pressing, a rat using model-based reinforcement learning principles also represents how states are interrelated. In the case of a world model of trial structure, the rat maintains a representation of the stability of the price and reward intensity in effect on a trial: every time the lever extends into the operant chamber, from the time the house lights flash until they flash again, the price and intensity will be constant.

As the function that maps the transition of one trial state (from lever-extension to lever-retraction) to the next is an identity function, the rat can effectively update the payoff from lever-pressing in a single step. Without such a state-transition function, the rat must gradually update the value of lever-pressing on every trial.

To test which of these two accounts best described the behaviour of rats, we compared two models of the evolution of PPPs and PRPs throughout the test trial. One model, the gradual change model, implies that pauses (both PPP and PRP) are sampled from a distribution of first-pauses whose mean changes smoothly over multiple reward encounters and whose mean remains stable following this suitably long transition. The second, the step model, implies that pauses are sampled from one distribution of first-pauses for the first n reward encounters, and sampled from a single distribution of pauses for subsequent reward encounters.

We further tested the hypothesis that the rat's estimate of payoff from self-stimulation is updated incrementally by considering the total proportion of time allocated to self-stimulation within each reward encounter. If the estimate changed gradually, as the result of a hill-climbing mechanism, the proportion of time allocated to self-stimulation would gradually reach a steady point. If the estimate had changed abruptly, the proportion of time allocated to self-stimulation would also change abruptly. By looking at the first derivative of time allocation during reward encounters with respect to the number of rewards earned, it is possible to determine whether well-trained rats behave according to a purely model-free mechanism (with consistent, gradual changes that slowly reach a derivative of 0) or a partly model-based mechanism (with an abrupt change followed by a derivative of 0). If the rat behaves as though the payoff is updated by small iterative changes, this analysis will also show the time scale over which learning the payoff from self-stimulation for a given trial occurs. If the rat behaves as though the payoff is updated step-wise, this analysis will also show how many exemplars are necessary before the payoff estimate

is updated.

4.1.2 Work bouts

Prior to the delivery of the first reward, there is no reason to believe performance is completely dictated by the payoff from self-stimulation. Indeed, post-priming pauses on test trials (Chapter 3) are often closer to those on leading bracket trials than halfway between leading and trailing bracket trials. It is possible, then, that something in addition to the expected payoff drives performance prior to the trial's stable period. Our hypothesis for the period of time prior to the delivery of the first reward is that both expected payoff and the desire to uncover the payoff drive time allocation in this period of time to a high value. In other words, when the rat is still unsure of the payoff it can expect from self-stimulation, the goal of exploitation (taking advantage of the source of reward with the greatest payoff) is aligned with the goal of exploration (identifying the payoffs from all possible sources).

To verify this, we determined the number of corrected work bouts, which include holds and all releases lasting less than 1s that return to a hold, on all trials of all payoffs in the period of time before the payoff is known. We hypothesize that in this period of time, the rat will tend to earn its first reward in a single, continuous work bout. A similar analysis was conducted on the number of corrected work bouts emitted from the time the first reward is delivered onward. When the price is low, the animal will necessarily earn its rewards by engaging in a single, continuous work bout, regardless of whether a reward has been delivered during the trial. At higher prices, the rat may partition its time among work and true leisure bouts. If the period of time prior to the first reward delivery is in some sense "special," we hypothesize that the price at which the rat begins to engage in multiple work bouts will be higher for these reward encounters than on subsequent reward encounters, after the rat knows (in principle) the subjective opportunity cost and reward intensity of the stimulation

it is to receive.

4.2 Methods

4.2.1 Behavioural protocol

The rats were the same as in Chapter 3. The data reported here were collected during the test trial type in the randomized-triads design already described. Test trials were preceded by leading bracket trials, during which the pulse frequency was as high as the animal would tolerate without interfering effects, such as forced movements and vocalizations, and the price was 1 second. Test trials were followed by trailing bracket trials, during which the pulse frequency was too low to support responding (10Hz) and the price was 1 second. Trials were cued by a 10 second inter-trial interval; 2 seconds before the end of the interval, a train of priming stimulation of pulse frequency equal to that delivered on leading bracket trials was delivered. The price, pulse frequency, and probability of reinforcement in effect for the test trial was drawn pseudo-randomly from a list, as has been described before.

Following successful completion of the work requirement—the lever was held for a cumulative amount of time defined as the price—the lever retracted, the trial clock stopped, and a BSR was delivered. Two seconds after lever retraction, it was extended again into the cage and the trial clock re-started. This period is defined as the blackout delay. The duration of the trial, without blackout delays, was set to the greater of 25 seconds or 25 times the price, allowing the rat to harvest a maximum of 25 rewards if it had been holding the lever continuously for the entire trial. Often, the rat obtained many fewer than 25 rewards, because it did not continuously hold the lever for the entire trial.

In the case of rats MA5, DE1, DE3, DE7, PD8, DE14, DE15, DE19 and DE20, probabilistic test trials were drawn pseudo-randomly along with test trials for which

the probability of obtaining a reward following successful completion of the work requirement was 1. Test trials for which the probability of reinforcement was less than 1 have been excluded from the present analysis for simplicity. For all rats but MA5, the active operant on probabilistic trials was mapped to a different lever, providing additional prior information about the payoff on the trial. Furthermore, reward encounters in the probabilistic case do not cleanly map onto identical exemplars of the payoff to be expected from lever-pressing. To simplify performance comparisons across rats, we focus here on the case that is universal to them all: trials on which the probability of reinforcement is 1, but whose payoff is still uncertain because it may take on any of a list of values.

4.2.2 Statistical analysis

When using a model-free reinforcement learning scheme, a reward prediction error will appear at earlier and earlier time steps following each reward encounter, until the prediction error reaches the first time step at which a reward can be predicted. In the period of time when this reward prediction error is back-propagating, an animal using model-free reinforcement learning mechanisms will base its action selection policy on state-action values that change from reward encounter to reward encounter. As soon as the reward prediction error reaches the first point at which a reward can be predicted, and the values of each state-action pair converge, the rat's action selection policy will no longer change from reward encounter to reward encounter.

When using a feed-forward scheme, the rat can simply "fill-in" a payoff, rapidly updating the value of state-action pairs as soon as the payoff can be known. In the period of time before the current trial's payoff can be filled in, an animal using model-based reinforcement learning mechanisms will base its action selection policy on state-action values that have not yet been updated; as soon as the payoff is known, there

is no need to slowly update the value of state-action pairs. The rat simply revises its action selection policy according to the new, rapidly-updated payoff.

To determine which of the two descriptions of the rat's performance throughout the trial best captured the behavioural data, we compared two models of the durations of post-priming and post-reinforcement pauses following each lever extension into the cage during test trials. In one of them, pauses are modelled as samples drawn from a distribution with an unstable mean, that is, with a mean that begins at some value for the first reward encounter, ends at some value for the last m reward encounters, and transitions smoothly between the two for at least 1 reward encounter. In the other, pauses are modelled as samples drawn from two distributions, that is, from one with a particular mean for the first n reward encounters and one with a different mean for the subsequent reward encounters $n + 1$ to the end.

In addition, we examined the proportion of time allocated to self-stimulation during each reward encounter, from the time the lever extended into the cage to the time it was retracted, either because the response requirement had been fulfilled or because the trial had come to an end. A purely model-free account of performance throughout the trial would predict a gradual change in time allocation across reward encounters until the rat's internal estimate matched the putative "true" value. A partly model-based account would predict no change until a sufficient number of reward exemplars had been delivered, followed by an abrupt change when the update was made, followed by no change.

4.2.2.1 Modelling gradual and step-like changes in PPP and PRP

To determine which of the two descriptions better accounted for performance, we fit two models of the post-priming and post-reinforcement pauses by maximum likelihood. Figure 4.1 provides an example of the two models, gradual-change (dotted red line) and step-change (solid black line) models. The figure shows the value of the

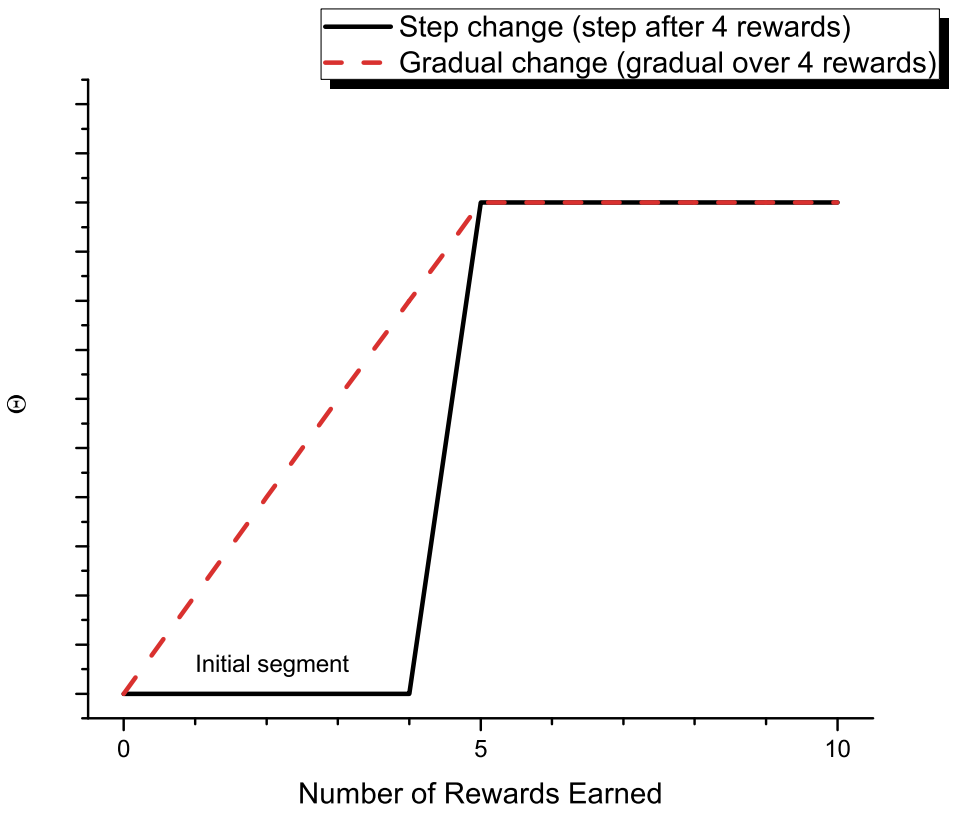

Figure 4.1. Comparison of gradual and step changes in distributional parameters across successive reward encounters. Parameters of the distributions generating post-priming (following 0 rewards) and post-reinforcement pauses (following rewards 1 through n) can either change gradually (red dashed line) or abruptly (solid black line). Here, Θ subsumes all parameters of the stochastic process that generates the pause, which in the case of a gamma distribution, would be the mean and standard deviation. Assuming the post-priming pause is drawn from a distribution parametrized with $\Theta_1 = \{\mu_1, \sigma_1\}$, and the last post-reinforcement pause is drawn from a different distribution parametrized by $\Theta_2 = \{\mu 2, \sigma_2\}$, we have modelled the non-stationarity in two different ways: either the parameters change gradually over m rewards, or there is a step change following m rewards. The initial segment refers to the period of time before the parameters of the process that produces post-reinforcement pauses reach their final Θ_2 values.

parameters of the distribution of post-priming and post-reinforcement pauses as a function of the number of rewards that have been earned in the trial. The ordinate is a place-holder for the parameters of the distribution of pauses, which in the case we have modelled below, subsumes the mean and standard deviation (in other words, $\Theta = \{\mu, \sigma\}$).

The dashed red line provides an example of the first model, a gradual-change model. According to the gradual-change model, the first pause (the post-priming pause) is drawn from a gamma distribution with mean μ_1 and standard deviation σ_1 (i.e., Θ begins at some initial value). Pauses m ($m = 6$ in the figure, after 5 rewards have been earned) to the end of the trial are drawn from a separate gamma distribution with mean μ_2 and standard deviation σ_2 (i.e., Θ reaches a final value). Pauses in between, from 2 to $m-1$ (in the figure, $m-1 = 5$, after 4 rewards have been earned), are drawn from a gamma distribution whose mean and standard deviation are straight-line functions of the reward encounter, starting at μ_1 for the pause on reward encounter 1 and ending at μ_2 for the pause on reward encounter m (i.e., Θ gradually changes from its initial to final value).

Pauses censored by the end of the trial were excluded from the analysis; if the rat had only ever collected one reward, we considered this to be infinite evidence in favour of a step-change model, since there was no way a gradual change model could account for this pattern of behaviour. The very infrequent case of a post-priming pause censored by the end of the trial, indicating that the rat simply never bothered to obtain a sample of the payoff, were excluded entirely from the analysis, as it would be impossible to arbitrate between the two models which was better at explaining the data. In total, only 459 trials (of 16007, or under 2.8%) across all animals, conditions, and price-frequency pairs were excluded.

The solid black line of figure 4.1 provides an example of the second model, the step-change model. According to this model, the first n pauses ($n = 5$ in the figure,

after 4 rewards have been earned) are drawn from a gamma distribution with mean μ_1 and standard deviation σ_1 (in other words, Θ begins at an initial value). Pauses $n+1$ (in the figure, $n+1=6$, starting on 5 rewards earned) to the end of the trial are drawn from a separate gamma distribution with mean μ_2 and standard deviation σ_2 (in other words, Θ reaches a final value). There are no transitional pauses.

In the case of a gradual change, we define an initial segment, extending from the PRP following reward 1 to the PRP following reward n. Throughout this initial segment, the parameters of the distribution from which PRPs are sampled varies smoothly as a straight-line function from the parameters that describe the distribution of PPPs to those that describe the stable segment that extends from the PRP following reward $n+1$ to that following the very last reward. For the step-change model, we can define an initial segment (from the post-priming pause, following reward 0, to the post-reinforcement pause following reward n), and a stable segment (from reward encounter $n+1$ to the last reward encounter).

In the case of the step-change model, for each initial segment of length S, from a length of one (that is, only the PPP is included in the initial segment) to two less than the maximum number of rewards delivered for trials on which each price-frequency pair was in effect, we identified the maximum-likelihood estimates of the parameters of a gamma distribution from which the pauses within the initial segment were drawn as well as the maximum-likelihood estimates of the parameters of a gamma distribution from which all pauses following the initial segment were drawn. These estimates provide the necessary information for calculating the probability of observing all pauses in the trial, assuming a step-like change, for an initial segment of length S.

In the case of the gradual-change model, for each initial segment of duration $S+1$, we identified the maximum-likelihood estimates of the parameters of a gamma distribution from which all pauses within the stable segment were drawn, as well as

the maximum-likelihood estimates of the parameters of a gamma distribution from which all post-priming pauses were drawn. These estimates provide the necessary information for calculating the probability of observing all pauses in the trial, assuming a gradual change, for a transition of duration $S + 1$. We then noted, for each initial step-change segment of duration S and initial gradual-change segment of duration $S+1$, the probability of the data assuming a step-like change (in the case of an initial segment of duration S), as well as the probability of the data assuming a gradual change (in the case of a transition segment of duration $S + 1$).

To compare the two models—gradual or step-like—we marginalized the likelihood of each model with respect to initial segment lengths and transition durations. We summed the probability of the data, assuming a step-like change, across initial segments of all lengths S, and the probability of the data, assuming a gradual change, across transition segments of all durations $S+1$. As a result, the first calculation provides the overall probability of the data, assuming a step-like change, when the initial segment contains 1, 2, or N reward encounters. The second calculation provides the overall probability of the data, assuming a gradual change, when the transition takes 1, 2, or $N - 1$ reward encounters to occur.

The ratio of the two probabilities provides the Bayes factor for one model compared to the other. The ratio of the probability of the data, given a step-change model, to the probability of the data, given a gradual-change model, for example, is the odds in favour of a step-change model. If this number is large (or, alternately, its logarithm is positive), then regardless of the duration of the initial segment of either model, a step-change model better accounts for the data than any gradual-change model. If this number is vanishingly small (or, alternately, its logarithm is negative), then regardless of the durations of the initial segment of either model, a gradual-change model better accounts for the data than any step-change model.

Following calculation of the Bayes Factor for each price-frequency pair in each

non-probabilistic condition for each rat, we extracted the value of S (if the common logarithm of the Bayes Factor was positive) or $S+1$ (if the common logarithm of the Bayes Factor was negative) for which the data were maximally likely. These values represent the maximum-likelihood estimate of the number of rewards the rat earned strictly before pauses could be said to be sampled from a single gamma distribution; this number therefore indicates the maximum-likelihood estimate of how many reward deliveries were necessary before behaviour could be said to have stabilized.

4.2.2.2 Time allocation difference

To describe the evolution of pauses in the trial, we computed the discrete equivalent of the derivative of the proportion of time allocated to self-stimulation in each reward encounter. For each reward encounter, defined as the period of time from lever extension (the end of the inter-trial interval or blackout delay) to lever retraction (the start of a new inter-trial interval or blackout delay), we summed the total time the lever was held with the total time the lever had been released for less than one second to obtain an estimate of the amount of time spent working for the BSR during the reward encounter. This work time was divided by the total duration of the reward encounter, including the total time the lever was depressed and the total time the lever was released, to obtain an estimate of time allocation for the reward encounter.

The time allocation difference was then calculated as the difference between time allocation on reward encounter n and time allocation on reward encounter $n-1$, the immediately preceding reward encounter, from the second reward encounter onwards. According to a model-free description of how the rats update the subjective opportunity cost and reward intensity in effect on a trial, this difference should gradually reach 0 over multiple reward encounters. If the rat has an estimate of the payoff based on the value of these key decision variables at the start of the trial, purely

model-free mechanisms would update this estimate incrementally as more rewards are earned. For example, time allocation might be relatively high on the first reward encounter from overestimating the payoff. When the first reward is earned, the payoff may be revised downward, producing a slightly lower time allocation. After n rewards, the rat's payoff estimate would reach the true payoff, and time allocation on all subsequent reward encounters would be the same. Taking the difference of time allocation with respect to reward encounter, a negative difference would be seen for the first n+1 reward encounters, reaching a value of 0 when the estimate of payoff became accurate.

In contrast, according to a model-based description, this difference will not be a smooth, continuous function. For the first n (where n could be as small as 1) reward encounters, time allocation will be constant, yielding a difference of 0. As soon as the payoff has been updated, time allocation will change, producing a large positive or negative number. Finally, for all reward encounters after the change, time allocation will be constant again, reflecting the new (accurate) payoff estimate, yielding a difference of 0 again.

4.2.2.3 Work bouts before and after the first reward is delivered

The above analyses imply two distinct periods of the trial: an adjustment period (when pauses can be understood as realizations of a stochastic process with distributional characteristics Θ_1 that are either stationary or non-stationary), followed by a stable period (when pauses can be understood as realizations of a stochastic process with stationary distributional characteristics Θ_2). At the boundary between the two, some process has occurred—either as a gradual or a step function—that has altered behaviour. It is natural, then, to ask whether the patterning of work bouts also differs between two periods of trial time we will demonstrate are different: prior to the delivery of the first reward, when the key determinants of decision-making

have not yet been revealed, and following the delivery of the first reward, when those key determinants have, in principle, been uncovered. Indeed, when trial parameters are not yet known, the usually competing goals of exploration and exploitation are aligned: exploration of the mapping between actions and rewards requires the rat to lever-press, thereby appearing no different than exploitation of the rewards derived from lever-pressing. Following the first reward delivery, the payoff from self-stimulation can (in principle) be known completely, so exploitation and exploration are once again antagonistic: he may greedily exploit the option (self-stimulation or extraneous activities) with the better reward, or sub-optimally explore whether the foregone option has increased in value.

To test whether the rat's willingness to work prior to the delivery of the first reward of a trial is different from that in subsequent periods of the trial, we extracted the number of corrected work bouts (holds and releases lasting less than 1s together) emitted by the rat resulting in a reward delivery, for either the time prior to or following the delivery of the first reward. This allows us to examine whether there are differences in how vigorously an animal will work for a reward it has not yet received, compared to when it knows what the payoff from self-stimulation will be. On each reward encounter that resulted in a reward delivery, both the objective price and the number of corrected work bouts (holds and taps, together, uninterrupted by a TLB) was recorded. For each objective price tested and in each of the two periods of trial time considered, the maximum-likelihood estimate of the mean and confidence interval of a Poisson process that generated the number of work bouts was determined. The confidence level for the intervals was adjusted to maintain a 5% family-wise error rate for the number of comparisons that were made: one for each unique objective price. As a result, the width of the confidence interval around each estimated number of work bouts per delivered reward was $(0.95)^{(1/c)}$, where c is the number of unique objective prices tested. We then determined the highest

price for which the maximum-likelihood estimate was significantly less than (that is, the upper bound on its confidence interval was strictly less than) 2 corrected work bouts per reward delivery. In other words, a 95% confidence interval about the mean number of bouts extending from 1.2 to 2 would indicate that on 97.5% of reward encounters resulting in a reward, that reward was obtained in fewer than 2 corrected work bouts. The highest price at which the 95% confidence interval (corrected for multiple comparisons) is strictly less than 2 provides the highest price at which the rat earns rewards in a single, continuous press. Any higher, and the rat unequivocally allocates at least two presses to obtaining the reward; any lower, and the rat cannot be unequivocally said to require multiple corrected work bouts to obtain a single reward.

4.3 Results

4.3.1 Two models of post-priming/reinforcement pauses

Figure 4.2 is a histogram of the common logarithm of the Bayes Factors, including those for which the Bayes Factor was infinite. In these latter cases, the animal only ever collected one reward before ceasing all responding, producing a single uncensored post-priming pause. In the upper inset, the region from -3 to 3 is highlighted. Commonly, Bayes Factors from 1 to 3 (common logarithms from 0 to 0.47) are considered trivial evidence, from 3 to 10 (common logarithms from 0.47 to 1) are considered substantial evidence, from 10 to 30 (common logarithms from 1 to 1.47) are considered strong evidence, from 30 to 100 (common logarithms from 1.47 to 2) are considered very strong evidence, and over 100 (common logarithm greater than 2) are considered decisive evidence (Jeffreys, 1998).

The median Bayes Factor was found to be 2.881 (common logarithm of 0.46), which indicates that regardless of how many transitional or initial post-reinforcement

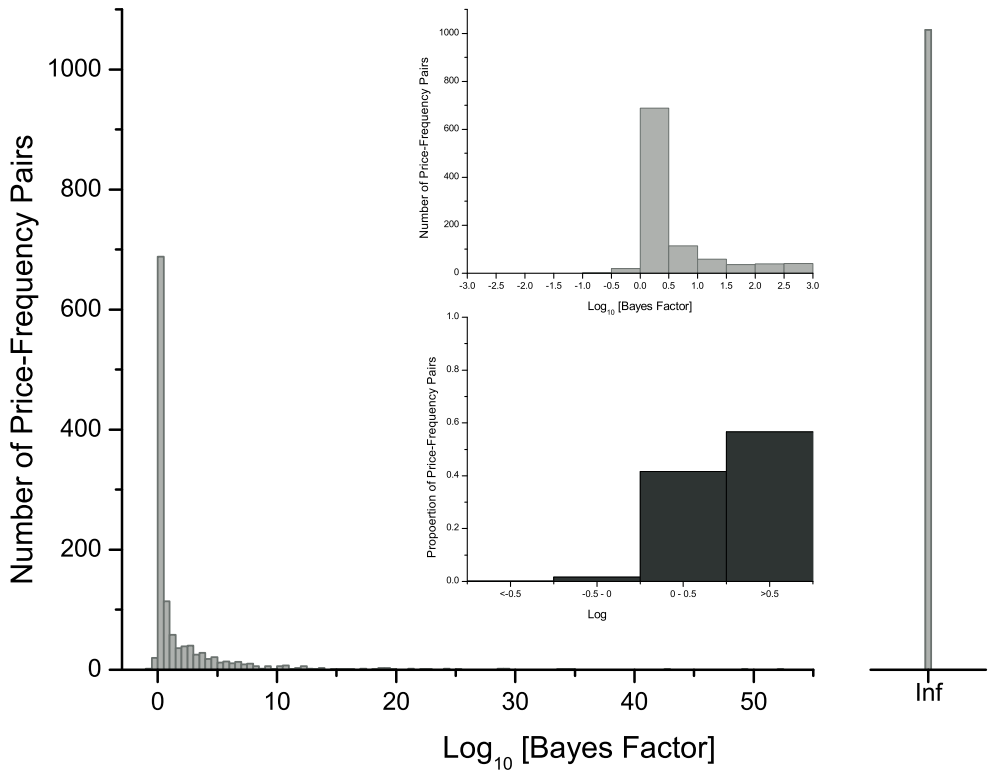

Figure 4.2. Bayes factors comparing step- to gradual-change models of pauses. Histogram of Log_{10}[Bayes Factors] observed in all price-frequency pairs of test trials of all animals in all conditions. Infinite Bayes Factors indicate that the rat collected a single reward and ceased responding, as the post-reinforcement pause following the first reward is censored by the end of the trial. The insets show the histogram in the region around a Log_{10} of 0 (top) in addition to a coarse grouping showing the number of Bayes Factors providing at least substantial evidence in favour of a gradual-change model (< -0.5), trivial evidence (-0.5 to 0.5), or at least substantial evidence in favour of a step-change model (> 0.5).

pauses one considers, a model according to which the distribution of pauses changes in step-wise fashion is just under 2.9 times more likely than a model according to which the distribution of these pauses changes gradually. This value of the Bayes Factor is just under what would be considered substantial evidence. Although many (533, or 43.2%) Bayes Factors are in the trivial range (with common logarithms ranging from -0.5 to 0.5), many (482, or 39.0%) also decidedly favour the step-change model, and a vanishingly small number provide any evidence that favours the gradual-change model. This is more clearly shown in the lower inset of figure 4.2, which depicts the proportion of price-frequency pairs that fall within the trivial and substantial-or-better ranges in favour of either the gradual-change model (negative values) or the step-change model (positive values). A considerable proportion of Bayes Factors provide at least substantial evidence in favour of a step-change model, while virtually none provide any evidence (trivial or better) in favour of a gradual-change model.

Indeed, 698 (56.6%) price-frequency pairs represent at least substantial evidence in favour of a step-change model, and only 2 represent substantial evidence in favour of a gradual change model, while no price-frequency pairs revealed strong or greater evidence in favour of gradual change. Figure 4.3 is a bar graph that depicts the number of price-frequency pairs for which the Bayes Factor falls into each qualitative category, for either the gradual-change (left) or step-change (right) models.

Finally, figure 4.4 is a histogram of the maximum-likelihood estimates of the number of reward deliveries required before pauses can be said to have been sampled from the same, underlying gamma distribution. Overwhelmingly, that estimate is 1: the median of the maximally likely number of reward deliveries required before the pauses the rat makes all come from the same (possibly payoff-dependent) gamma distribution is just one. Indeed, in only 32.3% of cases is the maximum-likelihood estimate greater than one reward delivery. The data provide a preponderance of evidence in favour of a step-change in post-priming and post-reinforcement pauses,

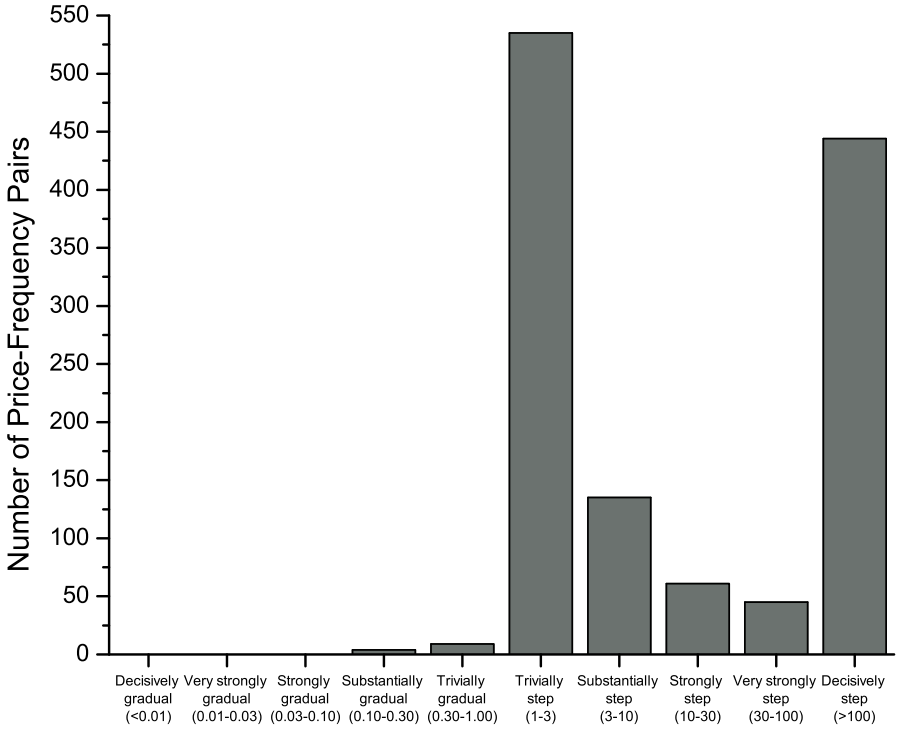

Figure 4.3. Qualitative categories of Bayes factors comparing step- and gradual-change models of pauses. Using Jeffreys' (1998) qualitative descriptions of Bayes factors comparing step- and gradual-change models of first-pause durations, there is little evidence to support a gradual-change model, very often trivial evidence to support a step-change model, and a preponderance of substantial, strong, very strong, or decisive evidence to support a step-change model.

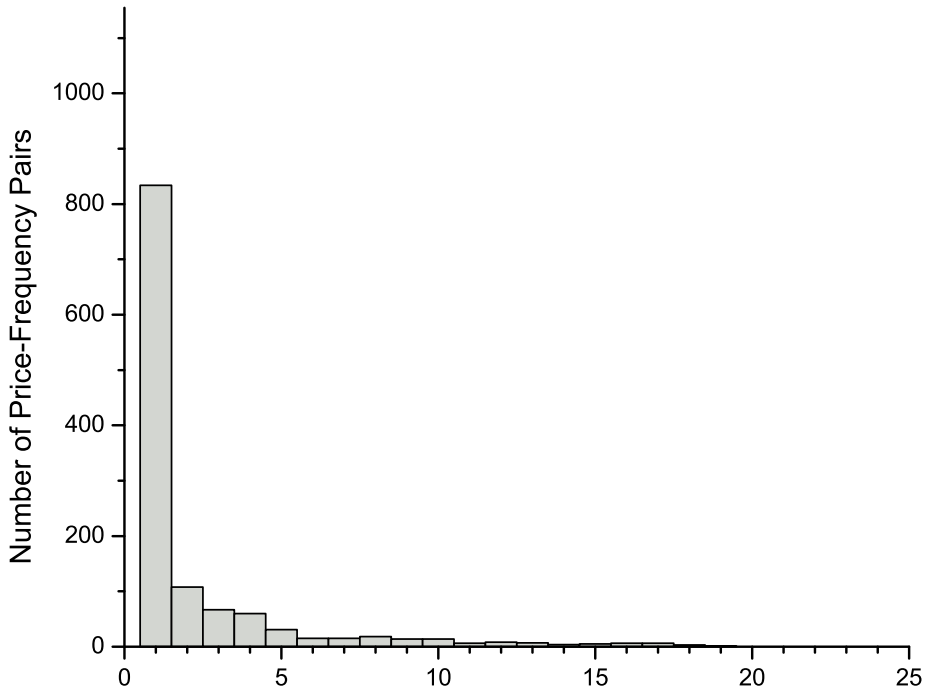

Figure 4.4. Maximum-likelihood estimate of the number of pauses in initial segments of either step- or gradual-change models. Histogram of the number of rewards required before a step- or gradual-change (the model with the best evidence) has been completed. The estimate is preponderantly one reward before performance changes, with comparatively very few instances of two or more rewards.

and the maximally likely time at which the step-change occurs is following the first reward delivery. It is still possible that although the post-reinforcement pauses are constant throughout the trial, the rat's behavioural policy may gradually change regardless. For example, the rat may interrupt its self-stimulation activities for longer and more frequent periods of grooming and exploring as its estimate of the payoff is gradually revised downwards, or shorter and fewer periods as the estimate is gradually revised upwards. It is to this question we shall now turn in identifying whether rats make a single, step-wise change to their behavioural allocation policy following a single exemplar of the payoff, or they make gradual changes throughout the trial as they update their estimate of the reward they can expect to receive and the opportunities they will have to forgo as they pursue electrical rewards.

4.3.2 Change in time allocation from reward encounter to reward encounter

The model comparison outlined above provides overwhelming evidence in favour of a step-change mechanism rather than a gradual-change mechanism. The natural next question becomes: how few reward deliveries are necessary before the rat behaves as though the estimate of payoff had been abruptly updated? We conducted an ANOVA on the difference in the proportion of time allocated to self stimulation with respect to reward encounter for every animal, for each combination of pulse frequency and price that was encountered on test trials. The step-change model implies that there will be a single reward encounter at which the change in time allocation is statistically different from zero, and will be zero otherwise. The reward delivery corresponding to that single reward encounter is the number of exemplars necessary before payoff is updated. For example, if the change in time allocation is statistically different from 0 following two rewards, and nowhere else, then the rat behaves as though it requires two exemplars of pulse frequency and price before it

abruptly updates its estimate of the payoff for the remainder of the trial. If, as will be demonstrated below, the change in time allocation is statistically different from 0 only following one reward, and nowhere else, then the rat behaves as though it requires a single exemplar of pulse frequency and price before it abruptly updates its estimate of the payoff for the remainder of the trial.

Figure 4.5 shows the mean absolute value of the difference of time allocation with respect to reward encounter number, from the first reward delivery to the tenth, for all animals. The general pattern, viewed across all animals, is consistent with the maximum-likelihood estimates of the step-change model derived above: time allocation changes suddenly (either becoming greater or smaller) between the first and second reward encounters, but ceases to change systematically from the second reward encounter onward. A within-subjects ANOVA conducted on the unsigned difference in time allocation with respect to reward number confirms this visualization, revealing a significant effect of reward encounter on the change in time allocation ($F(9, 106313) = 1792.3, p << 0.05, \eta^2 = 0.128$).

We further tested at which point time allocation ceased to change by conducting a series of comparisons between the change in time allocation following reward encounter n and all subsequent reward encounters. In other words, we compared the change in time allocation following the first reward delivery to the mean change following the second through tenth reward deliveries, then the change in time allocation following the second to the mean change following the third through tenth, and so on. To maintain a 5% family-wise error rate, we calculated the exact per-comparison probability for 9 comparisons as $1 - (1 - 0.05)^{(1/9)}$, or 0.0057. Although the change in time allocation following the first reward delivery was significantly different from all subsequent reward deliveries ($F(1, 106313) = 13.63, p < 0.05$ family-wise), time allocation ceased to change from the second reward delivery onward (F ranged from 6.6×10^{-4} to 0.38). Figure 4.6 depicts the mean squared deviations associated with

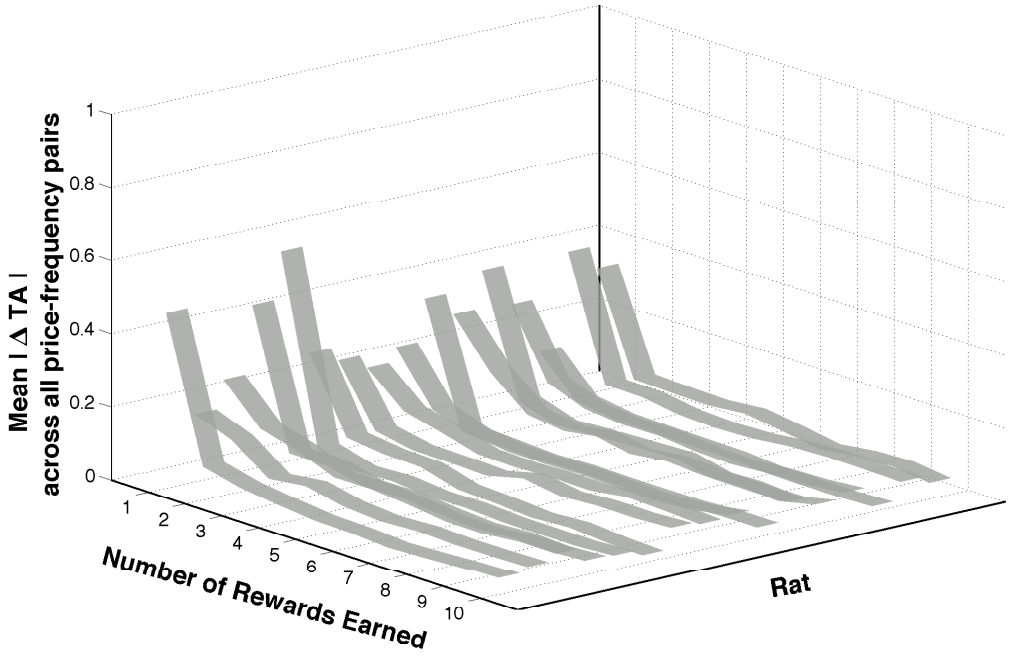

Figure 4.5. Absolute difference in time allocation for all animals. Mean unsigned change in time allocation from one reward encounter to the next as a function of the number of rewards earned, for the first 10 reward deliveries, for each rat. Time allocation changes drastically following the first reward delivery, but ceases to change thereafter.

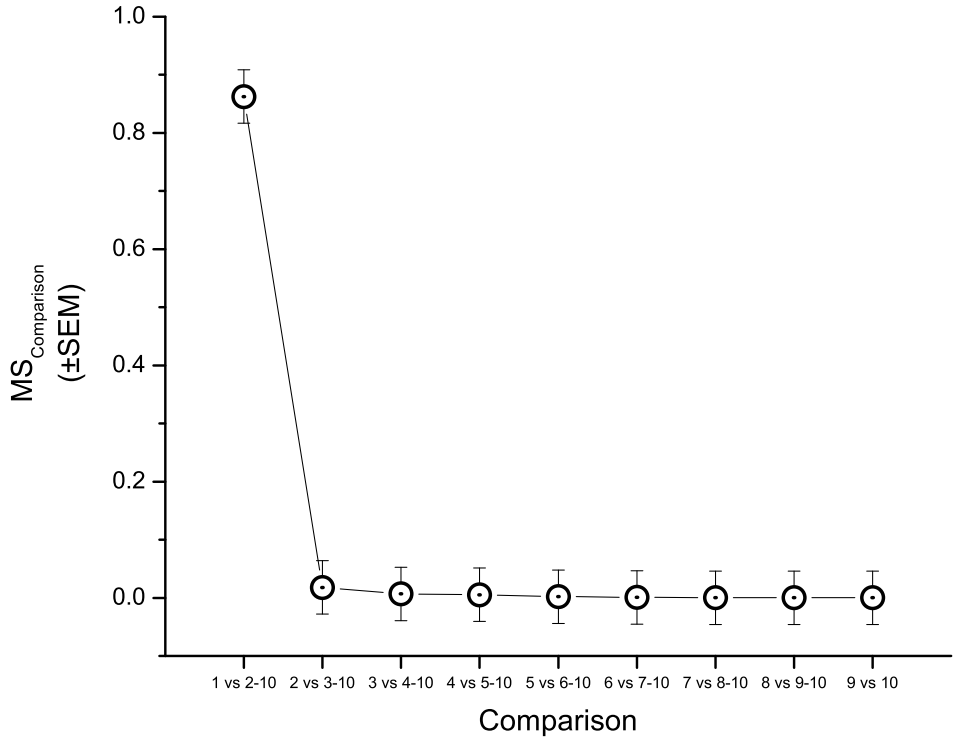

Figure 4.6. Post-hoc test of the within-subject change in time allocation across successive reward encounters. Mean squared deviation of each comparison (± SEM) as a function of the orthogonal comparison being made. While the change in time allocation following one reward delivery is significantly greater than the mean of all subsequent changes in time allocation, the change in time allocation following the second and subsequent reward deliveries are no different than each other and not reliably different from 0.

each single-*df* comparison (1 vs 2 through 10, 2 vs 3 through 10, etc.) along with their standard error. The mean squared deviation in the absolute value of the time allocation difference following one reward compared to the ten rewards that follow is significantly greater than 0, while the mean squared deviations of unsigned time allocation differences following two through nine rewards compared to those that follow are, for all intents and purposes, zero.

4.3.3 First reward encounter is different from subsequent reward encounters

Figure 4.7 provides a comparison of the crude estimates of the vigour with which a rat will work for electrical rewards before and following the first reward delivery. The upper left-hand panel is the mean number of corrected work bouts (per lever retraction) one representative rat engages in, for each price, before any rewards are delivered. Also indicated is the associated, 95% family-wise confidence interval. Asterisks indicate means that are significantly below 2 (maintaining a 0.05 family-wise error rate across all prices). The upper right-hand panel is the mean number of corrected work bouts (per lever retraction) that rat engages in from the first reward delivery onward. As with the left-hand panel, asterisks indicate means that are significantly below 2. Our crude estimate of the vigour is the highest price for which the estimated number of corrected work bouts is significantly less than 2: at lower objective prices, the rat usually (though not exclusively) engages in a single work bout to obtain a single reward, but at higher prices, the rat is always most likely to obtain the reward after two or more bouts. Dotted lines indicate, in each case, this crude estimate of vigour: the highest price at which the rat obtains rewards in a single bout of work.

The bottom panel of Figure 4.7 depicts the mean logarithm of the highest price at which the number of corrected work bouts per lever retraction is significantly

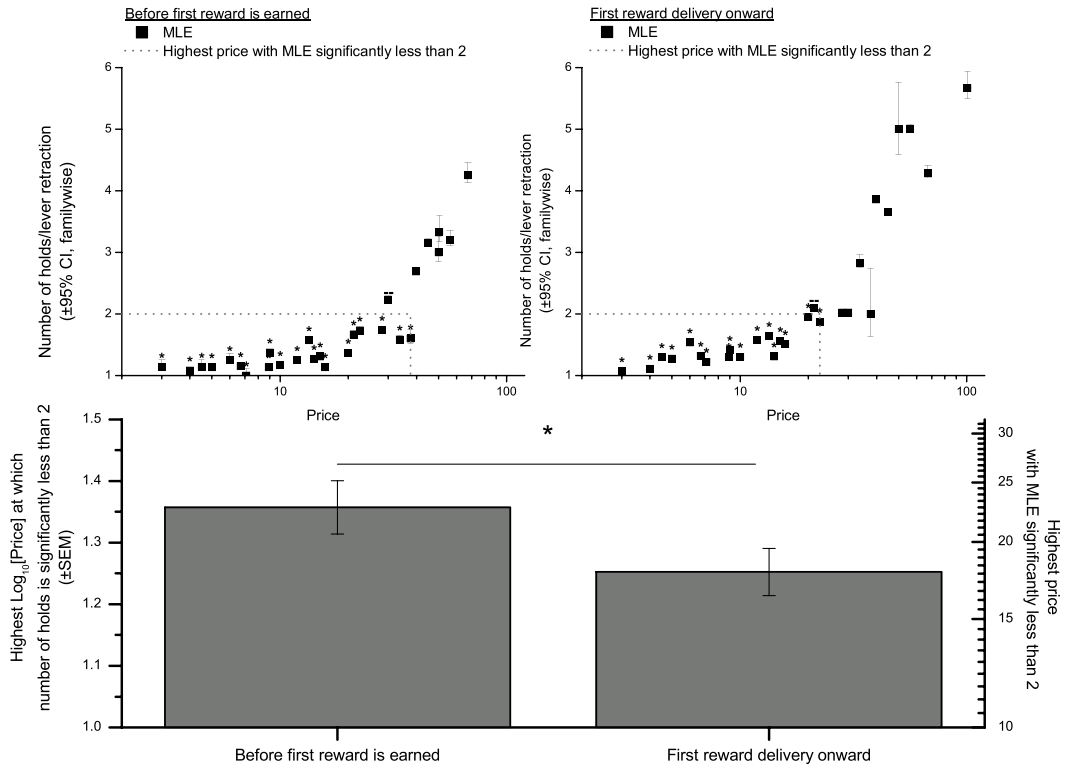

Figure 4.7. Number of work bouts required to obtain a reward. The top left panel shows the maximum-likelihood estimate (± 95 % confidence interval, family-wise) of the number of corrected work bouts (holds interrupted only by releases lasting less than 1s) required to obtain a reward as a function of the objective price, prior to the delivery of the first reward, for a single animal. The top right panel shows this estimate as a function of the objective price following the first reward delivery. Asterisks indicate estimates that are significantly less than 2 (two tailed, family-wise correction). Dotted lines indicate the highest price at which rewards are earned in significantly fewer than two bouts. The bottom panel is a bar graph of the highest price at which rewards are earned in fewer than two bouts for all animals. The asterisk here indicates a statistically significant within-subject difference in this price ($p < 0.05$).

less than 2 for all animals in all conditions, before and following the first reward delivery. A paired-samples t-test was performed on the difference in the logarithm of the highest price at which the number of corrected work bouts is significantly less than 2, to identify whether there was an effect of the first reward delivery. The t-test revealed a significant effect of the first reward delivery on this crude estimate ($t(26) = 2.77, p < 0.05$), indicating that rats were willing to work for higher prices before the payoff was completely known compared to after it was transparent in principle, across all conditions they encountered. This result, combined with the above two results concerning the model comparison and time allocation difference, implies that there are indeed two distinct periods of time, the boundary of which can be delineated by the delivery of the first reward of the trial.

4.4 Discussion

4.4.1 Stable trial responding

In our hands, performance for rewarding brain stimulation is remarkably stable following a step change between the time the payoff from self-stimulation cannot be known to the time it is (in principle) completely transparent. A contrast of two models—one in which performance in some initial period switches abruptly to a stable period to one in which performance changes gradually over a number of price-frequency exemplars until it is stable—provides evidence that, no matter how long or short the initial or transitional periods are, the model for which the data are more likely is generally the step-change model. The median Bayes Factor, across all price-frequency pairs encountered in all conditions for all rats, is just under what would be considered substantial evidence, with a considerable proportion of price-frequency pairs providing what we have termed "infinite evidence" in favour of a step change model. As soon as the rat "knows" what the subjective opportunity cost and reward

intensity of the electrical brain stimulation will be, the animal rapidly switches to a behavioural policy that reflects that payoff, rather than gradually adjusting it over a number of trials. The maximum likelihood estimate of the number of reward deliveries required for the animal to perform this switch is an equally overwhelming answer: following just one reward delivery, pauses can be said to come from a single, underlying distribution for the remainder of the trial, for the great majority of price-frequency pairs across all animals.

One could imagine that despite stable post-reinforcement pauses following the first reward encounter, overall performance reflecting an underlying behavioural policy could change gradually as the rat obtains more exemplars of the payoff from self-stimulation. For example, the rat may opt to maintain a constant post-reinforcement pause but interrupt its lever-pressing with more and longer bouts of leisure activities, such as grooming, resting and exploring as its estimate of the payoff from self-stimulation was revised to lower and lower values over a number of rewards. When the payoff is better than expected, the rat may opt to interrupt its lever-pressing with fewer and shorter bouts of non-lever pressing activities it can engage in throughout the trial. Such a behavioural policy would make the proportion of time allocated to lever-pressing change gradually as more rewards were delivered. Further confirming the model-comparison results and maximum-likelihood estimates extracted from them, the change in time allocation from reward encounter to reward encounter was significantly greater following the first reward delivery compared to all subsequent reward deliveries, and was no different for all subsequent reward deliveries.

Moreover, an animal's willingness to work is different from the time the reward is unknown to the time it is delivered. Whereas an animal will engage in a single, continuous work bout (releasing the lever only for a very short period of time to tap) up to high prices when the subjective opportunity cost and intensity have not yet been revealed, the same animal will engage in a continuous work bout up to a

significantly lower price when the subjective opportunity cost and intensity have been revealed compared to when they have no yet been revealed.

Taken together, these results imply that the rat can maintain a world model of the constancy of the payoff if that constancy is cued. As a consequence of this world model, if the payoff is not known when the trial begins, the rat can update its estimate of the payoff in a single step-like change following the delivery of a single reward. This would not be true if the rat had no world model of the constancy of the payoff throughout the trial: the rat would need to learn *de novo* what the subjective opportunity cost and reward intensity of the electrical reward are every time a new trial was presented.

The idea of world models in rodents is certainly not new. As a world model refers to a representation of how states transition to each other, this description is an extension to the operant conditioning domain of Tolman's original concept of a cognitive map. In Tolman's (1948) formulation of cognitive maps, the rat does not simply evaluate the net reward associated with an action. Instead, the rat forms a representation of the relationship between external environmental cues and the path that has been taken through them. For example, consider a hungry rat that is allowed to explore a Y-maze in which one arm was baited with water and the other with food. Assuming a purely model-free learning mechanism, the total net reward from visiting the water arm of the Y-maze is 0, and thus, no reward prediction error is made: the rat expects no reward, receives no reward, and does not update the value of heading to the water-baited arm. The total net reward from visiting the food arm comes as a surprise: the rat expects nothing at the end of the arm, receives food, and updates the value of heading to the food-baited arm. Such an animal cannot know what the state will be when it heads left or right, because that information is neither learned nor represented. Despite this, rats trained in this way were found to quickly head toward the water-baited arm of the Y-maze when subsequently water deprived,

indicating that they had a cognitive map of where to find food and water. The idea of a cognitive map is isomorphic to a model-based, feed-forward model of operant performance: the rat acts based on a mapping, not of food and water locations, but rather, of the varying demands that will be placed on the rat to obtain rewards of varying strength.

If rats indeed behave as though they have a model of the constancy of the subjective opportunity cost and reward intensity throughout the trial, then this model must be learned from the time the animal first encounters the flashing light of a trial. When the rat is first placed in an operant chamber for training, the flashing house lights are necessarily meaningless, as the rat has never seen the house light cue before. Over the course of multiple reward deliveries, the rat forms a representation of the amount of time the lever will have to be held and of the magnitude of the reward to be delivered, in addition to representations that may or may not change like the probability of reinforcement. As a new trial begins, the house light cue is presented again, and the rat must maintain a new representation of the new subjective opportunity cost and reward intensity in effect for the new trial. Given that the subjective opportunity cost, reward intensity, and (in the case of rats undergoing the probability experiment) probability of reinforcement vary considerably from trial to trial, the only reliable signal provided by the flashing house light is that a change in any of the key determinants of decision may have occurred, and the variance in those key determinants will be zero from the offset of the flashing house light to its onset at trial's end. It is possible that the rat maintains a representation of this change (as implied in Chapter 3) and a representation of the non-variance in the determinants of the payoff from self-stimulation.

Testing where this representation is maintained is a straightforward empirical question in principle, provided one were recording in the correct location. Rats could be presented with trials differing with respect to the variability of an easily controlled

determinant of the payoff: the subjective opportunity cost. Rats would be presented with the randomized-triads design, though in addition to the pseudo-random selection of the test trial from a list of prices and pulse frequencies, any given trial in the sequence could require a fixed or variable work requirement before a reward was delivered. In other words, an appropriately-cued lead, test, or trailing trial would deliver stimulation when the lever had been held for a fixed, cumulative amount of time (as is the case for the present experiment), or for a variable, cumulative amount of time. Assuming a sufficiently well-trained rat, putative neurons responsible for signalling that the payoff may have changed would predictably fire at the start of every trial type regardless of the variability in opportunity cost. Putative neurons responsible for signalling the variance (or lack thereof) in opportunity cost would show differential activity following each cue type. Actually conducting such an experiment, however, would be a herculean task, as the number of price exemplars the rat would need to sample before it had a reasonably accurate estimate of the average subjective opportunity cost on any given trial is very large indeed. Nonetheless, the question is empirical and would provide a mechanism for the neural basis of model-based decision-making.

4.4.2 Fast learning or filling in?

Given that rats require only one exemplar of the payoff on a particular trial in order to set a behavioural policy that will guide performance for the remainder of the trial, it is unlikely any feed-back mechanism can explain these data. Either each time step is so large that it is meaningless (there is only one time step for which to update the value of state-action pairs), or there is truly a simple representation of within-trial stationarity somewhere in the brain: if the house light begins to flash, then the key determinants of the decision will stay constant until it flashes again. Much like the previously-described world model of the triad structure of the session,

we have presented considerable evidence here that the rat behaves as though it has a world model of the constancy of the key determinants of its decision to press within a trial.

Our results are closely related to and extend findings regarding performance for brain stimulation rewards in the rat when those rewards were delivered at rates that changed in either a cued or un-cued fashion (Gallistel et al., 2001; Mark and Gallistel, 1994; Mazur, 1995a). Gallistel et al. (2001) found that when the average rate of reward changed in un-cued fashion following either long (weeks) or short (hours) periods of time, rats' stay durations at each of the levers of a concurrent variable interval schedule tracked fairly closely the performance of an ideal, Bayesian observer. When changes in the average rates occurred on the order of hours, rats adjusted phenomenally quickly to the new contingency, while when the average rates of each lever were stable on the order of days and weeks, rats adjusted considerably more slowly. Their analyses also demonstrate that performance does not track the immediately local rate as suggested by Mark and Gallistel (1994), *per se*, but rather, reflected the average rate.

One way that model-free reinforcement learning could arguably account for this step-like change is by assuming that the payoff from self-stimulation acts as a discriminative stimulus. In this case, the rat has learned over the course of training every combination of subjective opportunity cost and reward intensity it is likely to encounter, extracted the appropriate total future reward from taking every action in every trial state, and implements that pre-learned behavioural strategy as soon as it encounters the combination again. Given the combinatorial explosion that would be involved, and previous results implying that rat performance approximates an ideal detector of change in concurrent variable-interval schedules (Gallistel et al., 2001), we find this proposition unlikely. As we argue below, a more satisfying account of the process underlying the observed step-change from post-priming to post-reinforcement

pause, as well as from overall performance before the first reward delivery to following first reward delivery, can be found in a feed-forward, "filling in" model.

4.4.3 An extension of ideal-detector theory

When conditions are stable, it would behove an animal working under a variable-interval schedule of reinforcement to require a large number of exemplars before inferring that the average rate has changed, since the experienced rates of reinforcement are the result of an exponential process. When conditions are unstable, a small number of exemplars ought to be necessary. The duration for which conditions have been stable places a high prior probability for the average rate of reinforcement. In order to infer that these rates have changed, an animal would require substantial evidence before the mean was updated. In contrast, variable conditions place a flatter prior probability for the average rate of reinforcement, which require less evidence to infer that there has been a change in rate. When there is a great deal of evidence that a patch should deliver berries at a rate of 3 per minute, a few anomalous observations should not lead a foraging animal to infer that a change in rate has occurred. Similarly, when conditions change rapidly, and there is no evidence of a particular rate of reinforcement, less evidence should be necessary for an animal to infer that a change in rate has occurred.

In the present experiment, an ideal detector of change would necessarily require a single exemplar of the payoff to infer that a change has occurred. As there is no variability in the work requirement, and therefore in the rate during work, any change, even if it were unsignalled, would require only one reward to be detected. This is a proposition that could be, in principle, tested empirically: if the price of the electrical reward, which sets the reciprocal of the rate of reinforcement during work, were to abruptly change midway through the test trial, but remain constant for the remaining duration of the trial, the rat would re-adjust its performance in step-wise

fashion. Similarly, if the subjective intensity of the electrical reward were to change step-wise midway through the test trial, and then remain constant, the rat would re-adjust its performance in an equally step-wise fashion. In this sense, our proposal that rats in the randomized-triads design develop a world model of the constancy of the trial from reward encounter to reward encounter is simply a special case of the proposal that rats behave as though they are ideal detectors of change.

Since both initial pauses and overall performance change in step-like fashion following the first reward, the question becomes: what do these behavioural measures reflect about the rat's policy? Model-free learning mechanisms require a back-propagating mechanism to the earliest point a reward can be predicted, which would result in gradual changes in the animal's action selection policy. It is therefore unlikely that our results can be explained by a simple hill-climbing process that updates the estimate of reward over multiple reward deliveries, revising the magnitude of the reward to the same degree on every iteration and the subjective opportunity cost as the lever is held for the required cumulative amount of time. Instead, we propose that if such an update occurs, the rat takes into account statistical properties of the environment to tune the rate at which estimates of key determinants of decisions are revised. In the case of brain stimulation rewards delivered in cued trials for which these determinants are constant, the statistical properties of this environment (formalized here as a world model of the reward encounter) allow the rat to update its estimates of the determinants that matter in a single step.

If we assume that rats use a strategy that is at least partly model-based, then there is necessarily a "filling in" mechanism driving the rat's behavioural policy: as soon as the payoff is known completely and with certainty, as it is following the first reward delivery, the rat can simply "fill in" this payoff into what it ought to do, from the moment the lever extends back into the cage at the end of the blackout delay. Prior to this point in time, before the payoff is completely known, the rat may have updated

its estimate of the subjective opportunity cost as it held the lever and adjusted how it partitions its time on the basis of this on-going estimate. However, from the time the first reward is delivered until the trial ends, performance is completely stable. Post-reinforcement pauses can be said to have come from the same underlying gamma distribution, and the proportion of time allocated to lever-pressing (which is the ratio of corrected hold time to the sum of post-reinforcement pause, corrected hold and release times) for each reward delivered after the first cease to change. Since the amount of time the lever is held per reward will be constant, only the time the lever is released can make the proportion of time allocated to lever-pressing per reward change. Since two of the three components of time allocation do not change, and time allocation itself does not change, the third component (all releases following the first lever-press following a reward delivery) also cannot change. In other words, the entire behavioural policy—when to start pressing, how long to press, and how long to release—is fixed from the time the first reward is delivered and the payoff is known.

This "filling in" mechanism has been described elsewhere (Gallistel et al., 2001) as a feed-forward model of Thorndike's (1898) law of effect. Rather than waiting for observable consequences of behaviour to inform the animal about the best course of action, the feed-forward description implies that an internal world model of the animal's situation informs the animal about the best course of action. We demonstrate here that rats indeed behave as though they have an internal world model of the price-frequency constancy, and that a single reward is sufficient in providing payoff information that will feed forward to the rat's behavioural policy for the combination of subjective opportunity cost and reward intensity it can expect to receive throughout the trial.

4.4.4 The decision to quit

In many instances, rats earn a single reward and cease to respond, our so-called "infinite evidence" conditions of the contrast between gradual-change (purely model-free) and step-change (at least partly model-based) descriptions of the task of learning the subjective opportunity cost and reward intensity of the electrical stimulation on offer during a trial. This work-then-quit behaviour could be the result of either of two processes. The animal may not have ceased responding, *per se*, but has engaged in a post-reinforcement pause drawn from a distribution with a central tendency sufficiently long that its duration is censored by the end of the trial. Alternately or in tandem, armed with a world model of how trials progress (Chapter 3), the animal may have opted to wait until the leading bracket trial, with known high payoff, would be presented again. Regardless of which process drives the animal to apparently cease responding entirely, both imply the above-described "filling in" mechanism: the currently expected payoff from self-stimulation, updated in step-wise fashion following the delivery of a single reward, sets the duration of the pause to be taken before the animal begins responding. As the proportion of time allocated to self-stimulation per reward is also set by this single-step updated expected payoff, it is possible that all components of the behavioural policy that make up the molar measure of performance are also set when the payoff is "filled in." In other words, it is altogether possible that the internal representation of payoff sets not only the duration of the post-reinforcement pause, but also of each lever-press to make, of each short lever release, and of each bout of the various other activities the rat could perform in the operant chamber. We shall return to this hypothesis in Chapter 5.

4.4.5 Conclusions

The results presented here suggest that rats behave as though they have a world-model of how a test trial will progress. Not only does a single, step-change model account for initial (post-priming and post-reinforcement) pauses better than a gradual-change model, no matter how long it takes for the step- or gradual-change to occur, but that step-like change usually occurs following a single exemplar of the reward to be delivered, at least in the case of risk-less rewards. This finding runs contrary to the predictions of a purely model-free description of performance for rewarding brain stimulation in the randomized-triads design, according to which the association between actions (namely lever-pressing) and consequences is gradually adjusted as the reward prediction error is driven to zero over the course of many reward deliveries. If such a "gradual" updating process occurs, the rate at which estimates of the determinants of the decision are revised must at least be subject to the statistical properties of the environment.

Given the presumed involvement of ventral tegmental dopamine neurons in signalling the temporal difference reward prediction error that is critical to model-free accounts of performance for rewards, it would be interesting to use the methods and analyses developed here to gauge how interference with dopamine efflux affects model-free and model-based performance, if it does at all. For example, it would be possible to ascertain whether dopamine receptor antagonists like pimozide alter the rapidity of adjustments to new subjective opportunity costs and reward intensities over the long time scales over which they act, and whether that adjustment occurs step-wise or gradually.

Similarly, new optogenetic techniques allow for the selective excitation or inhibition of dopamine neurons on the millisecond time scale over which reward prediction errors are presumed to have their effect. If phasic dopamine neurons truly encode

this reward prediction error, then inhibiting them in the appropriate time window following each reward would artificially indicate to the rat that the reward it had received was always less rewarding than expected. The post-reinforcement pause would decrease continually throughout the trial as the reward was predicted to be ever smaller, until it was censored by the trial. Meanwhile, the overall proportion of time allocated to self-stimulation per reward would decrease until it reached a lower asymptote, despite the reward's constant pulse frequency.

In contrast, if phasic dopamine signalling is related to dopamine tone, which itself is related to the absolute intensity of the rewarding effect, its effort cost, or the payoff that can be expected from engaging in any other activity while in the operant chamber, as suggested by Hernandez et al. (2010), then selectively activating dopamine neurons in the same time window will continue to produce the step-like changes in initial pauses and overall performance described here. If the effect of decreased dopamine tone is on the payoff from self-stimulation or leisure activities, rather than the process by which the payoffs are updated, the rat will continue to update the payoff from self-stimulation in a single step. Enhanced dopamine signalling will simply change what that payoff is. The results, methods, and analyses presented here provide a fertile starting point for understanding how manipulations affect the animal's on-going behavioural policy. A profound understanding of how various neural circuits store, process, and implement the various components of this policy is a truly daunting task, but the tools and results described here are easily applicable to the further understanding of how human and non-human animals decide and choose.

Chapter 5

A molecular model of performance for brain stimulation reward

5.1 Introduction

Although there have been many attempts to characterize the moment-to-moment action selection problem (Pyke et al., 1977; Staddon, 1992; Montague et al., 1995; Sutton and Barto, 1981; Niv et al., 2007), few have attempted to apply their descriptions to brain stimulation rewards. This is surprising, as intra-cranial self-stimulation provides direct access to neural machinery involved in implementing the decision. The electrode is in an identifiable location and provides a reward that is not only devoid of extraneous sensory characteristics, but also from which the animal will not become sated.

Traditional models of real-time performance have been framed in the context of temporal-difference reinforcement learning, describing punctate actions (Montague et al., 1996), or punctate actions accompanied by latencies (Niv et al., 2006), as a Markov or semi-Markov decision process. In these models, the animal holds a representation of a small number of states of the world and the decisions it can make in each of those states, and either bases action selection on a cached value of the action in a state (if using model-free learning) or a tree search (if using model-based learning).

Previous chapters have revealed two major behavioural processes that seem to operate in the well-trained animal.

1. An expected payoff from self-stimulation on trial T can be inferred from cached values of the price and intensity of rewards on trial $T-1$. In other words, the rat

maintains a simple one-trial look-back world model of the state transition function from trial $T-1$ to trial T. Unlike what is proposed by model-free learning schemes, the rat does not base its action-selection policy on the reward it received on the last trial. Instead, the rat bases its action-selection policy on the payoff for the trial to come based on its estimate of the trial type that has just come to pass (trial $T-1$) and its estimate of the trial sequence it encounters (the state transition function $\hat{\mathcal{T}}$).
2. Following a very small number of reward deliveries, which may potentially be the very first reward delivered, the payoff on a test trial is known for certain and no longer requires updating. Unlike feedback-based reinforcement learning schemes, the rat does not gradually adjust its action-selection policy in response to a backward-propagating reward prediction error. Instead, the rat "fills-in" the appropriate payoff as soon as it can be known. In other words, the rat maintains a simple one-shot update rule for the payoff from self-stimulation on test trials. The ultimate result of this one-shot update process is that reward delivery provides the necessary stimulus to set the action selection mechanism for the remainder of the trial.

In light of these two findings, we provide a new account of the action selection problem. The rat develops a world model with two components: the next trial type it can expect, and the stability of the subjective opportunity cost and reward intensity on any given trial. As soon as the costs and rewards on a trial can be known, the rat then allocates its time among the competing activities that can be observed. The various activities directly observable to an investigator who has access to the record of lever presses and releases are the result of various "hidden" processes. These hidden processes, or behavioural states, generate stay durations with characteristic distributional properties, terminating on other observable activities. The rate at which one hidden behavioural state is terminated when the animal engages in a particular activity is entirely set by the payoff the animal can expect from self-stimulation activities during the trial. Real-time performance is the result of the ongoing hidden

behavioural states that have control over what the animal does, and action selection is determined by which hidden behavioural state will take control for a given level of payoff.

5.2 Introduction to the CTMC

5.2.1 Continuous-time Markov chains

Action selection is usually described in the reinforcement learning literature as a series of point events, leading from one trial state to another as decisions are made. In contrast to these approaches, we present a portrayal in which the rat has a model of the trial—that is, parameters remain constant—and a model of the session—that is, the session's triad structure—which provide the rat with an expected payoff. In our model, the rat is in one of a variety of behavioural states when it engages in a particular activity for a period of time, and the duration of time the rat spends in some states is purely a function of the payoff the rat expects to get according to its world models. The concept of a "state" moves from the external world (like the state "One tenth of a second from the state which brings rewards") to the internal world (e.g., "I am holding patiently").

Let us assume that the rat has a model of the world based on a few simple rules, like a) the first trial the rat encounters will have a high payoff, b) the occurrence of a flashing house light signifies that the state of the world has changed, c) the payoff following the flashing light depends in some way on the price and reward intensity that has just been encountered or inferred, and d) the payoff will not change until the house light flashes again. If the rat expects a high payoff, it will devote its time almost exclusively to acquiring trains of brain stimulation. If the rat expects a low payoff, it will devote its time almost exclusively to all other activities it can perform in the operant chamber, such as grooming and exploring.

When the rat spends its time doing other things, it still faces an action selection problem: it can groom, explore, or rest, but it cannot perform all those actions at once. If some process sets how long the animal performs each of those actions, an experimenter with access only to the stream of holds and releases cannot distinguish between these activities.

A continuous-time Markov chain (CTMC) describes the behaviour of a system as a series of processes that terminate with a characteristic rate. If the system truly obeys the Markov property, all one needs to know is the current state of the system to predict what the future states may be. As a result, each state of the system must have a constant rate of failure. If every state of the system has a characteristic, constant rate of failure, then the dwell times in every state will be exponentially distributed. In other words, CTMCs describe an underlying system as a set of exponentially distributed processes which each terminate with characteristic rate onto each other.

Suppose the following simple CTMC holds true: the rat begins a trial with a post-priming pause, which terminates with characteristic rate on a hold, which terminates with characteristic rate on a release, which terminates with characteristic rate on a hold, until the whole system is stopped when the cumulative amount of time spent holding reaches the price or the cumulative amount of time in all these activities reaches the trial time. In this chain, future activities can be predicted on the basis of only the current activity: there is some constant probability that the current activity will fail and its termination will lead with a particular known probability to another activity. Action selection in this chain is simply the implementation of a characteristic termination rate when the animal is performing a particular activity. Dwell times in each activity would be the directly observable result of the action selection process in real time. Overall trial performance in this case emerges from the differential probabilities that the rat will begin and cease to perform various activities, such that trials where the payoff from self-stimulation is high will favour holding and

discourage pausing or releasing, while trials where the payoff is low will favour pausing and releasing over holding.

The directly-observable activities the rat engages in are not likely to be purely Markov processes. If that were so, the dwell times in the various activities we can detect would be sampled from an exponential distribution that peaks at 0 and decays constantly over time. Instead, the rat likely spends some minimum amount of time performing an action. At the very least, it spends at least as much time performing the action as we can detect. The amount of time the rat spends in any particular activity might therefore be better modelled as the result of a non-exponential process whose termination rate increases over small time frames and approaches exponential behaviour the longer the rat has been performing that action.

When the distribution of dwell times in directly observable activities is not exponential, the chain can no longer be strictly called "Markovian," because the probability of switching from activity A to activity B is no longer constant over time. The probability of switching from A to B when A terminates may still be constant, so there is still a "Markovian" element to the chain: assuming A has stopped, there is a constant probability that the next activity will be B. Chains of this type are called "semi-Markov," since termination rates may not be independent of time, but are constant at transition points. The model we present here is a continuous-time, semi-Markov chain involving activities which can be directly observed, and behavioural states which may not.

5.2.2 Previous models

5.2.2.1 Melioration

Matching refers to the observation that on concurrent variable-interval (VI) schedules of reinforcement, animals will tend to allocate their relative rates of re-

sponding to each alternative in proportion to the relative rates of reinforcement. The response requirement on any given operant is sampled from an exponential distribution or a geometric approximation thereof, the effect of which is that there is a constant probability per unit time that engaging with the manipulandum will deliver rewards. Importantly, the VI schedule of reinforcement traditionally used to study matching has an infinite hold: as soon as it is armed, the manipulandum remains armed until the next response is made. These two features ensure that matching is very nearly optimal: no other pattern of relative responding to the concurrently available alternatives will substantially increase the number of rewards an animal may obtain. In single-operant contexts, matching has been generalized (de Villiers and Herrnstein, 1976; McDowell, 2005) such that response rate is related to the rate at which experimenter-controlled rewards are delivered relative to the other rewards that an animal may derive from the operant chamber context, such as grooming, exploring, and resting. Although some of the assumptions of matching have been found not to hold in the case of brain stimulation rewards (Conover et al., 2001b), it has been a useful framework for studying decision-making in animals.

Since matching is very nearly the optimal strategy in concurrent VI schedules, some (Williams and Royalty, 1989) have argued that matching is the result of a maximizing algorithm. At any point in time, the animal may emit responses with the highest probability of reinforcement, which would result in approximately matching behaviour. However, Herrnstein and Heyman (1979) tested animals on a concurrent VI/VR schedule of reinforcement, under which pigeons allocated their pecking between one key providing 3.2 seconds of access to a food hopper according to variable intervals with means of 15, 30, or 45 seconds and another providing access to the food hopper according to variable ratios with means of 30, 45, and 60 responses. In their hands, although pigeons matched their relative responding to the relative obtained rates of reinforcement, their performance did not maximize the overall rate of rein-

forcement. Pigeons had a strong bias toward the VI schedule, rather than the VR schedule predicted by a maximization account.

One proposed mechanism (Vaughan, 1981) by which matching occurs is melioration. According to this hypothesis, if the local rate of reinforcement (the number of reinforcements obtained while working at a manipulandum per unit time spent there) of an alternative is less than the local rate of reinforcement obtained elsewhere, the animal will be pulled toward the richer of the two. For example, if pecking key A provides grain at a rate of 1 every 45 seconds and pecking key B provides grain at a rate of 1 every 15 seconds, as the pigeon spends more and more time pecking at key A, the probability that a response at B will be reinforced increases. At the point where the local rate of reinforcement from A begins to fall below that of B, the pigeon will switch to responding on key B. As it spends a greater amount of time at key B, the probability that a response at A increases, until the local rate of reinforcement from A exceeds that of B, resulting in another switch. This result obtains because the reward is held in trust so long as the rat has not yet lever-pressed; as time marches on, the probability that the lever is armed will increase and the rat will surely be rewarded for a lever press. If the pigeon were to peck at a rate of once every 15 seconds on key A (delivering rewards at rate 1/45), it would receive a reward once every three pecks; if the pigeon simultaneously pecked at a rate of once every 45 seconds on key B (delivering rewards at rate 1/15), it would receive a reward once every press. Melioration proposes that because the local rate of reinforcement on pecking key B is greater than that on pecking key A, the pigeon will allocate more pecks to B. If the pigeon now pecks once every 5 seconds to B and once every 60 seconds to A, the local rate of reinforcement will be 1 pellet per 3 responses at B and 1 pellet per response at A. When the local rate of reinforcement from A exactly matches that of B—say the pigeon pecks at A at a rate of once every 15 seconds (local rate of 1/3) and at B at a rate of once every 5 seconds (local rate of 1/3)—the pigeon is also

matching its relative rate of responding ($\frac{1/15}{1/5}$) to the relative rate of reinforcement ($\frac{1/45}{1/15}$). Melioration proposes that the differential local rate of reinforcement drives the animal to alter its responding to the option providing rewards at the higher local rate until there is no difference in the local rate of reinforcement from the two options available.

Data from Gallistel et al. (2001) imply that the factor controlling allocation of responses to alternatives is not their relative local rates of reinforcement. Stay durations for levers on either side of an operant chamber would be expected to fluctuate not only with changes in the programmed rate of reinforcement, but also with unusually long intervals sampled from the variable-interval distributions. If local rates of reinforcement were the source of matching behaviour measured on the molar level, then rats working under concurrent VI/VI schedules delivering electrical brain stimulation would be influenced by unusually low rates of reinforcement that may occur simply by randomly sampling intervals from an exponential distribution. Unfortunately for the melioration hypothesis, the distributions of stay durations (as assessed by log-survivor plots) at a wide range of immediately-preceding inter-reward intervals were virtually superimposable. It mattered little whether the immediately previous local rate was low (that is, the inter-reward interval was high) or high: the probability of remaining at the lever per unit time was identical regardless.

In explaining these data, the authors advocated a feed-forward control of performance, rather than the feed-back system implicit in melioration. Feed-back systems require the rat to select the action entirely according to immediate consequences: responding at a rate of 1/15 seconds at A produces a local rate of reinforcement of 1/3 responses, while responding at responding at a rate of 1/45 seconds at B produces a local rate of reinforcement of 1/1 responses. In contrast, a feed-forward system maintains a representation of the environment that allows an animal to quickly react when it perceives changes to its situation. They propose that one system evaluates

the rate of reinforcement from the options, and another evaluates whether that rate of reinforcement is likely to have changed. If a change in the overall rate is likely to have occurred, the animal adjusts its performance as quickly as could be done in principle. If the environment is highly variable or highly stable, the threshold for identifying a change in rate is tuned accordingly. Note that in the feed-forward case, the rat's actions are not the result of the immediate consequence of its actions. Instead, the rat's actions are the result of a model of its environment.

Given the evidence suggesting a feed-forward, rather than feed-back mechanism for the emergence of matching, how can one model the moment-to-moment basis for the behaviour observed on a molar scale?

5.2.2.2 Model-based reinforcement learning

One feed-forward description of real-time performance involves endowing the experimental subject with a model of the identity, magnitude, and cost of the rewards to come. For example, a master chess player may have a model of all board configurations their opponent may respond with for each piece they could move. The current configuration of the board provides a stimulus that informs the player which action is optimal if he wants to achieve the goal of a check-mate only if the player also maintains a mapping of how the board *can* evolve and projects this mapping sufficiently into the future. In an operant context, a pigeon may have learned that responding at a blue pecking key an average of 30 times leads to 4 seconds of access to a water bowl, while responding at the red pecking key leads to 3 seconds of access to a food hopper approximately every 45 seconds. In this case, the mapping of pecking keys to the magnitude of the water and food rewards, their identity, and their costs are learned and encoded in memory.

Model-based reinforcement learning has been especially useful in the context of goal-driven evaluative systems. Importantly, a habit-driven (that is, not model-based)

evaluative system maintains a representation of magnitude but not identity, making it resistant to reinforcer devaluation. If all that is learned is that the blue pecking key delivers 1 arbitrary unit of reward while the red pecking key delivers 2 arbitrary units of reward, then subsequent devaluation of the food delivered by the red pecking key but not the water delivered by the blue pecking key will not affect a habit-driven system. In contrast, a goal-driven (that is, model-based) evaluative system maintains a representation of both identity and magnitude, making it vulnerable to reinforcer devaluation.

In the model-based formulation, some stimulus is necessary to provide the appropriate information. As discussed in Chapters 3 and 4, rats working for brain stimulation rewards in the randomized-triads design behave as though they have a model of the triad progression of trials within a session, cued by the inter-trial interval, and a model of the progression of reward encounters within a trial. The appropriate stimulus, however, is a vector of subjective determinants of decision: trials during which the reward intensity is high and subjective opportunity cost is low are cued by the occurrence of an inter-trial interval and the combination of low reward intensity and subjective opportunity cost on the trial that preceded the inter-trial interval cue. Moreover, this vector of intensity and opportunity cost does not need to be explicitly uncovered, as rats rarely respond on trailing bracket trials of a repeating triad of leading, test and trailing trials. This implies that the expected combination of intensity and opportunity cost may be updated, but even when it isn't, it provides the appropriate signal of which of the three trial types will follow.

Within a test trial, the payoff is unknown prior to the delivery of the first reward. As soon as the first reward is delivered, at least in the case of risk-less rewards, rats behave as though they have a constant behavioural policy for the duration of time they are earning rewards. Post-priming and post-reinforcement pauses, as well as the overall proportion of time spent harvesting rewards, change abruptly between

the time the subjective opportunity cost and reward intensity are unknown to the time they are, in principle, completely determined. Here again, key determinants of decision-making (opportunity cost and reward intensity) provide the necessary stimulus for implementing a complete behavioural policy.

If the stimuli that signal the appropriate action to be taken are the determinants of the decision itself, a model-based description of the task simplifies to a simple filling-in process. When there is no objective variability in the opportunity cost, reward intensity, and probability of reinforcement, the rat can simply pursue a behavioural policy based on its best current estimate of what the scalar combination of these decision variables will be. On this view, the representation of the payoff from self-stimulation competes with that of extraneous activities, which each set the probability that rats will engage in "work-related" or "leisure-related" activities.

A similar model has been formulated by Gallistel et al. (2001). In modelling performance on concurrent VI/VI schedules of reinforcement, they assumed that stays on either side were stochastic, exponential processes related to the combination of reinforcement rate and magnitude (the income) for that side. When conditions fluctuate frequently, the rat obtains a small number of exemplars following the point at which a change is perceived to have occurred, re-estimates the obtained incomes, and adjusts performance accordingly. When they are stable for long periods of time, the rat's estimates reflect a compromise between new and old rates.

5.2.3 Description of the mixture model

In the model we present here, the scalar combination of subjective opportunity cost, reward intensity, and probability determine the leaving rate (and thus its reciprocal, the stay duration) of a wide array of behavioural states composing the various observable activities in which the rat engages. External stimuli, such as the flashing house light, signal a change in trial type and that conditions will be stable

until the light flashes again. During test trials, the first reward delivery allows the rat to revise its initial estimate of the payoff in a single step, providing the rat with an appropriate internal stimulus that dictates how it ought to allocate its time. Prior to this, the rat's performance is strongly influenced by its uncertainty of the payoff to come, setting a consistent behavioural policy at this phase of the trial that differs from the later, stable phase of the trial that follows the first reward delivery.

5.2.3.1 Activities

In our treatment of real-time performance, the animal may be engaged in one of six activities that are directly evident from the stream of holds and releases at the lever. These behavioural activities are: post-priming pauses (PPPs), post-reinforcement pauses (PRPs), holds, taps, true leisure bouts (TLBs) and censored true leisure bouts. We discuss each activity below.

Post-reinforcement and post-priming pauses occur between the time the lever extends into the operant chamber to the first time the lever is held. The first such pause in a trial, as defined in Chapter 3, is the PPP, as it occurs 2 seconds following the onset of the constant priming stimulation that is delivered during the inter-trial interval. Post-reinforcement pauses, in contrast, occur following every subsequent lever extension, usually (and in the case of rewards delivered with certainty, always) preceded by the delivery of reinforcement.

Holds refer to every depression of the lever. When the subjective opportunity cost of the reward is sufficiently low, there may be many holds, all censored by lever retraction. As the opportunity cost increases, there may be fewer holds, and when it is sufficiently high, any given hold may be much less likely to result in the completion of the work requirement.

Finally, lever releases occurring after the lever has been held may either be short or long (Conover et al., 2001b). When the reward intensity is high and the

subjective opportunity cost is moderately low, rats will often release the lever for short periods of time, quickly re-engaging the lever. These short pauses, uncensored by the end of the trial and lasting under a second, we define as taps. They may be included in the psychological perception of work, since the rat is close to the lever, and may even still have a paw on the manipulandum while the lever is released. In contrast, the rat may switch to a different goal after depressing the lever, leaving it to explore, groom and rest. These longer pauses, lasting a second or more, we define as true leisure bouts, as we presume that the animal is engaged in activities unrelated to self-stimulation and related, instead, to the extraneous leisure activities it can perform in the operant chamber.

Censored true leisure bouts (CTLBs) refer to pauses of any type (following lever extension at the start of a reward encounter or following a hold in the midst of a reward encounter) that are censored by the end of the trial but last longer than 1s; when they occur, the animal has engaged in an activity that lasts at least as long as the entire trial. A 1s criterion was used to ensure that very short pauses that were censored by the end of the trial were not counted as CTLBs. We have isolated CTLBs from PPPs, PRPs, taps and TLBs to facilitate the extrapolation of performance in real time to the whole trial.

5.2.3.2 Hidden states

Since there is no way to know exactly what the animal is doing in the operant chamber from glancing at the stream of lever releases and holds that is available, each behavioural activity (PPP/PRP, CTLB, hold, tap, and TLB) may comprise various "hidden" behavioural states. For example, TLBs may include grooming, exploring and resting bouts that each have characteristic termination rates. The distribution of dwell times in TLB activities would therefore be a mixture of multiple behavioural states that the animal is in and which would be directly observable only given the ap-

propriate equipment. Moreover, hidden behavioural states may produce behavioural patterns that would not be observable even in principle. For example, the hold activity may include a mixture of "patiently waiting" and "impatiently tapping" states, or tapping-related releases may include long-interval and short-interval distributions. In each case, only a careful interrogation of the distributions of dwell times making up each activity would reveal the number and characteristics of hidden behavioural states that compose them.

5.2.3.3 Expected payoff

In our modelling, we have assumed that the most important determinant of dwell times within each hidden behavioural state is the expected payoff from self-stimulation. Given the degree to which the mountain model, a molar model of performance for rewarding brain stimulation, can explain performance for risk-less and risky options based on a scalar combination of subjective reward intensity, opportunity cost, and probability of reinforcement (the payoff), we consider the most important determinant of the behavioural allocation function to be the payoff that can be expected from self-stimulation activities and the (presumably constant) payoff from non-self stimulation activities. Our choice of determinant for the action-selection process was motivated by a feed-forward model (Gallistel et al., 2001) according to which the leaving rate on one side of a concurrent VI/VI schedule is proportional to both the expected income (or payoff) derived from the other side and the expected income relative to all sources of reward. Similarly, we have assumed that the rate at which a hidden behavioural state is terminated is a function of payoffs that may be obtained from self-stimulation and from extraneous leisure activities.

The model proposed below assumes that performance on the molar level is the result of the rat filling in a value for the payoff it can expect from self-stimulation, which biases the durations of the underlying behavioural states it will engage in.

When a hidden state terminates uncensored, a hidden state from a different behavioural activity is begun. Although each activity we can observe is composed of the same underlying behavioural states with particular distributional characteristics (like a shape and scale), their mean depends on the expected payoff. The animal's pattern of responding is therefore completely, though non-deterministically, described by the payoff it can expect to receive. In other words, the payoff from self-stimulation will bias the animal to spend more time in some activities and less time in others as a result of altering the rate at which those activities terminate.

When that payoff is not known with certainty, as is the case on test trial types in the randomized-triad design, the first reward encounter will have different, though consistent characteristics: the usual trade-off between exploration and exploitation is disrupted. During this first reward encounter, the outputs of exploration and exploitation both consist of work. As a result, the animal would be biased toward working longer. Until the subjective opportunity cost is sufficiently high to indicate to the animal that no matter how great the subjective intensity of the reward to come, self-stimulation will not be worthwhile, the rat will depress the lever until a reward is delivered.

We present here a two-phase model of performance for rewarding brain stimulation, in which the first phase is characterized by a short PPP and a single, long, continuous work bout (including holds and taps), while the second is characterized by a continuous-time semi-Markov chain of hidden behavioural states, the probability of which is determined by the payoff from self-stimulation, and for which leisure-related activities (PRPs and TLBs) are highly sensitive to alterations in payoff while work-related activities (holds and taps) are not.

5.3 The CTMC

The model we shall present below is, in essence, hierarchical. The various directly-observable activities (PRP, hold, tap, TLB) are made up of various hidden behavioural states. The probability that an animal engaged in an activity has entered a particular state is constant across payoffs. The probability that the particular state terminates may or may not be payoff-dependent. Behavioural states can be collectively observed through various activities that transition to each other in often trivial ways. For example, the probability of transitioning from a PRP, tap or TLB to a hold is one. If a TLB is going to terminate uncensored by the end of the trial, it must necessarily be followed by a hold. Similarly, if a hold is going to terminate uncensored by lever retraction following successful completion of the work requirement, it will necessarily terminate on either a tap or a TLB. The relative probability of each transition (hold to tap contrasted with hold to TLB) is also a function of the current expected payoff. Finally, the probability per unit time of ceasing responding for the remainder of the trial may also be a function of the currently expected payoff.

Figure 5.1 is a schematic of how we have modelled the entire reward encounter, from the time the lever is extended into the chamber to the time it is retracted, starting from the time the rat has estimated the payoff. Prior to the time the subjective opportunity cost and reward intensity are revealed, the rat's performance reflects the payoff it expects to receive.

Five labelled white boxes represent the five different activities available after the first reward has been delivered. Within four of those activities (post-reinforcement pause, hold, true leisure bout, and tap), there are multiple hidden behavioural states. Although activities can transition to each other, for simplicity of modelling, hidden behavioural states do not. When one hidden behavioural state terminates (for example, a short hold), a completely different activity is begun (for example, a tapping-related

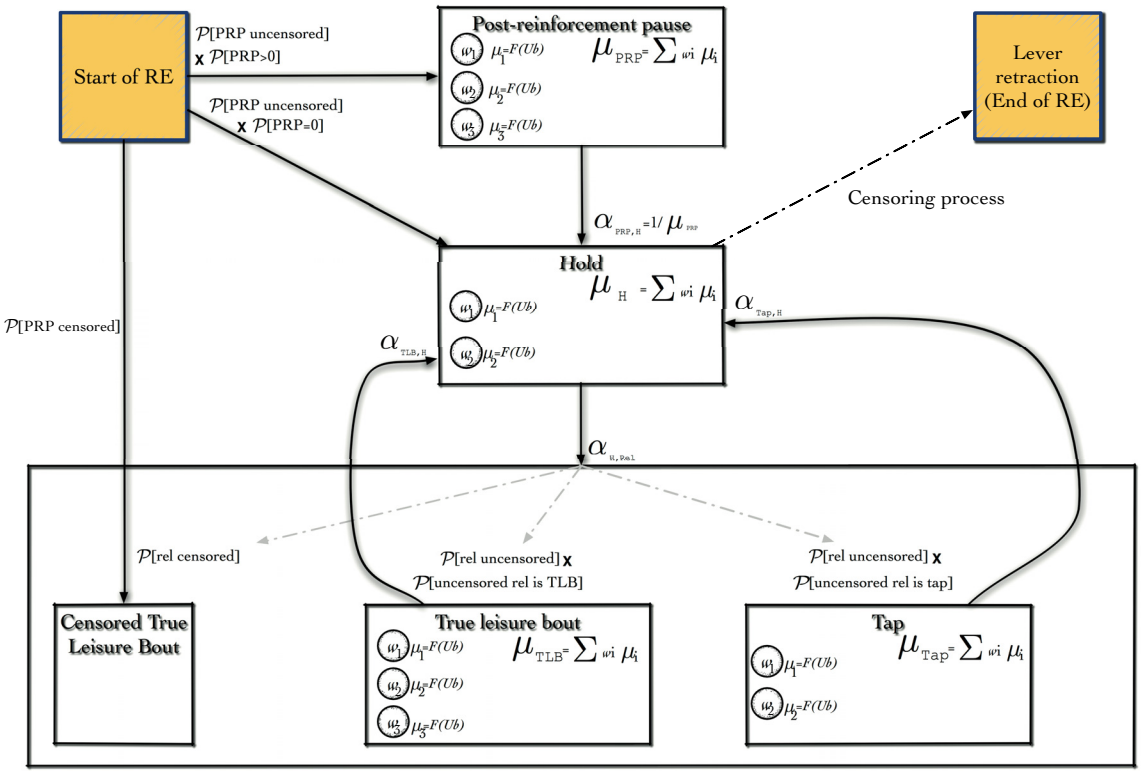

Figure 5.1. Molecular model of a reward encounter. When the RE begins (top left corner), the rat may initiate an uncensored post-reinforcement pause, a hold, or a censored pause. These activities (white boxes) comprise multiple hidden behavioural states (small circles) with different probabilities of occurrence (w_i) and within which the rat will spend a possibly payoff-dependent time ($\mu_i = F(U_b)$). When the rat leaves a hidden behavioural state, it transitions at rate $\alpha_{S,S'}$ to a new activity, like a hold, or one of multiple different releases, such as TLBs, taps, and CTLBs. When the hold has been visited for a cumulative amount of time, the hold time is censored by lever retraction, at which point a new reward encounter may begin.

release). The preponderance of a hidden behavioural state in generating dwell times for the activity in question is w_i (its weight). The payoff estimate for the trial determines the mean μ_i of the distribution from which dwell times of a hidden behavioural state will be drawn. The expected dwell time in an activity is the convex combination of its component parts ($\sum_i w_i \mu_i$ where $\sum_i w_i = 1$), and activity S transitions to activity S' by following the arrows in the activity-transition diagram presented in figure 5.1 at a rate of $\alpha_{S,S'}$.

The reward encounter proceeds as follows. The lever extends into the operant chamber, which begins the reward encounter (yellow box, top left corner). The rat may cease responding altogether with probability $\mathcal{P}$[PRP censored] set by the payoff, or if not (with probability $\mathcal{P}$[PRP uncensored]), may begin pressing immediately with probability $\mathcal{P}[PRP = 0]$. If the first pause in responding is neither censored by the end of the trial, nor shorter than can be measured, the rat begins a PRP of non-zero duration. This PRP activity is actually the result of multiple (in the schematic, three) hidden behavioural states. The probability that the rat is in behavioural state i while performing the PRP activity is w_i, and the duration for which the rat is in said state i is given by μ_i. As a result, the expected duration of the PRP (μ_{PRP}) is the convex combination $\sum_i w_i \cdot \mu_i$, where $\sum_i w_i = 1$. The rate at which the PRP is terminated is $1/\mu_{PRP}$, and this termination leads to a hold. When a hold begins, it, too, is the result of multiple (in the schematic, two) hidden behavioural states, and the expected duration of the hold (μ_H) is the convex combination of these holding-related hidden behavioural states.

The hold terminates, when uncensored, on a release. The probability that a lever release will be censored by the end of the trial is

$$\mathcal{P}[\text{Rel censored}];$$

if the release is not censored by the end of the trial (with probability $\mathcal{P}$[Rel uncensored]), it will either be a short, tapping-related release occurring with probability

$$\mathcal{P}[\text{uncensored rel is tap}]$$

or a long, leisure-related TLB release occurring with probability

$$\mathcal{P}[\text{uncensored rel is TLB}].$$

Both the TLB and tap activities are made up of hidden behavioural states, and as with the PRP and hold activities, their termination rate is the reciprocal of the convex combination of their respective hidden behavioural states. The only state in which the rat is presumed to remain once entered is the CTLB state, which can be reached with probability

$$\mathcal{P}[\text{PRP censored}]$$

if the rat begins the CTLB at the start of a reward encounter, or

$$\mathcal{P}[\text{Rel censored}]$$

if the rat begins the CTLB following a hold. When the cumulative time that the lever has been held reaches the price, the hold bout will be censored by the retraction of the lever, possibly followed by reward delivery. At this point, a new reward encounter may begin.

In addition to indirectly setting the termination rate of an activity, we have assumed that the payoff can set four separate probabilities:

1. the probability that the rat begins the reward encounter in the CTLB state

$$\mathcal{P}[\text{PRP censored}]$$

2. the probability that the rat begins the reward encounter with a PRP of 0 duration

$$\mathcal{P}[PRP = 0]$$

3. the probability that a long pause in responding is censored by the end of the trial

$$\mathcal{P}[\text{Rel censored}]$$

4. the probability that an uncensored pause in responding is less than 1s due to lever tapping

$$\mathcal{P}[\text{uncensored rel is tap}]$$

All other transition probabilities are either trivial (1 or 0) or calculable from the above four relationships (e.g., $\mathcal{P}[\text{PRP uncensored}] = 1 - \mathcal{P}[\text{PRP censored}]$). It is possible that one or more of these probabilities is payoff-independent, as we will empirically demonstrate later.

Figure 5.2 is a schematic of how we have modelled the dependency of dwell times in an activity on the expected payoff. Following the first reward delivery, and every reward delivered thereafter, the rat has an up-to-date estimate of the payoff from self-stimulation. This estimate sets the duration of an activity: while the relative composition of the hidden behavioural states making up an activity remains the same, their hazard functions will reach a different final level, thereby changing the expected duration of the activity. The rat then remains in a given behavioural state for a period of time sampled from the appropriate distribution of dwell times or until its sojourn is censored by an experimenter-enforced event, such as retracting the lever

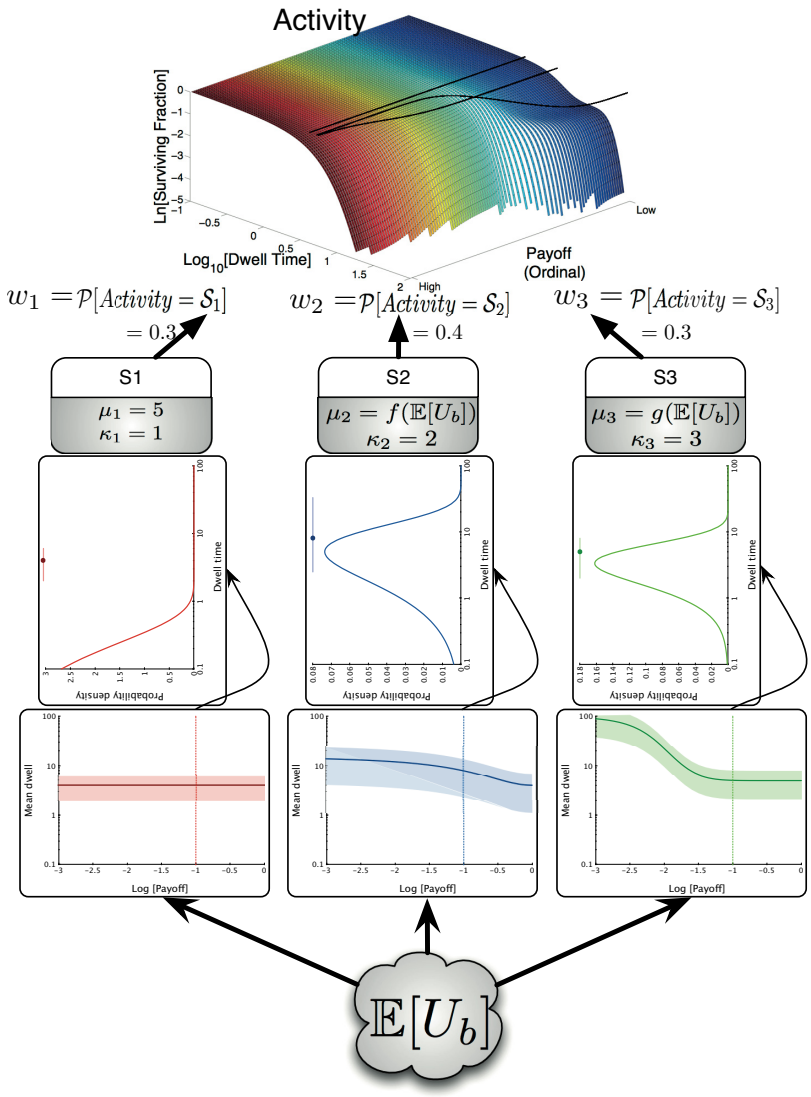

Figure 5.2. Dependency of dwell times on payoff. The expected payoff from self-stimulation ($\mathbb{E}[U_b]$) is presumed to set the mean dwell time in any payoff-dependent hidden behavioural states. As any observed activity is the result of a dwell time chosen with some probability from one of its component hidden states, the dwell times in an activity will be stochastic realizations of the distribution that characterizes a particular state at a particular payoff level, with mean μ_i and shape κ_i. The distribution from which the dwell time is selected is chosen with probability $w_i = \mathcal{P}[Activity = \mathcal{S}_i]$. Bottom row shows the dependence of the mean of each hidden state (y-axis) on payoff (x-axis); horizontal lines and shaded areas represent the predicted mean and standard deviation of the distribution setting dwell time durations. Vertical lines indicate from where, in the panels above, dwell times are sampled. The middle row shows the probability density function (rotated 90 degrees) of each hidden state. At the top, an example of the resulting log-survival plot is depicted as a function of the log-dwell time and payoff, where hot colours signify high payoffs. Overlain on this 3D log-survival plot are the three functions setting the means of the hidden states that make up this hypothetical activity.

and initiating a blackout delay or inter-trial interval. In the schematic example, high self-stimulation payoffs reduce the amount of time spent in states S_2 and S_3 when performing the activity, but the mean dwell time in state S_1 is constant across payoffs. As a result of its effect on the means of individual hidden states of an activity, the payoff will alter the expected sojourn in that activity: overall, the rat will spend less time performing this hypothetical activity by virtue of spending less time in each of its component behavioural states. At the top of the figure, we show the log-survival function plotted at each log-dwell time (y-axis) and payoff (x-axis, it is inverted such that ribbons closer to the reader represent high payoffs and those farther away are low payoffs) that results from the mixture of payoff-dependent and independent states.

For simplicity and symmetry, we expected that dwell times in leisure-related activities would be sensitive to the payoff from self-stimulation, and dwell times in work-related activities would be sensitive to the payoff from everything else. To test this, we first checked whether the dwell times in each activity were dependent on the payoff from self-stimulation. We could therefore assess whether the payoff from self-stimulation affected the activities we presumed they would (PRP, TLB) and did not affect the activities we presumed they wouldn't (holds, taps).

The CTMC assumes that dwell times in any given state will be independent of history. This includes the time spent in the PRP; in other words, the duration of the PRP on one reward encounter can not be dependent on performance during prior reward encounters if the Markov property is to hold. One would expect the PRP on reward encounter n to be highly similar to that on reward encounter $n-1$ if the constant payoff from self-stimulation throughout the trial had set them both. In a scenario where the Markov property does not hold, the duration of the PRP bears some relationship, above and beyond the payoff, to activities performed on the preceding reward encounter, such as the amount of time spent holding and tapping. To test this, we performed a two-stage hierarchical linear regression on the logarithm

of each PRP. The first stage of the hierarchical linear regression predicted the log-PRP duration based on the logarithm of the payoff from self-stimulation for the trial from which it came. The second stage predicted the log-PRP duration from both the log-payoff and the log-corrected work bout (which sums consecutive holds and taps) from the reward encounter that just preceded it. If this second model predicts additional variance in log-PRP durations that cannot be attributed to the log-payoff, the Markov property does not hold in the case of PRPs with non-zero duration.

In the case of TLBs, the Markov assumption means that the time spent engaged in the TLB activity is sampled with the same probability from any of the behavioural states that make it up, no matter how long the animal spent holding or tapping beforehand. One could argue that following a longer work bout (that is, holding and tapping), the exhausted rat will take a longer true leisure bout. We assessed the degree to which this was true by correlating the duration of every TLB activity with the duration of the corrected work bout that immediately preceded it. A strong positive correlation would imply that these two activities do not obey the Markov rule: the duration of a TLB would be dependent on the duration of time spent in hold and tap activities. In contrast, the absence of a correlation would imply that the duration of a TLB is independent, at least to a first-order approximation, of the rat's lever-pressing history.

5.4 Methods

5.4.1 Behavioural protocol

Rats were the same as in Chapters 2, 3 and 4. Rats were presented with triads of trials, starting with a leading bracket trial presenting high-frequency stimulation at a 1s price, a test trial presenting stimulation of a frequency and price that was pseudo-randomly drawn without replacement from a list, and a trailing bracket trial

presenting low-frequency stimulation at a 1s price.

In the case of rats DE1, DE3, MA5, DE7, PD8, DE14, DE15, DE16, DE18 and DE19, test trials were drawn from a list containing 3 pseudo-sweeps for each pair of reinforcement probabilities presented (1 vs 0.75, 1 vs 0.50, 1 vs 0.50 with lever locations switched, and 1 vs 0.75 with lever locations switched). Performance during each probability condition of each pair was analysed separately. Pseudo-sweeps were sets of pairs of prices and frequencies either arrayed along the frequency axis at the same 4s price (frequency pseudo-sweep), arrayed along the price axis at the same high frequency (price pseudo-sweep), or arrayed along a ray that passed through the intersection of the previous two sweeps and the presumed F_{hm} and P_e values for the mountain fit described in Chapter 2 (radial pseudo-sweep).

In the case of rats F3, F9, F12, F16, F17 and F18, test trials were drawn from a list containing 9 pseudo-sweeps. Three of the pseudo-sweeps were identical to those presented above. The other 6 were price-frequency pairs arrayed along the frequency axis at 0.125s, 0.25s, 0.5s, 1s, 2s, and 8s prices.

Performance presented throughout this chapter was obtained exclusively during test trials of the randomized-triads design.

5.4.2 Statistical procedures

We began by first testing some of the assumptions of the model by determining the proportion of variance, if any, that could be accounted for by previous performance when controlling for the payoff in effect during the trial. It was then possible to identify the various components of the continuous-time semi-Markov chain (CTMC), by fitting increasingly complex mixtures of exponential and gamma distributions to the dwell times observed in four, non-overlapping categories of behavioural activities (PRPs, holds, taps and TLBs) and transitions to a fifth (CTLBs) from the first reward delivery onward. These components, and the payoff-dependent probabilities

of transitioning from one activity to the next, were individually fit for each animal.

When the various components of the CTMC had been identified, we extrapolated what molar performance, from the first reward delivery onward, would be if the rat's behavioural policy were governed by this continuous-time semi-Markov model. The extrapolation provides both a qualitative and quantitative description of how well the CTMC accounts for performance not only on the molecular level, which is its stated purpose, but also on the molar level.

5.4.2.1 Testing the Markov assumption for PRPs

Our CTMC assumes that every reward encounter following the first reward delivery is essentially the same: whether one considers the beginning, middle, or end of the trial, the animal behaves as though it passes through a number of hidden behavioural states, measurable as activities, and either obtains a reward or stops responding altogether. We sought to test whether the duration of the PRP was dependent in any way on previous lever-pressing activities by performing a linear regression of the logarithms of the PRP durations (excluding, of course, nil durations) against the logarithms of the total corrected work bouts that preceded them.

If the time spent working and the post-reinforcement pause are both related to the payoff from self stimulation, then the time spent in each of those activities will necessarily be related to each other. If the Markov property does not hold for PRPs, then there will be variability in PRP duration that can be attributed to the time spent in previous activities, such as the last work bout, and not to the payoff alone. We calculated the proportion of variance in log-PRP duration that could be uniquely attributed to the payoff and not the duration of the corrected work bout that resulted in the reward delivery, and the proportion that could be uniquely attributed to the last corrected work bout and not the payoff. A single corrected work bout was the sum of all hold times and tap times that were uninterrupted by an intervening

TLB. For example, if the rat held for 1s, released for 0.5s, and held again for 0.75s before engaging in a TLB for 10s, the total corrected work bout that preceded the 10s TLB was 2.25s. We isolated the bouts from the period of time following the first reward delivery, that is, from reward encounters 2 onward, and included only those work bouts that were censored by lever retraction. These proportions were assessed by comparing a "larger" model to two "smaller" models:

1. Using only log-payoff as a predictor,
2. Using only the log-duration of the last corrected work bout as a predictor, and
3. Using both log-payoff and log-duration of the last corrected work bout as a predictor.

The proportion of variance accounted for by model 3 results from two predictors. Subtracting the proportion of variance accounted for by payoff (model 1) provides an estimate of the increase in predictive power that the last corrected work bout alone contributes to log-PRP duration. Subtracting the proportion of variance accounted for by the last corrected work bout (model 2) provides an estimate of the increase in predictive power that the payoff alone contributes to log-PRP duration. If the Markov property did not hold, the duration of the last corrected work bout would have to have had at least some influence on the duration of the subsequent PRP. As such, we extracted the proportion of variance in log-PRP that could be uniquely attributed to the log-last corrected work bout, as well as that which could be uniquely attributed to the log-payoff.

5.4.2.2 Testing the Markov assumption for TLBs

A key assumption of the CTMC is that transitions from one state to another are independent of previous states. Although the rat may enter a state that is characterized by a gamma distribution, and its dwell time in that state will therefore not be independent of the amount of time already in the state, when that state terminates,

the Markov property assumes that the rat transitions to other states with constant probability. It does not matter that the rat has been transitioning between holds and taps for a long or a short time—the duration of the TLB when the work bout terminates uncensored is presumed to be drawn from the same mixture of distributions that each characterize a hidden behavioural state.

To determine whether the duration of any TLB depended on the duration of time spent working (that is, all holds and taps) before that point, we correlated TLB durations with the total corrected work bout duration that preceded them. The total proportion of variance in TLB duration accounted for by the last corrected work bout was extracted across all price-frequency pairs presented to each animal in each condition.

5.4.2.3 Inferring hidden behavioural states

As a first step, we identified whether the expected dwell time in an activity was payoff-dependent by extracting the maximum-likelihood estimate of the mean dwell time for the activity at each payoff in all subjects and conditions together. The data were then restricted to the range of payoffs that was common to all animals and conditions, and we performed a linear regression of the log-maximally likely mean dwell time onto the logarithm of the payoff. The deviance of this regression is a measure of error, and the deviance of a null model for which payoff has no effect on dwell time (an intercept-only model) is a measure of total variability. Their difference is the variability predicted by payoff. The proportion of the total variability that was explained (a pseudo R^2 statistic) was then calculated for each activity. It is on the basis of these R^2 statistics that we chose to model dwell time as payoff-dependent or independent.

Each non-overlapping behavioural activity (PRP, hold, tap, TLB, CTLB) could comprise multiple hidden components. In order to balance comprehensive-

ness with parsimony, we fit increasingly complex mixtures of exponential and gamma distributions to the set of dwell times in each activity, for which the mean of each component of the mixture was potentially a logistic function of the logarithm of the payoff, with constant shape parameter (in the case of gamma distributions) and mixture proportions. Exponentially-distributed components were assumed to be payoff-invariant, while the logarithm of the mean of gamma-distributed components could be a logistic function of payoff. Payoff-sensitive, *exponentially*-distributed components are a special case of payoff-sensitive, gamma-distributed components, where the shape parameter equals one, and payoff-insensitive gamma-distributed components are a special case where the slope of the logistic function of payoff is 0.

If payoff-dependent dwell times could be justified, we fit up to 5 components, starting with one gamma-distributed component whose mean was a scaled sigmoid function of payoff. We then attempted a fit with one exponential component whose mean was constant across payoffs and one payoff-dependent gamma-distributed component. In each attempt, first the number of gamma components was incremented (from 1 to the number of components), then the total number of components. This continued until the weight of any component was below 0.1 on a fit attempt or the number of components would have become 6.

If the assumption of a payoff-dependent dwell time was not justified, we fit up to 5 components, starting with one exponentially-distributed component whose mean was constant across payoffs. We then attempted a single gamma-distributed component. In each attempt, first the number of gamma-distributed components was incremented (from 0 to the number of components), then the total number of components. This continued until the weight of any of the components of a fit attempt was under 0.1, or the number of components would have become 6.

We then identified among all combinations of exponentially- and gamma-distributed hidden states that which had the lowest Akaike Information Criterion

(AIC, Akaike, 1976). Attempts to fit too complex or too parsimonious a model to the set of dwell times in an activity would have high AIC values, while a model that was sufficiently complex to explain the data without extraneous parameters would have low AIC values.

The AIC is defined as twice the number of parameters minus twice the logarithm of the likelihood function for the maximally likely set of fit parameters:

$$AIC = 2n_{parm} - ln(\mathcal{L}).$$

The probability of n data points, as a function of k mixing proportions w and distributional parameters Θ (where Θ corresponds to the mean in the case of exponential distributions, a vector with shape and scale parameters as entries for payoff-independent gamma distributions, and a vector with logistic and shape parameters as entries for payoff-dependent gamma distributions) is the likelihood,

$$\mathcal{L} = \mathcal{P}[x_1, \ldots, x_n | \Theta_1, w_1, \ldots, \Theta_k, w_k] = \prod_{i=1}^{n} (\sum_{j=1}^{k} (w_j \mathcal{P}[x_i | \Theta_j])).$$

In other words, it is the product of the convex combination of the probability of each dwell time according to each hidden behavioural state weighted by its mixing proportion. In the case of censored observations, the survivor function was used. In the case of uncensored observations, the density function was used. For example, the likelihood of a mixture of two components, with weights of 0.4 and 0.6, and means of 1s and 2s, for a single uncensored dwell time (x_1) of 1.6s is given by

$$\mathcal{P}[x = x_1 | \Theta_1, \Theta_2] = \mathcal{P}[x = 1.6 | \mu_1 = 1] \times 0.4 + \mathcal{P}[x = 1.6 | \mu_2 = 2] \times 0.6.$$

Similarly, the probability of a single censored dwell time (x_2) of 1.6s is given by

$$\mathcal{P}[x \geq x_2|\Theta_1, \Theta_2] = \mathcal{P}[x \geq 1.6|\mu_1 = 1] \times 0.4 + \mathcal{P}[x \geq 1.6|\mu_2 = 2] \times 0.6.$$

The probability of observing both x_1 and x_2 is the product of the above probabilities. The fitting algorithm identified the parametrization of Θs (which comprises all the distributional characteristics of each hidden behavioural state) and ws (which sets the preponderance of each hidden behavioural state in an activity) for which this likelihood $\mathcal{L}$ was maximal.

The number of parameters for any given model of an activity is two or five for every gamma-distributed component. In the case of components with payoff-dependent means ($n_{\gamma_{var}}$), there is one for the shape, one for the intercept of the logistic, one for its slope, one for its minimum asymptotic value and one for its maximum asymptotic value. In the case of components with payoff-independent means ($n_{\gamma_{const}}$), the slope is fixed at 0, the maximum and minimum are fixed at the mean, and only the intercept is free to vary. There is one parameter for every exponentially-distributed component ($n_{Exponential}$, the shape has a value fixed at one and one parameter determining the mean is free to vary). Finally, there is one parameter for all but one of the total number of distributions: k-1 of the components have a parameter determining their mixing proportion, and the mixing proportion of one of the components is not free to vary. This number of parameters, $n_{parm} = 2n_{\gamma_{const}} + 5n_{\gamma_{var}} + n_{Exponential} + (k-1)$, is what we used to evaluate the parsimony of our model. To ensure parsimony in how a behavioural activity was described, we considered only the mixture model which produced the lowest AIC value.

In summary, we identified sets of w and Θ values (one for each hidden state/mixture component) for which the likelihood function was maximal and for which the

AIC was minimal, where

$$\Theta_i = [\mu_i]$$

for payoff-invariant exponential components, where μ_i is the mean of distribution i,

$$\Theta_i = [\kappa_i, \mu_i]$$

for payoff-invariant gamma components, where κ_i is the shape parameter of distribution i,

$$\Theta_i = [\kappa_i, \beta_{0_i}, \beta_{1_i}, Min_i, Max_i]$$

for payoff-dependent gamma components, where β_{0_i} is the intercept, β_{1_i} is the slope, Min_i is the minimum and Max_i is the maximum predicted log-mean dwell time for component i at a particular payoff.

We identified the maximum-likelihood estimates of the parameters describing each activity iteratively in a manner similar to the expectation-maximization algorithm. In each iteration of a fit, assuming the most recently fit set of k constant (for payoff-independent exponentially- and gamma-distributed components) and logistic functions (for payoff-dependent gamma-distributed components), and k shape parameters (in the case of gamma distributions), we first obtained the maximum-likelihood estimates of the mixture proportion of the first $k-1$ components.

This fit was constrained such that the sum of the proportions could not be greater than 1 or less than 0. Once a maximum-likelihood fit of the mixing proportions was complete, the algorithm used the most recently fit values to identify the maximum-likelihood estimates of the logistic and constant functions setting the logarithm of the mean. This fit was constrained such that the expected dwell time for any hidden behavioural state was at most the longest observed dwell time for that activity and greater than 0. As the survivor function is a monotonically increasing function of the mean (for a score x, the survival probability of x never decreases as the mean

increases), the probability of a set of censored observations can be made arbitrarily large. For example, for a score $x = 0.1$s, the survivor function is 0.37 (i.e., 37 % of dwell times are at least 0.1s) when the mean is 0.1s, 0.90 when the mean is 1.0s, and 0.99 when the mean is 10.0s. As a result, the mean of any component at any payoff was constrained to be at most the longest dwell time observed for that activity across all payoffs, and at minimum, the shortest dwell time. Shape parameters of gamma distributions were constrained to be at least 1. A value of 1 indicates an exponential distribution. When a state is composed of exponentially-distributed sub-states, both of which must terminate in order for the state to terminate (for example, observing a "tapping release" may require "release paw from lever" and "resume lever-press" to be completed), dwell times in the state will have a gamma distribution with a shape parameter greater than 1. Shape parameters between 0 and 1 imply that a state is made up of a mixture of sub-states, with one component terminating more quickly than others. As uncovering the components of a mixture is the purpose of the algorithm described here, states were not allowed to have shape parameters that were less than 1.

To ensure that each component was uniquely specified, behavioural states were sorted from largest mixing proportion to smallest. This ensured that component 1 always referred to the most preponderant component of the mixture of distributions, component 2 the second, and so on. Following this second phase of the algorithm, the logarithm of the probability of all dwell times in the activity given the fit parameters (the log-likelihood) was calculated and compared to its value on the previous iteration. The algorithm continued to fit w, then Θ, values until the log-likelihood changed by less than 0.001 (the parameter values converged) or 500 iterations had been performed. All fits converged well before the 500 iteration limit was reached. Three starting values were used for this process: where payoff-dependent, logistic functions were set to have a slope of 0, -50 and 50 and an intercept that set the midpoint of the logistic function

at the midpoint of the logarithm of the payoffs tested. In all cases, the starting shape parameter of all components was set to 1 and the starting maxima and minima of the logistic functions, where applicable, were set to the maximum and minimum observed dwell times in the activity.

5.4.2.4 Inferring transition probabilities

Inference of the preponderance of each state (its mixing proportion w_i) provides the probability that a dwell time in an activity is sampled from the distribution that is characteristic of that state. Inference of the parameters that set each state's distributional characteristics (Θ_i), which could be

1. a mean μ_i for exponential,
2. a shape-mean combination $[\kappa_i, \mu_i]$ for gamma, or
3. a shape-scaled logistic combination $[\kappa_i, \beta_{0_i}, \beta_{1_i}, Min_i, Max_i]$ for payoff-dependent gamma distributions,

provides the probability per unit time that any state will terminate. The total expected duration of a sojourn in the activity is the weighted combination (convex because the weights sum to one) of the duration of each of its component states $\sum_i(w_i\mu_i)$. The termination rate of an activity is the reciprocal of its total expected duration. The transition rate from one activity to the next is thus almost completely specified by the set of w_is and Θ_is.

For most links in the chain, the probability of a transition from one observable activity to another, and from one hidden behavioural state to another, is trivial: it is either 0 or 1. For example, when a PRP has terminated, no matter which behavioural state may have actually occurred, it necessarily terminates on a hold with probability 1. When a tap or TLB terminates, no matter which behavioural state may have actually occurred, it must also necessarily terminate on a hold with probability 1. However, a hold could terminate on a long, censored release or an uncensored release,

and the uncensored release could be either a tap or a TLB. We therefore needed to explicitly model two probabilities: the probability that the rat would not return to work after releasing the lever, and if it did, the probability that an interruption to lever-pressing resulted in either a tap or TLB. We performed a logistic regression, using log-payoff as the sole predictor for a binomial outcome: either that a release would be censored by the end of the trial and last longer than 1s, in the case of the probability of a CTLB, or that an uncensored release was either a tap or a TLB, in the case of the probability of a tap. The former logistic regression provides the probability of transitioning from a hold to a CTLB, assuming a hold has terminated ($\mathcal{P}[Rel_{cens}]$), while the probability of transitioning from a hold to either a tap or TLB is obtained by subtracting this number from one ($\mathcal{P}[Rel_{unc}] = 1 - \mathcal{P}[Rel_{cens}]$). The second logistic regression provides the probability that an uncensored release is a tap ($\mathcal{P}[Rel_{unc} = tap]$) at each level of payoff; subtracting this number from 1 provides the probability that an uncensored release is a TLB ($\mathcal{P}[Rel_{unc} = TLB] = 1 - \mathcal{P}[Rel_{unc} = tap]$).

5.4.2.5 Inferring starting probabilities

Two cases require further explicit modelling. As stated, the CTMC assumes that following reward delivery, there is a PRP of non-zero duration that terminates on a hold. Although this is often the case, it is not a universal occurrence. For a select few rats with highly effective electrodes for which stimulation does not appear to induce any additional movement, the rat depresses the manipulandum as the lever re-enters the operant chamber. In this case, the first activity the rat performs in a reward encounter is a hold rather than a PRP. We address this reality by modelling zero-duration PRPs explicitly. Following reward delivery, there is some payoff-dependent probability that the CTMC will begin directly in a hold activity, thereby forcing the PRP to last zero seconds. We performed a logistic regression on the proportion of not

immediately-aborted reward encounters that were begun directly in the hold activity, as a function of the logarithm of the payoff.

A second special-case of the PRP concerns the so-called CTLB activity. It is possible that there are actually multiple hidden behavioural states related to censored true leisure bouts. For example, a rat may return to work after tens of minutes in anticipation of the next trial. In order to simplify the analysis, we considered only PRPs and releases that were censored by the end of the trial and lasted longer than 1s. We performed a logistic regression on the proportion of reward encounters that were immediately aborted (that is, the PRP was longer than 1s and censored by the end of the trial), as a function of the logarithm of the payoff.

5.4.2.6 Extrapolating to the entire trial

Once the rat has obtained an estimate of the payoff, this estimate sets the duration of the various hidden behavioural states that compose each activity. The activity will have an expected dwell time equal to the convex combination of the weight of each behavioural state and its mean. The reciprocal of this dwell time defines the rate at which the activity is left. From the above-mentioned hidden behavioural states, transition probabilities, and special-case transitions, a complete picture of performance on a reward encounter can emerge.

The rate $\alpha_{A,A'}$ at which an activity A transitions to another activity A' is the product of the reciprocal of its mean dwell time multiplied by the probability that A' follows A ($\alpha_{A,A'} = 1/\mu_A \cdot \mathcal{P}[A \rightarrow A']$). In many instances, the transition probability $\mathcal{P}[A \rightarrow A']$ is either 0 or 1. Uncensored PRPs, taps, and TLBs are always followed by holds. In the case of transitions from holds, the probability of a transition to a tap, TLB or censored true leisure bout may be somewhere in the interval [0,1]. Holds could terminate on censored (CTLB) or uncensored releases; if uncensored, a release could be a tap or a TLB. Both of these probabilities have been explicitly modelled above.

As a result, every potential transition rate is completely specified in the model.

We can define a series of differential equations for the evolution of performance in the CTMC in terms of these transition rates: at any time t, the rate at which an activity is left is the sum of all transition rates leading away from it, and the rate at which it is entered is the sum of all transition rates leading toward it. If α_{ij} is the transition rate from activity j to i, and α_{ii} is the (negative) transition rate away from activity i, then the probability at any time t that the rat engages in activity i is a differential matrix equation of the form

$$\frac{\partial}{\partial t}\mathcal{P}[activity = i \text{ at time } t] = \mathcal{A}\mathcal{P}[activity = i \text{ at time } t]$$

where $\mathcal{A}$ is a matrix whose entries are the elements α_{ij} and α_{ii}. For example, suppose the probability of being engaged in activity A is 1, and A' is 0, and further suppose that the rate $\alpha_{A,A'} = 0.25$. After 1s, the probability of being engaged in activity A' is 0.25, and the probability of still being engaged in activity A is 0.75. Another 1s later, the probability of transitioning from A to A' is still 0.25, so of the 0.75 probability that remains in performing activity A, one quarter of it ($0.75 \times 0.25 = 0.1875$) transitions from A to A'. Thus, there is now a 0.5625 probability ($0.75 - 0.1875$) of still being engaged in activity A, and a 0.4375 probability ($0.25 + 0.1875$) of being engaged in activity A'. In other words, the absolute change in the probability of being engaged in activity A (0.25 at 1s, and 0.1875 one second later) depends on the probability of already performing the activity (1.00 at 1s, and 0.75 one second later), which is an ordinary differential equation of the form $\partial/\partial t\mathcal{P}[A \text{ at } t] = \alpha_{A,A'}P[A \text{ at } t]$. Since there are multiple activities that could transition to each other, the CTMC defines a system of these ordinary differential equations that can be elegantly condensed into the matrix equation defined above, with the transition rates $\alpha_{A,A'}$ as entries to the matrix $\mathcal{A}$.

If we enumerate the activities as PRP (1), hold (2), tap (3), TLB (4) and CTLB (5), this matrix becomes:

$$\mathcal{A} = \begin{bmatrix} -\alpha_{PRP,H} & 0 & 0 & 0 & 0 \\ \alpha_{PRP,H} & -\alpha_{H,R} & \alpha_{tap,H} & \alpha_{TLB,H} & 0 \\ 0 & \alpha_{H,tap} & -\alpha_{tap,H} & 0 & 0 \\ 0 & \alpha_{H,TLB} & 0 & -\alpha_{TLB,H} & 0 \\ 0 & \alpha_{H,CTLB} & 0 & 0 & 0 \end{bmatrix}$$

and the probability of engaging in activity i at time t is a column vector with entries arranged in the same order (PRP, hold, tap, TLB and CTLB).

The general solution to differential equations of the form $\partial/\partial t X(t) = AX(t)$ is

$$X(t) = C_1 e^{\lambda_1 t} U_1 + \ldots + C_n e^{\lambda_n t} U_n$$

where C_k is a constant, and λ_k and U_k are the kth eigenvalue and eigenvector, respectively. In the case of engaging in some activity at any given point in time, this becomes

$$\mathcal{P}[activity = i \text{ at time } t] = C_1 e^{\lambda_1 t} U_1 + \ldots + C_n e^{\lambda_n t} U_n,$$

with C_k, λ_k and U_k the k^{th} constant, eigenvalue, and eigenvector.

The constant is solved according to the appropriate starting probabilities in each state. The probability of starting in the CTLB activity is $\mathcal{P}[PRP_{cens}]$. The probability of starting in either of the two non-CTLB activities (PRPs and holds) becomes $1 - \mathcal{P}[PRP_{cens}]$, which sets the probability of starting on a hold to $(1 - \mathcal{P}[PRP_{cens}]) \cdot \mathcal{P}[PRP_{unc} = 0]$, and of starting on an uncensored PRP to $(1 - \mathcal{P}[PRP_{cens}]) \cdot (1 - \mathcal{P}[PRP_{unc} = 0])$. All other starting probabilities are 0. At $t = 0$, the exponential terms of the probability equation disappear (as $e^0 = 1$), so the

starting probabilities define a simple linear system of form

$$\text{starting probability} = C_1U_1 + \ldots + C_nU_n = [U_1 \ldots U_n] \begin{bmatrix} C_1 \\ \ldots \\ C_n \end{bmatrix},$$

which can be solved for the vector of C_ks directly.

This general solution provides the probability, at any time t, that the rat is performing each of the 5 activities, when approximating the overall termination of an activity as a true time-invariant Markov process (i.e., all activities are exponentially distributed). This seemed a reasonable approximation, as the modal shape parameter for an activity was usually one, which is indicative of an exponentially-distributed process. The time spent engaging in each of the above-mentionned five activities is simply the integral of the probability of performing the activity in question over time. For example, if at time $t = 0.0s$, there is a 0.8 probability of performing a PRP activity, and a 0.2 probability of performing a hold activity, the rat has effectively spent 0.08s in a PRP and 0.02s in a hold between $t = 0$ and $t = 0.1s$. If, at time $t = 0.1s$, there is now a 0.7 probability of performing the PRP activity and a 0.3 probability of performing a hold, the rat has effectively spent 0.07s in the PRP and 0.03s in the hold activity between $t = 0.1$ and $t = 0.2s$. Summing these together, the rat has accumulated a total of $0.08 + 0.07 = 0.15s$ in the PRP and $0.05s$ in the hold activity since the start of the reward encounter. We performed a numerical integration, at 0.1s time steps (the resolution of our behavioural apparatus) until one of the following conditions was true:

1. The total time accumulated in hold activities reached the price, or

2. The total time accumulated in all activities together reached the reward encounter time predicted by a linear regression of log-reward encounter duration on log-objective price and log-subjective reward intensity.

The linear regression used above to predict the duration of the reward encounter used the log-subjective intensity, as a more valuable reward will drive the reward encounter to be shorter, as well as the log-objective price. This was done because even if the subjective opportunity cost of rewards will be equal when the price is 0.01s or 0.02s, the rat will have to actually hold the lever for 0.01s or 0.02s, and thus, the reward encounter will last, at minimum, twice as long in the 0.02s case than it does in the 0.01s case. At a constant objective price, increases in reward intensity should decrease the duration of the reward encounter, and at a constant subjective reward intensity, increases in objective price should increase the duration of the reward encounter.

This allowed us to extrapolate the results of our molecular model, which specifies behaviour on a moment-to-moment time scale, to the molar level, which describes behaviour on the scale of reward encounters and whole trials. The numerical integration of the system of differential equations provides the evolution of performance over arbitrary time scales, while the stopping criteria (either the objective price or the predicted reward encounter duration) provide a reasonable time scale over which to evaluate the integration. This modelling allows us to estimate the proportion of time allocated to holding and tapping, but only for the period of time the rat is actually working. In order to extrapolate time allocation to the entire trial, this time allocation was multiplied by the probability of not starting on a CTLB ($1 - \mathcal{P}[PRP_{cens}]$), because all reward encounters begun with a CTLB will necessarily have a time allocation of 0.

5.5 Results

5.5.1 Testing the assumptions of the model

5.5.1.1 PRP is independent of previous work bout

To test the Markov assumption with respect to the duration of the PRP, we performed a linear regression of the logarithm of the PRP duration as a function of the logarithm of the payoff and the logarithm of the last corrected work bout. This analysis did not include PRPs of 0 duration (for obvious reasons), and excluded instances where the last corrected work bout was equal to the objective price. This exclusion was performed to maximize the proportion of variance that can be uniquely attributed to each predictor: last corrected work bout duration or payoff, which is a scalar combination of the objective price with other key determinants of the decision to press.

Figure 5.3 is a box-whisker plot of the proportion of variance in log-PRP durations that can be accounted for by only the log-payoff (Log[Ub]) and by only the log-last corrected work bout (Log[CWB]) when the rat did not obtain a reward following a single, continuous hold. Overall, the log-last corrected work bout accounts for no variance (median of 0.00145) above and beyond what can be accounted for by the log-payoff, while the log-payoff uniquely accounts for considerable variance (median of 0.17696). The far-right box-whisker plot provides the ratio of the proportion of variance uniquely predicted by log-payoff compared to log-last corrected work bout. The median of these ratios is very large (161.6), reflecting the much greater degree to which our proposed determinant (log-payoff) can account for the decision to wait before pressing again than previous performance.

This linear regression provides an important confirmation that the duration of the PRP is relatively independent of the duration of the states that came before it,

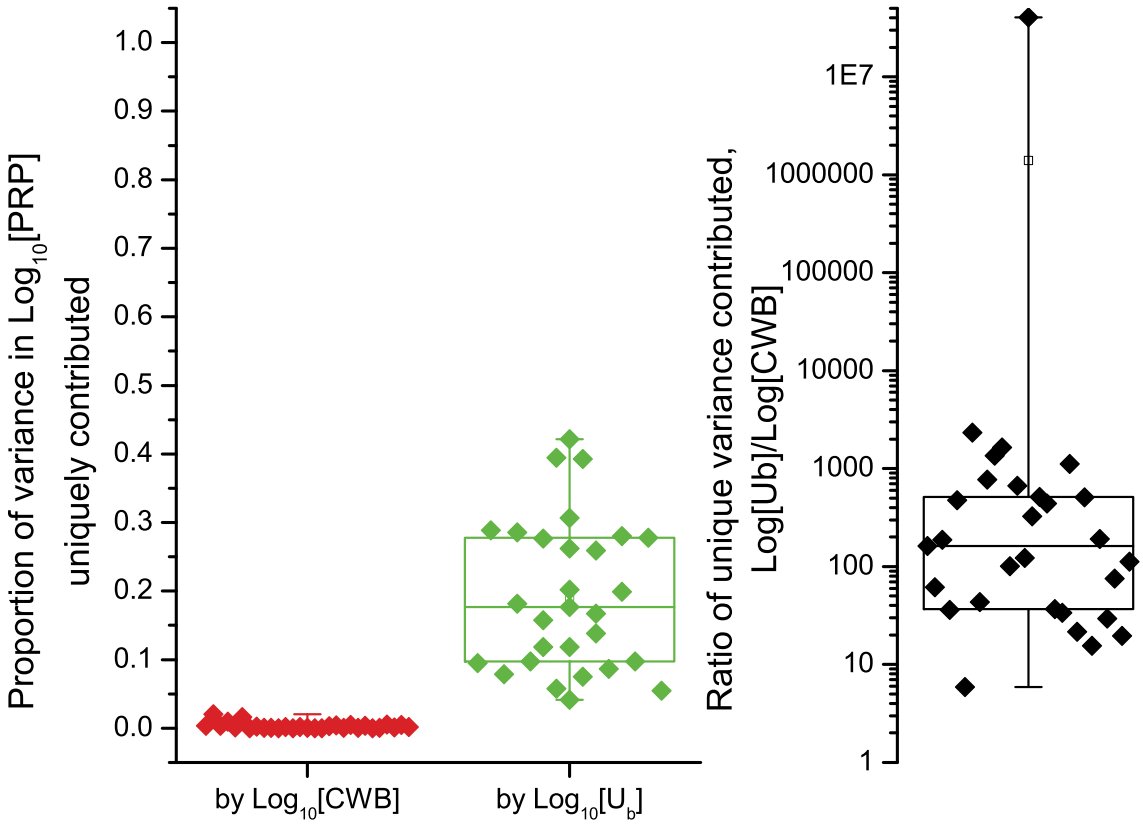

Figure 5.3. PRP durations are independent of previous work bout. Box whisker plot of the proportion of variance in Log_{10}[PRP duration] that can be uniquely attributed to the duration of the last corrected work bout (red) or to the payoff from self-stimulation (green) and not the other. In the right-hand panel, the ratio of the proportion of variance in Log_{10}[PRP duration] that can be uniquely attributed to payoff to the proportion of variance that can be uniquely attributed to the duration of the last corrected work bout is shown in a box-whisker plot. In all cases, post-reinforcement pauses do not bear any meaningful relationship with the duration of the last work bout, when controlling for payoff, while being much more heavily dependent on payoff while controlling for the duration of the last work bout.

thereby confirming (at least to a first-order approximation) that the Markov property holds in the case of the post-reinforcement pause. We now turn to the second type of pause that may not be time-invariant: the true leisure bout.

5.5.1.2 TLB duration is independent of previous work bout duration

To test the Markov assumption with respect to TLBs, we performed a simple linear regression of the log-TLB durations as a function of the log-corrected work bout that preceded them. For the Markov assumption to hold, TLB durations (and therefore, their logs) would need to be independent of history. As a first-order approximation, then, we set to determine whether there was any relationship between how long the rat had worked and the duration of the pause, lasting longer than 1s, that followed.

Figure 5.4 demonstrates this independence. The left-hand panel shows a scatter (for the animal showing the strongest dependence of TLB on the last corrected work bout) of the duration of the TLB as a function of the immediately preceding corrected work bout: there is little evidence a systematic relationship. The right-hand panel is a box-whisker plot of the proportion of variance in log-TLB that can be accounted for by the log-last corrected work bout. It has a median value of 0.011, and in no case is the relationship statistically significant at the traditional 0.05 level. We can conclude from these data that, indeed, the Markov assumption holds (to a first-order approximation) in the case of TLBs: they are relatively independent of the corrected work bouts that preceded them, no matter what the payoff may be. Regardless of whether the rat spent a great deal of time working or a short amount of time working, the rat will spend no more or less time engaging in uncensored releases lasting over 1s.

Having justified the Markov assumption for PRPs and TLBs, we shall now describe the results of modelling real-time performance in the period of time following

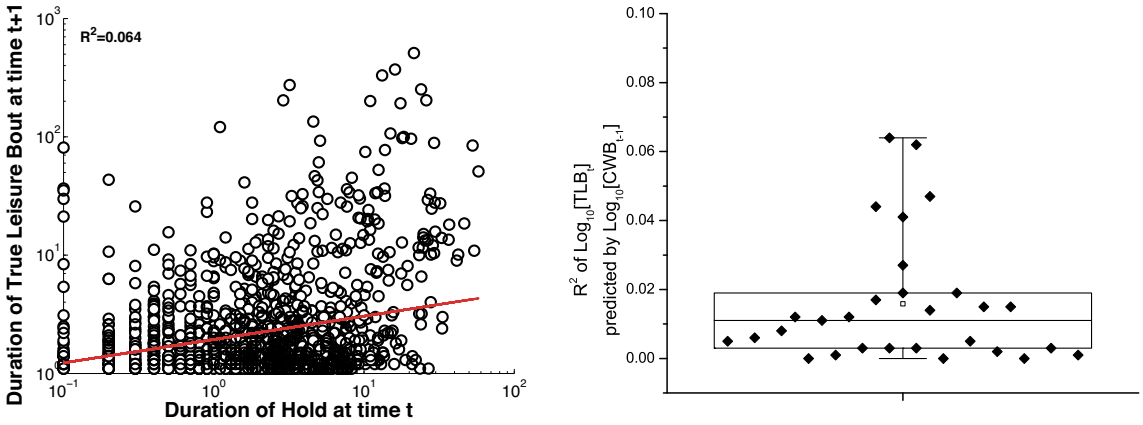

Figure 5.4. TLB duration is independent of previous work bout. The left-hand panel provides a scatter plot of the data from F17, the animal showing the strongest dependence of log-TLB duration on the log-last corrected work bout duration. The regression line is indicated in red. The right hand panel provides a box-whisker plot of the R^2 values from all animals in all conditions; the median value occurs at approximately 0.01 (1% of the variance in log-TLB duration can be explained by the duration of the last corrected work bout) and in no case is the regression statistically significant. Points indicate individual R^2 values.

the first reward delivery as a CTMC, painting a portrait of each activity.

5.5.2 Modelling performance

5.5.2.1 Payoff-dependent and independent activities

Figure 5.5 provides our justification for allowing PRPs and TLBs to have hidden behavioural states with payoff-dependent means, while restricting the hidden behavioural states of holds and taps to payoff-independent means. Although the maximum-likelihood estimates of the PRP (top left) and TLB (bottom right) duration consistently decrease with payoff across rats (R^2 values of 0.48 and 0.38, respectively), the maximum-likelihood estimates of the tap (top right) and hold (bottom left) durations have much weaker relationships with payoff (R^2 values of 0.07 and 0.11, respectively). All correlations were significant, but at a 5% type-I error rate, with the number of data points observed (at minimum 1260), an R^2 value of 0.003 would be statistically significant. As a result of their weaker payoff relationships, we judged that allowing holds and taps to consist of payoff-dependent components was not justified by the overall pattern of dwell times in these activities across all animals.

5.5.2.2 Post-reinforcement pause

Figures 5.6 and 5.7 provide a portrait of the PRP. Figure 5.6 shows the log-survival function of the dwell times at each payoff for one animal (DE15) at each payoff, with dwell times on a log scale. Hot colours indicate high payoffs and cool colours indicate low payoffs. In linear space, the log-survival function is a straight line when the underlying process is a single exponential distribution; it is convex for mixtures of multiple distributions and concave for gamma distributions. Plotting the log-survival (z-axis) as a function of the log-dwell time (y-axis), as has been done here, exaggerates the convex/concave relationship for single gamma- and mixtures of

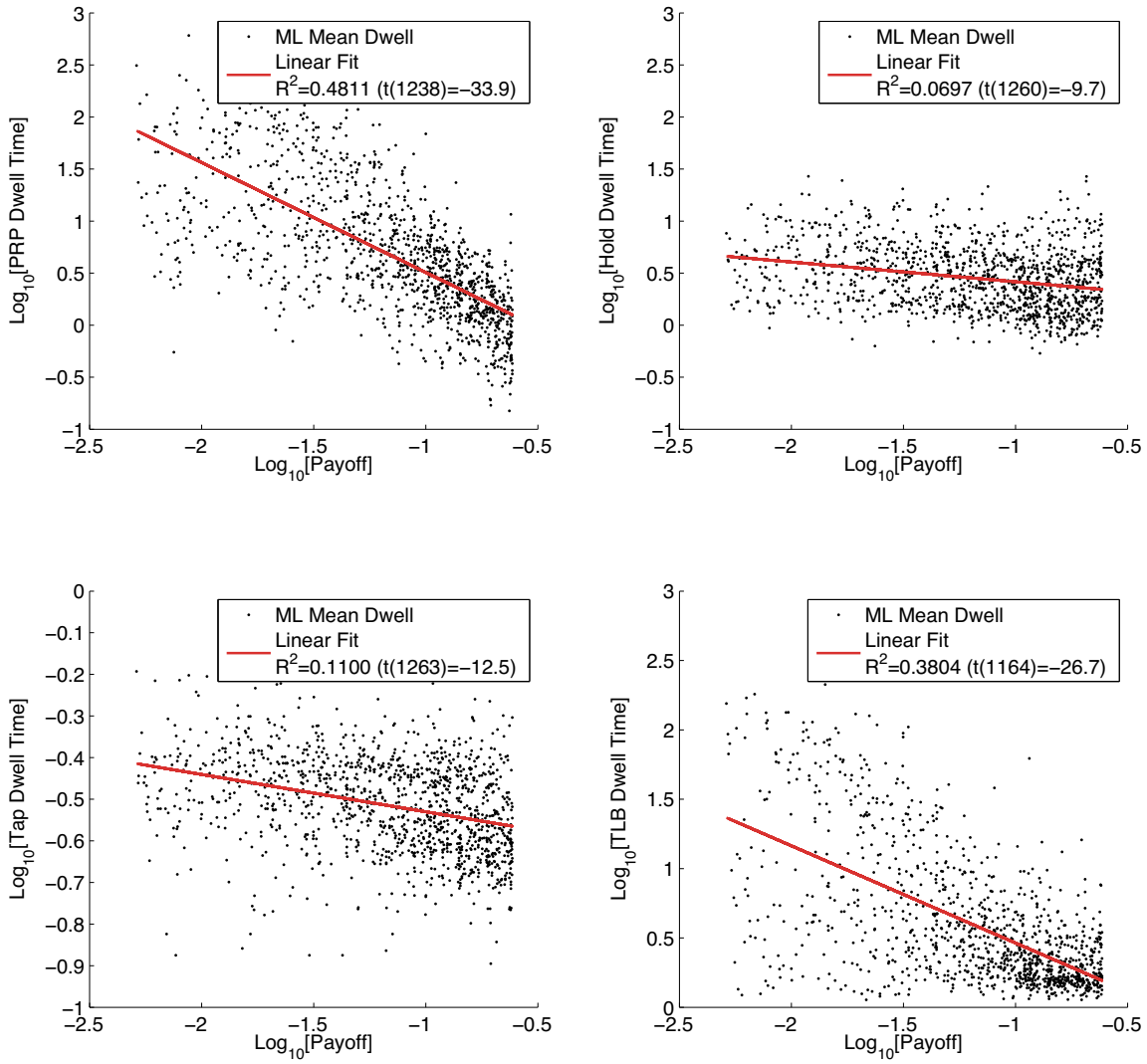

Figure 5.5. Maximum likelihood dwell times as a function of payoff. Scatter plots of the log-maximum likelihood dwell times in PRP (upper left), hold (upper right), tap (lower left) and TLB (lower right) activities for all animals in the range of payoffs common to all animals, as a function of the log-payoff. While the relationship between log-dwell time and log-payoff is strong for both PRP and TLB activities (R^2 values are 0.48 and 0.38, respectively), the relationship is weak for both holds and taps (R^2 of 0.07 and 0.11, respectively).

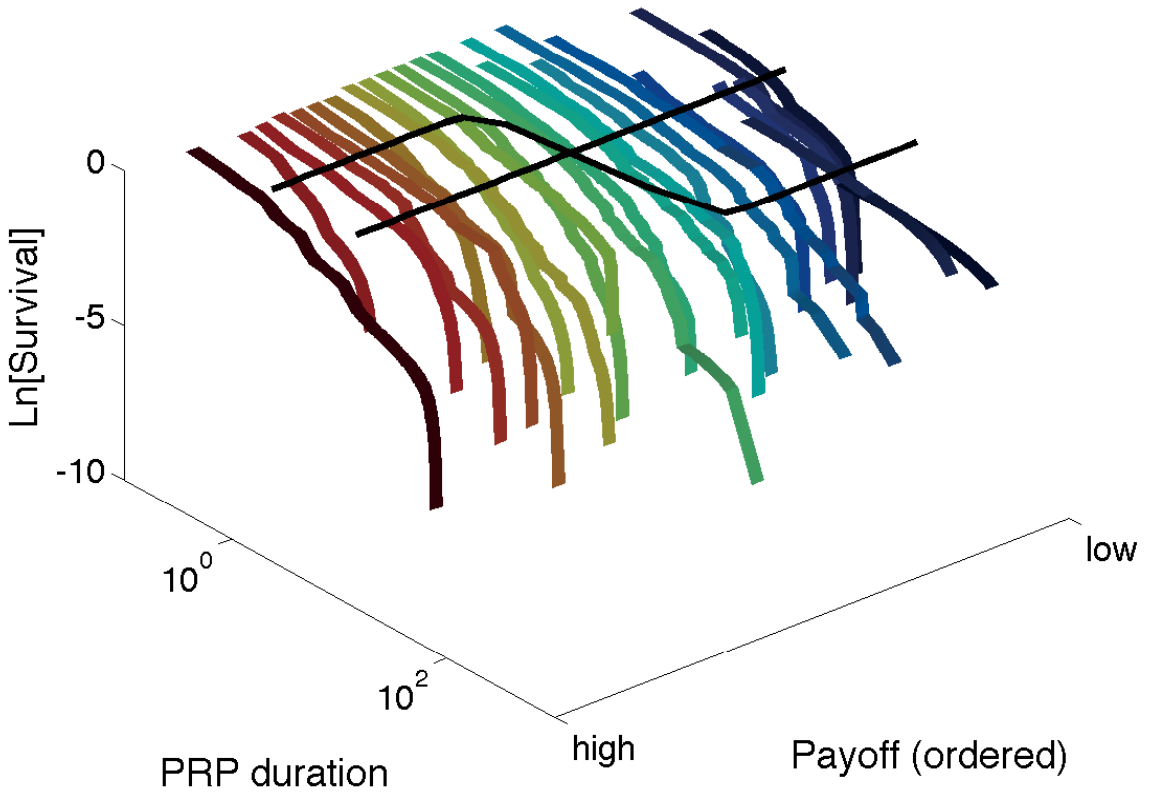

Figure 5.6. Log-survival function of the PRP dwell time. The log-survival function is shown here as a function of PRP dwell time (in logarithmic space) and payoff (in ordinal space) overlain with the predicted dwell times of the hidden states. Hot colours indicate high payoffs while cold colours indicate low payoffs. At high payoffs, the PRP rarely survives beyond a few seconds, while at low payoffs, PRP dwell times on the order of hundreds of seconds have a higher probability of survival.

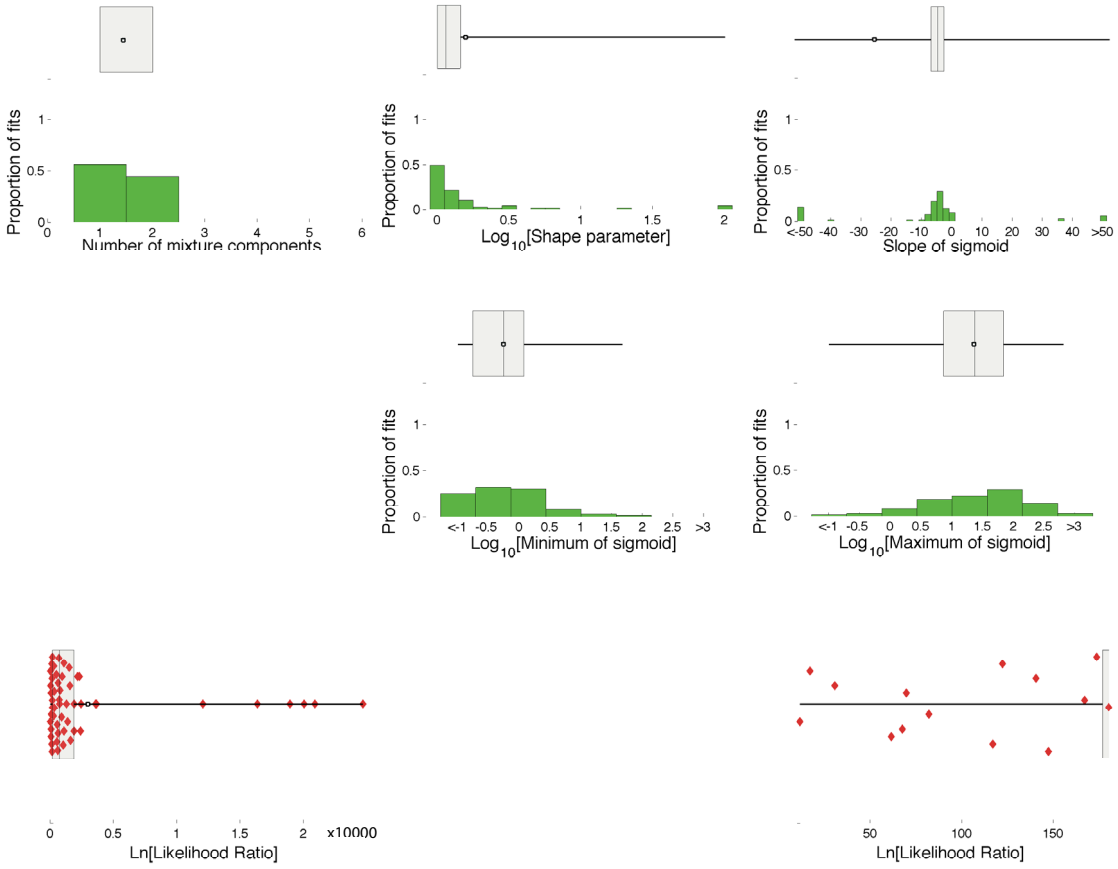

Figure 5.7. Portrait of the post-reinforcement pause activity. Top panels show histograms of the number of hidden states (upper left), the log-shape parameter (upper centre), the slope of the sigmoid (upper right), the log-minimum dwell time (middle centre), and log-maximum dwell time (middle right). The bottom two panels show a box-whisker plot of the Ln-likelihood ratio of the CTMC model to a model in which PRP durations are stochastic realizations of a single, payoff-independent exponential process. The bottom right panel focuses on the region from 10 to 180. Red points indicate individual Ln[Likelihood Ratio] values. Note that the scale on the bottom left plot is in units of ten thousand.

multiple gamma distributions; exponential distributions become convex. Gamma and exponential (that is, gamma distributions with a shape parameter of 1) distributions thus appear to have more concave (downward-accelerating) log-survival functions, and mixtures have more convex (downward-decelerating) functions.

Overlain with the fit, we have indicated the hidden behavioural states in black. In this animal, two hidden PRP behavioural states were extracted: one with a negative sloping payoff dependence and one payoff-independent component.

The plots on the first row of figure 5.7 show the number of hidden states (left), their shape parameter (middle), and the slope of the scaled logistic function that sets their mean (right), for each animal in each condition. In each case, we provide a histogram and box-whisker plot of the extracted value. The post-reinforcement pause appears to consist of one or two states (median number is 1, inter-quartile range spans 1 to 2) that is largely exponentially distributed (median κ is 1.14, or a logarithm of 0.06; inter-quartile range spans 1 to 1.44, or logarithms from 0 to 0.16). This hidden state has a payoff-dependent mean, and the slope of the payoff-mean relationship is negative (median slope is -4.77, inter-quartile range spans -6.9 to -2.7).

The middle row shows the minimum of the scaled logistic (middle) and the maximum of the scaled logistic (right), for each animal in each condition. We provide, as above, a histogram and box-whisker plot of the extracted values. The post-reinforcement pause appears to have a very low minimum (often as low as 0.1s), with a median log of -0.26 (0.55s) and inter-quartile range spanning -0.86 (0.17s) to 0.07 (0.42s). At low payoffs, the post-reinforcement pause rises to a maximum of 1.37 (23.20s), with an inter-quartile range spanning 0.86 (7.28s) to 1.83 (68.08s).

The bottom panel of figure 5.7 shows the log-likelihood ratio of our model, in which dwell times are sampled from a mixture of exponential and payoff-dependent gamma distributions, compared to a null model in which dwell times are sampled from a single, payoff-independent exponential distribution. As the range of log-likelihood

ratios extends to 2.5×10^4, we have expanded the plot in the right-hand panel, to focus on the region from logarithms of 10 to 180. In all cases, the log-likelihood ratio is large, with a minimum log of approximately 11, indicating that in the worst case, the probability of our data is approximately 60000 times (e^{11}) more likely if we assume our model is correct than if we assume a null model is correct, even when taking the extra parameters into account.

5.5.2.3 Hold

Figures 5.8 and 5.9 provide a portrait of the hold activity. Figure 5.8 shows a typical log-survival plot of the dwell times at each payoff for one animal (F9). The payoff-independent behavioural states are indicated with a black line overlay. In this animal, two hidden holding-related states were extracted: one with a low (log of 0.2, or 1.6s) mean and one with a high (log of 1.1, or 12.6s) mean.

The top plots of figure 5.9 show the number of hidden states (left), their shape parameter (middle), and their mean (right), for each animal in each condition. In each case, we provide a histogram and box-whisker plot of the extracted value. Holds appear to consist of two or three (median number is 2, inter-quartile range spans 2 to 3), mostly exponentially distributed (median κ is 1.25, inter-quartile range spans 1 to 1.91) hidden states. The log-dwell time in holding-related states is broadly distributed, with a median at a log of 0.34 (or 2.19s) and an inter-quartile range spanning a log of -0.14 (or 0.7s) to a log of 0.84 (or 6.92s).

Because multiple hidden behavioural states related to holding were identified, and these states were forced to have payoff-independent mean dwell times, we further identified the mean dwell time for the two most preponderant components (those with the greatest w_i). The two most preponderant components were ordered by their mean; if only one component was identified, it was assumed to be the longest. The medians and inter-quartile ranges of the mean dwell times for each of these two components

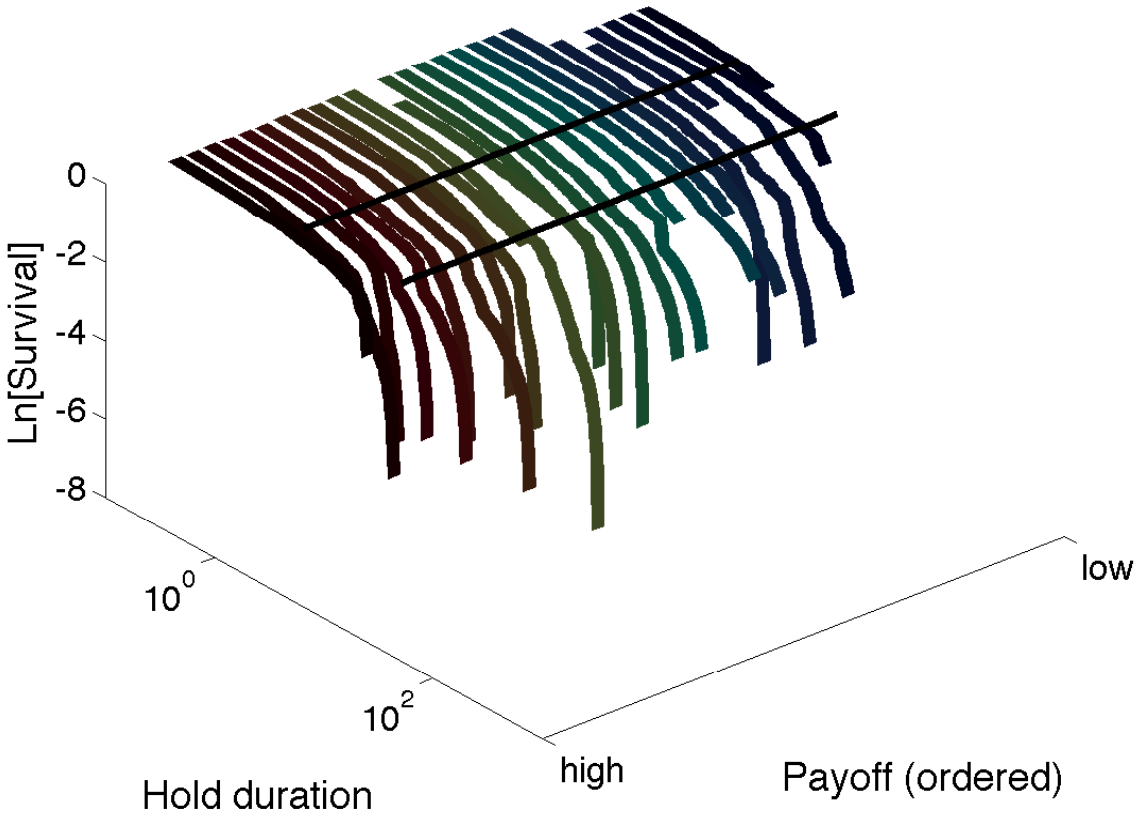

Figure 5.8. Log-survival function of the hold dwell time. The log-survival function is shown here as a function of hold dwell time (in logarithmic space) and payoff (in ordinal space) overlain with the predicted dwell times of the hidden states. Hot colours indicate high payoffs while cold colours indicate low payoffs. Two components are apparent: one with a mean that is a few seconds long, and a second with a mean that is on the order of tens of seconds.

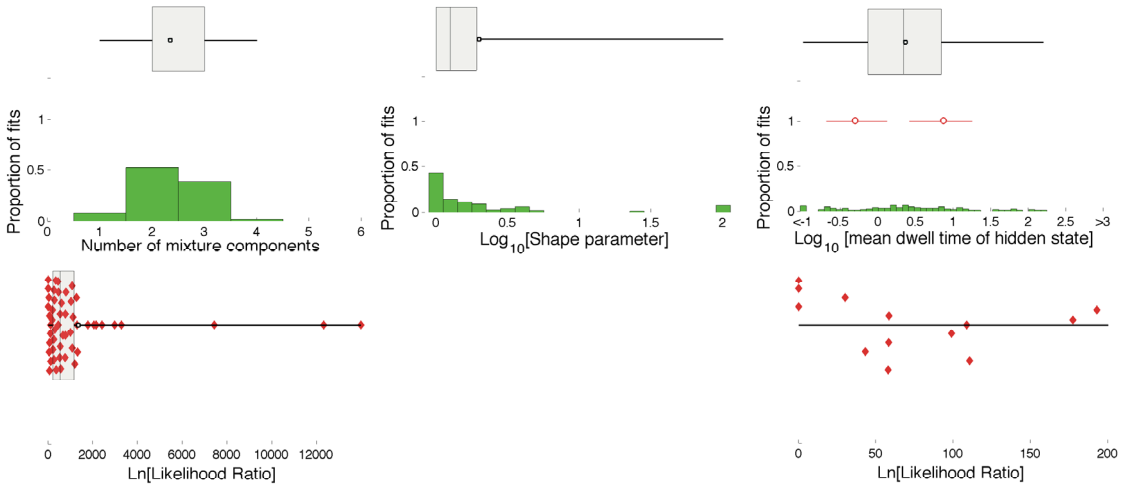

Figure 5.9. Portrait of the hold activity. Top panels show histograms of the number of hidden states (left), the log-shape parameter (centre), and predicted mean dwell time (right). The middle two panels show a box-whisker plot of the Ln-likelihood ratio of the CTMC model to a model in which hold durations are stochastic realizations of a single exponential process. The bottom right panel focuses on the region from 0 to 180. Red points indicate individual Ln[Likelihood Ratio] values.

are indicated in red in the top right panel of figure 5.9. The dwell time of the shorter component had a median log of −0.31 (0.49s) with an inter-quartile range spanning a log of −0.69 (0.20s) to 0.12 (1.31s). The dwell time of the longer component had a median log of 0.87 (7.35s) with an inter-quartile range spanning a log of 0.412 (2.58s) to 1.25 (17.7s).

The bottom panel of figure 5.9 shows the log-likelihood ratio of our model, in which dwell times are sampled from a mixture of exponential and payoff-independent gamma distributions, compared to a null model in which dwell times are sampled from a single, payoff-independent exponential distribution. As above, because of the large range spanned by the log-likelihood ratios (from 0 to 13000), the right-hand panel focuses on the region from 0 to 180. While 3 of 52 fits (6%) have a log-likelihood ratio of 0, indicating that our two-component model is no better than a null (single exponential component) model, the next smallest log-likelihood ratio is 30. In 94% of cases tested, the probability of the data according to our model is at least $e^{30} \approx 10^{13}$ times better than a model in which all hold times are drawn from a single exponential distribution.

5.5.2.4 Tap

Figures 5.10 and 5.11 provide a portrait of the tap activity. The upper left-hand panel shows a typical log-survival function of the dwell times at each payoff for one animal (F9). Overlain with the fit, we have indicated the hidden behavioural states in black. In this animal, two hidden tapping release-related behavioural states were extracted: one with a very short mean (log of −1, or 0.1s) and one with a longer mean (log of −0.45, or 0.35s).

The top plots of figure 5.11 show the number of hidden states (left), their shape parameter (middle), and their mean (right), for each animal in each condition. In each case, we provide a histogram and box-whisker plot of the extracted value.

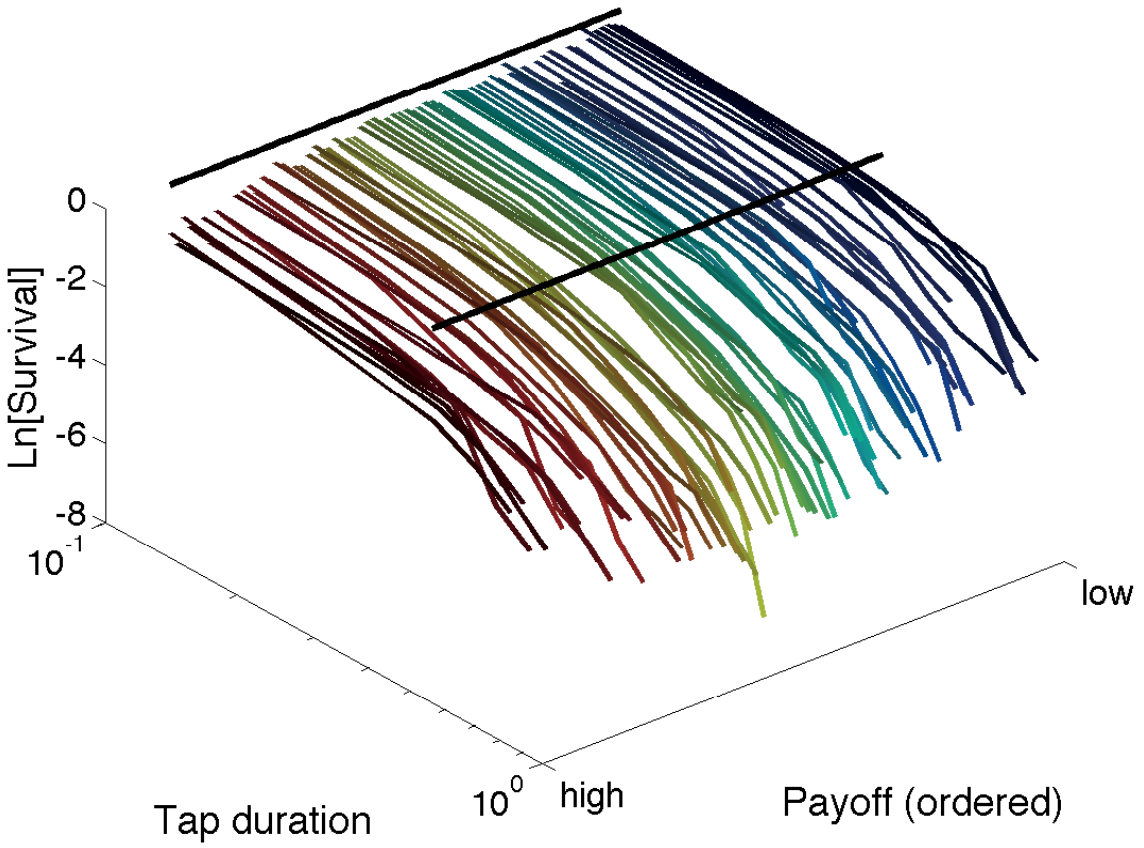

Figure 5.10. Log-survival function of the tap dwell time. The log-survival function is shown here as a function of tap dwell time (in logarithmic space) and payoff (in ordinal space) overlain with the predicted dwell times of the hidden states. Hot colours indicate high payoffs while cold colours indicate low payoffs. Two components are apparent: one with a mean that is 0.1 seconds long, and a second with a mean that is on the order of 0.4 seconds.

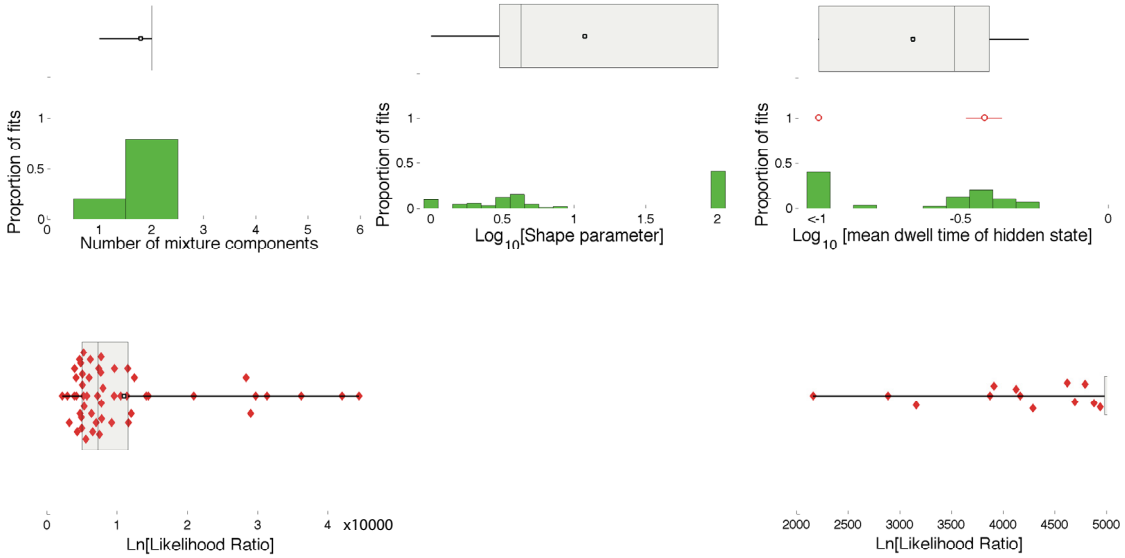

Figure 5.11. Portrait of the tap activity. Top panels show histograms of the number of hidden states (left), the log-shape parameter (centre), and predicted mean dwell time (right). The bottom two panels show a box-whisker plot of the Ln-likelihood ratio of the CTMC model to a model in which hold durations are stochastic realizations of a single exponential process. The bottom right panel focuses on the region from 2000 to 5000. Red points indicate individual Ln[Likelihood Ratio] values. Note that the scale on the bottom left is in units of ten thousand.

The tap appears to consist of two (median number is 2, inter-quartile range spans 2 to 2), largely gamma-distributed (median κ is 4.23, inter-quartile range spans 2.98 to 99.99) hidden states. The log-dwell time in tapping-related release states has a median at -0.53 (or 0.30s) and an inter-quartile range spanning from -1 (or 0.1s) to -0.44 (or 0.36). Additionally, this distribution appears to be highly bimodal, with one mode near the median, at approximately 0.3s (log of -0.53), and another mode at the shortest tapping-related release that can be detected, 0.1s (log of -1).

Because multiple hidden states related to tapping were identified, and these states were forced to have payoff-independent mean dwell times, we further identified the mean dwell times of the two most preponderant components (those with the greatest w_i). The two most preponderant components were ordered from shortest to longest mean; if only one component was identified, it was assumed to be the longest. The medians and inter-quartile ranges of the mean dwell times for each of these two components are indicated in red in the middle right panel of figure 5.11. The dwell time of the shorter component had a median log of -1.00 (0.1s) with an inter-quartile range spanning a log of -1.00 (0.1s) to -1.00 (0.1s). The dwell time of the longer component had a median log of -0.43 (0.37s) with an inter-quartile range spanning a log of -0.49 (0.32s) to -0.37 (0.39s).

The bottom panel of figure 5.11 shows the log-likelihood ratio of our model, in which dwell times are sampled from a mixture of exponential and payoff-independent gamma distributions, compared to a null model in which dwell times are sampled from a single, payoff-independent exponential distribution. Similarly to figures 5.7 and 5.9, the right-hand panel focuses on the region from 2000 to 5000. In all cases, the log-likelihood ratio is large, with a minimum of approximately 2158, indicating that in the worst case, the data is approximately $e^{2158} > 2^{64} - 1$ (greater than can be represented by a 64-bit unsigned integer) times more likely if we assume our two-component model is correct than if we assume a null (single exponential component)

model is correct.

5.5.2.5 True Leisure Bout

Figures 5.12 and 5.13 provide a portrait of the TLB. Figure 5.12 shows a typical log-survival function of the dwell times at each payoff for one animal (DE15). Overlain with the fit, we have indicated the hidden behavioural states in black. In this animal, two hidden TLB behavioural states were extracted, both with negative sloping payoff dependencies.

The plots on the top row of figure 5.13 show the number of hidden states (left), their shape parameter (middle), and the slope of the scaled logistic function that sets their mean (right), for each animal in each condition. In each case, we provide a histogram and box-whisker plot of the extracted value. The post-reinforcement pause appears to consist of two states (median number is 2, inter-quartile range spans 2 to 2) that is largely gamma-distributed (median κ is 5.11, inter-quartile range spans 1 to 13). These hidden states have a payoff-dependent mean, and the slope of the payoff-mean relationship is negative (median slope is -8.1, inter-quartile range spans -153.4 to 0)

The middle row shows the minimum of the scaled logistic (middle) and the maximum of the scaled logistic (right), for each animal in each condition. We provide, as above, a histogram and box-whisker plot of the extracted values. True leisure bout activities appear to have a very low minimum (often as low as 1s), with a median log of 0.23 (1.68s) and an inter-quartile range spanning 0.12 (1.31s) to 0.47 (2.92s). When payoff is low, the dwell time in TLB-related behavioural states rises to a median log of 1.38 (23.82s), with inter-quartile range spanning a log of 0.52 (3.34s) to 1.92 (83.18s). However, the maximum dwell times appear to be almost uniformly distributed: they are widely distributed from a log of 0, corresponding to the 1s tapping criterion (which is the *lowest* possible dwell time in TLB activities), to a log of 2, corresponding to a

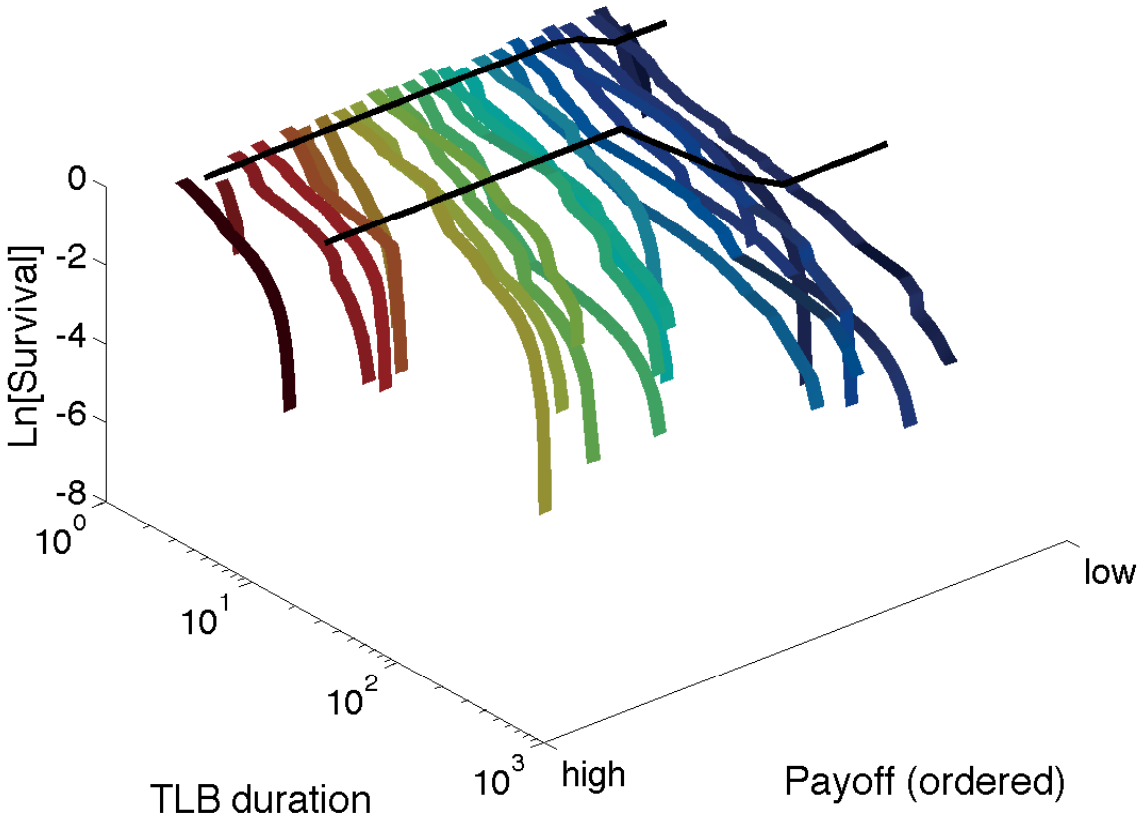

Figure 5.12. Log-survival function of the TLB dwell time. The log-survival function is shown here as a function of TLB dwell time (in logarithmic space) and payoff (in ordinal space) overlain with the predicted dwell times of the hidden states. Hot colours indicate high payoffs while cold colours indicate low payoffs. At high payoffs, the probability that a TLB will survive beyond 10s is very low, while at low payoffs, the probability that a TLB will be at least tens of minutes is very high. One component of this two-component mixture is close to 1s and relatively payoff-independent, reflecting a possible artifact of the arbitrary 1s criterion. The other component of this mixture is highly payoff-dependent.

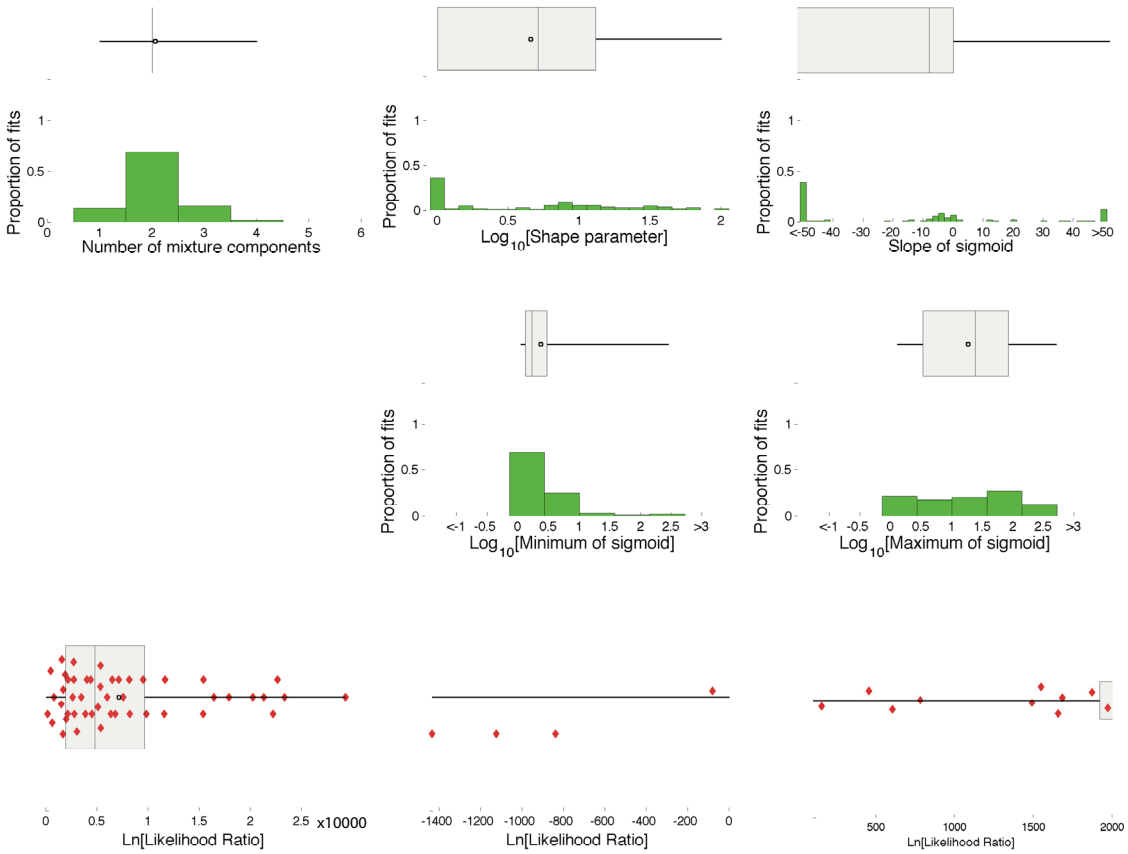

Figure 5.13. Portrait of the true leisure bout activity. Top panels show histograms of the number of hidden states (upper left), the log-shape parameter (upper centre), the slope of the sigmoid (upper right), the log-minimum dwell time (middle centre), and log-maximum dwell time (middle right). The bottom two panels show a box-whisker plot of the Ln-likelihood ratio of the CTMC model to a model in which TLB durations are stochastic realizations of a single exponential process. The bottom centre panel focuses on the region from -1400 to 0, while the bottom right panel focuses on the region from 100 to 2000. Red points indicate individual Ln[Likelihood Ratio] values. Note that the scale on the bottom left is in units of ten thousand.

100s true leisure bout. We shall return to this point later.

The bottom panel of figure 5.13 shows the log-likelihood ratio of our model, in which dwell times are sampled from a mixture of exponential and payoff-dependent gamma distributions, compared to a null model in which dwell times are sampled from a single, payoff-independent exponential distribution. In all but 4 (7%) cases, the log-likelihood ratio is large, with a minimum of 154.7, indicating that in 93% of cases, the probability of the data assuming our two-component, payoff-dependent model is true was at least $e^{154.7} \approx 1.53 \times 10^{67}$ times greater than the probability of the data assuming a null (single payoff-independent component) model. In the remaining 7% of cases, the probability of the data assuming a null model was greater than when assuming the mixture model we have proposed.

5.5.2.6 Starting probabilities

The top left panel of figure 5.14 shows, for a representative rat (DE1), the relationship between the logarithm of the payoff and the probability that a reward encounter begins with a CTLB. The prediction of the logistic regression is depicted along with the observed probabilities and their associated 95% confidence intervals; the proportion of variance is indicated in the legend. The top right panel of figure 5.14 shows a box-whisker plot of the overall proportion of variance in CTLB probability accounted for by payoff across all rats and conditions. The median R^2 is 0.15, with an inter-quartile range spanning 0.09 to 0.21.

The bottom left panel of figure 5.14 shows, for a representative rat (F3), the relationship between the logarithm of the payoff and the probability that a reward encounter begins with a hold. The prediction of the logistic regression is depicted along with the observed probabilities and their associated 95% confidence intervals; the proportion of variance accounted for by payoff is indicated in the legend. The bottom right panel shows a box-whisker plot of this proportion across all rats and

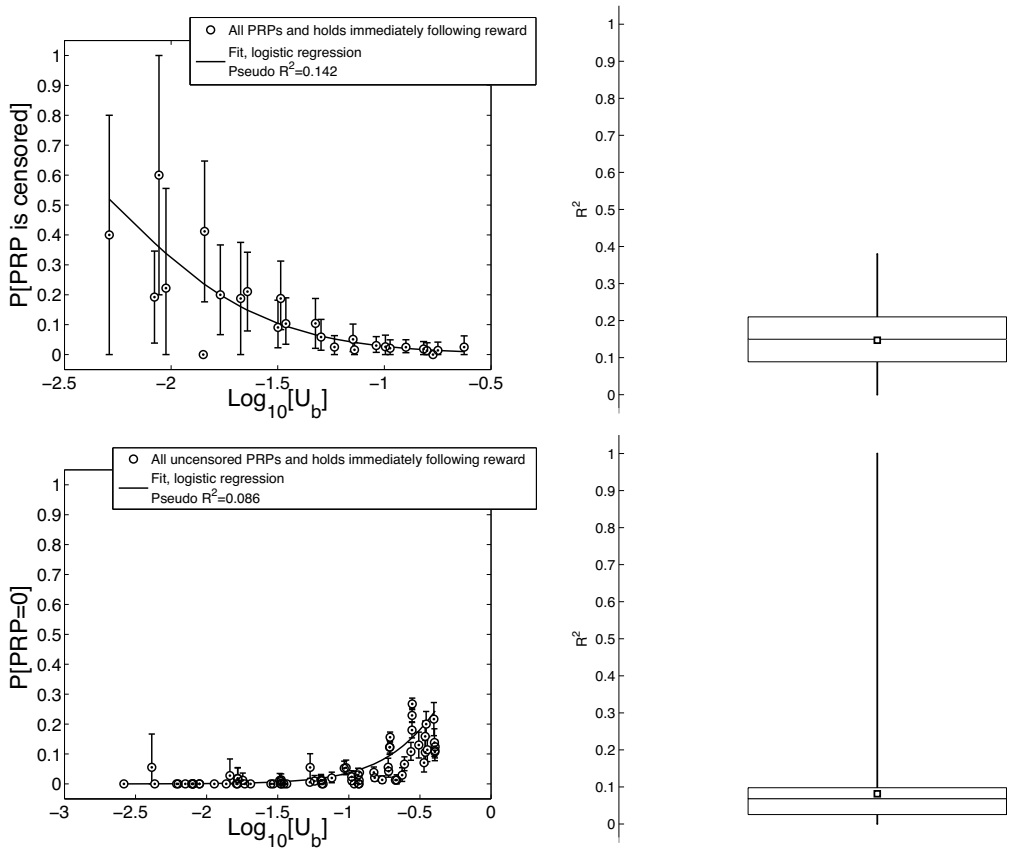

Figure 5.14. Probability that the PRP is censored, and if not, that it lasts 0s. The upper left panel provides an example relationship of the probability of a CTLB when the reward encounter begins as a function of payoff for a typical animal (DE1). Circles indicate proportions of times the rat began a CTLB at the start of the reward encounter (± 95% confidence interval), solid lines indicate the logistic regression line. The upper right panel is a box-whisker plot of the R^2 values of this relationship across all animals and conditions. The bottom left panel provides an example relationship of the probability of starting a reward encounter immediately in the hold activity (that is, the duration of the PRP is 0) as a function of payoff for a typical animal (F3). Circles indicate proportions of times a reward encounter began on the hold activity (± 95% confidence interval), solid lines indicate the logistic regression line. The bottom right panel is a box-whisker plot of the R^2 values of this relationship across all animals and conditions.

conditions. The median R^2 is 0.07, with an inter-quartile range spanning 0.03 to 0.10.

In both cases, there is considerable variability in the proportion of variance that can be accounted for by these logistic regressions. Some rats very rarely begin reward encounters in a CTLB. As a result, the R^2 values for the logistic regressions of these animals is close to 0. Similarly, because of interfering motor effects, some rats rarely begin reward encounters by immediately holding the lever down, while some rats with highly effective electrodes are capable of depressing the lever as soon as it begins to extend into the chamber.

5.5.2.7 Transitions from holds to releases

The top left panel of figure 5.15 shows, for a representative rat (DE15), the relationship between the logarithm of the payoff and the probability that an uncensored hold terminates on a CTLB. The curve of the logistic regression is depicted along with the observed probabilities and their associates 95% confidence intervals; the proportion of variance is indicated in the legend. The top right panel shows a box-whisker plot of the overall proportion of variance in CTLB probability accounted for by payoff across all rats and conditions. Overall, rats very rarely stop responding entirely part-way through a trial, and the probability of doing so is largely payoff-independent. The median R^2 is 0.01, with an inter-quartile range spanning 0.002 to 0.03. These estimates are much lower than the results seen above, implying that if the rat is going to wait until the next trial begins before the rat returns to work, it will do so as soon as the reward encounter begins.

The bottom left panel of figure 5.15 shows, for a representative rat (DE20), the relationship between the logarithm of the payoff and the probability that an uncensored hold terminates on an uncensored release lasting one second or less. The curve of the logistic regression is depicted along with the observed probabilities and their associated 95% confidence intervals; the proportion of variance accounted for

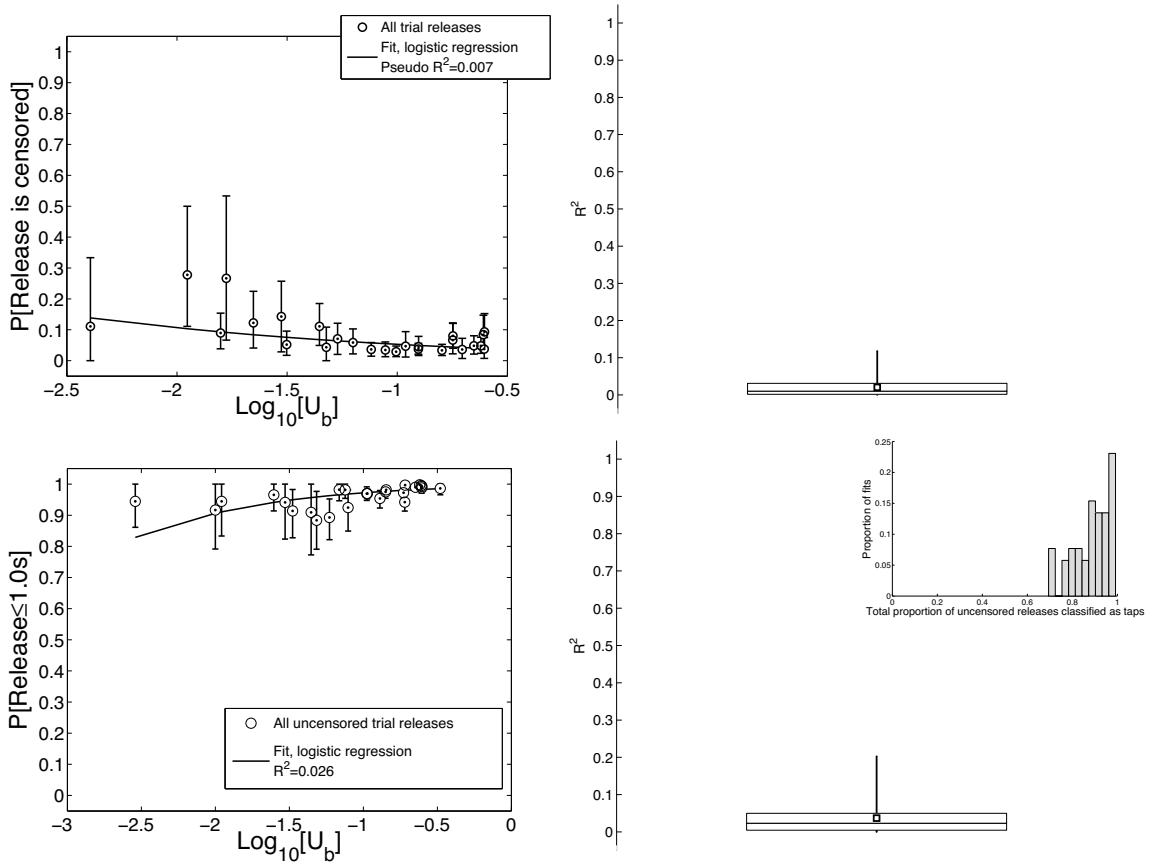

Figure 5.15. Probability that a hold terminates on a CTLB, and if not, that it terminates on a tap. The upper left panel provides an example of the relationship between the probability that a hold that is terminated before reward delivery will be followed by a CTLB for a typical animal (DE15). Circles indicate proportions of times an uncensored hold was followed by a CTLB (± 95 % confidence interval), solid lines indicate the logistic regression line. The upper right panel is a box-whisker plot of the R^2 values of this relationship across all animals and conditions. The bottom left panel provides an example of the relationship between the probability that a hold that is not interrupted by a CTLB will be a tap for one representative animal (DE20). Circles indicate proportions of times an uncensored hold that wasn't followed by a CTLB was interrupted by a tap (± 95 % confidence interval), solid lines indicate the logistic regression line. The bottom right panel is a box-whisker plot of the R^2 values of this relationship across all animals and conditions. In the inset, we have provided a histogram of the proportion of uncensored releases that have been classified as taps.

by payoff is indicated in the legend. The bottom right panel shows a box-whisker plot of this proportion across all rats and conditions. The median R^2 is 0.02, with an inter-quartile range spanning 0.0045 to 0.05. The inset to the bottom right panel shows why: the probability that a release will be classified as a tap is above 70% in all fits, and in over half of the cases, the probability that a release will be classified as a tap is above 90%. While the probabilities of a long, censored PRP, and a short, zero-duration PRP are payoff-dependent, these results imply that the probabilities of a tapping-related release, and long, censored leisure bouts are both payoff-independent.

5.5.3 Molar predictions of the molecular model

Figure 5.16 depicts the first step in extrapolating from the molecular-level performance described by the CTMC to molar, whole-trial performance. The left-hand panel shows the linear regression of log-reward encounter duration onto log-price and log-subjective intensity for a representative rat (DE15). Objective price was used because at a constant reward intensity, no matter how low the objective price, its value will affect the duration of the reward encounter: a trial for which the objective price is 0.02s will take longer to lead to reward than one for which the objective price is 0.01s, even though the rat may treat those two as subjectively equally costly.

As expected, there is a strong relationship between price, intensity, and reward-encounter duration. As a result, we used this estimate of how long the rat took to complete a reward encounter that was not censored by the end of the trial.

This estimate provides the appropriate time frame for which to evaluate the evolution of the CTMC as described. Each activity i in the CTMC will transition to another activity j at a rate of α_{ji}, and the rate at which it is left will be α_{ii}. For example, if holds transition to taps, TLBs, and CTLBs at rates

$$\alpha_{tap,hold}, \alpha_{TLB,hold}, \alpha_{CTLB,hold}$$

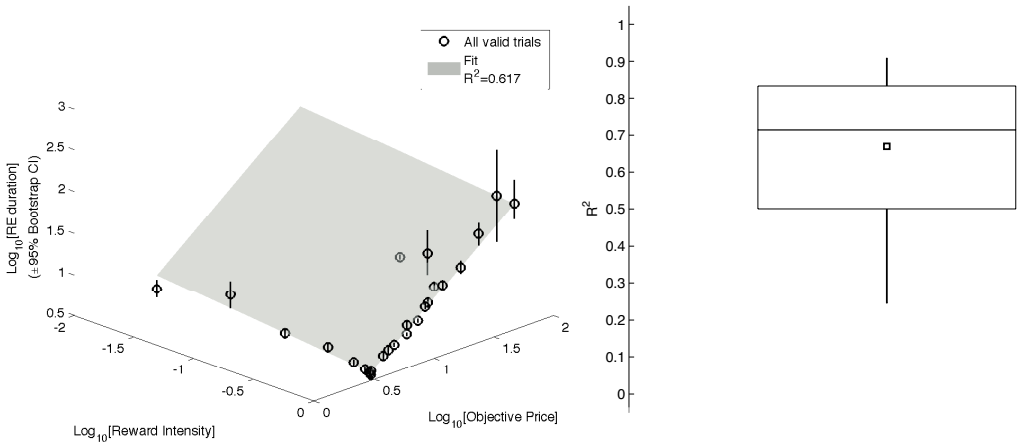

Figure 5.16. Predicting reward encounter duration for whole-trial extrapolation. The left-hand panel provides an example regression of log-reward encounter duration as a function of log-objective price and log-reward intensity for a representative animal (DE15). Circles indicate mean log-reward encounter duration (± 95 % bootstrap confidence interval), surface indicates the regression plane. The right-hand panel is a box-whisker plot of the R^2 values across all rats and conditions.

then, necessarily, the rate at which the rat leaves the hold state will be the (negative) sum over transition rates to all states that it leads to, or

$$\alpha_{hold,hold} = -\sum_{j}(\alpha_{j,hold}).$$

These rates can be summarized in a simple, though powerful matrix differential equation

$$\frac{\partial}{\partial t}\mathcal{P}[activity = i \text{ at } t] = \mathcal{A}\mathcal{P}[activity = i \text{ at } t]$$

where $\mathcal{A}$ is a matrix whose entries are α_{ij} and α_{ii}. The solution to this system of 5 linear, ordinary differential equations gives the probability, at any point in the reward encounter, that the rat engages in any activity i, assuming the transition rates are time-invariant. Since the average amount of time accumulated in each activity from the start of the reward encounter to any arbitrary time t is an integral of the probability of engaging in each activity, integrating the time spent in each activity from 0 to the point in time at which the reward encounter is ended provides an estimate of how much time is spent in each activity. The time allocation thus extrapolated (time spent in hold and tap activities divided by the time spent in all activities) provides an estimate of the proportion of time allocated per reward encounter. Multiplying this value by the probability of starting a reward encounter without a CTLB provides an estimate of what the time allocation would be, were we to extrapolate molecular performance to the level of the entire trial.

Figures 5.17 and 5.18 show the time allocation predicted by the model if we were to extrapolate the CTMC from the reward encounter to the whole trial, for each of the rats that underwent the subjective price experiment. The extrapolation was performed on these animals because of the much larger quantities of data available from which to estimate the payoff-dependent and independent hidden states. The left panels of each row are contour plots of the time allocation extrapolated from the CTMC. Often (though not always), the prediction is close to the data points, and the

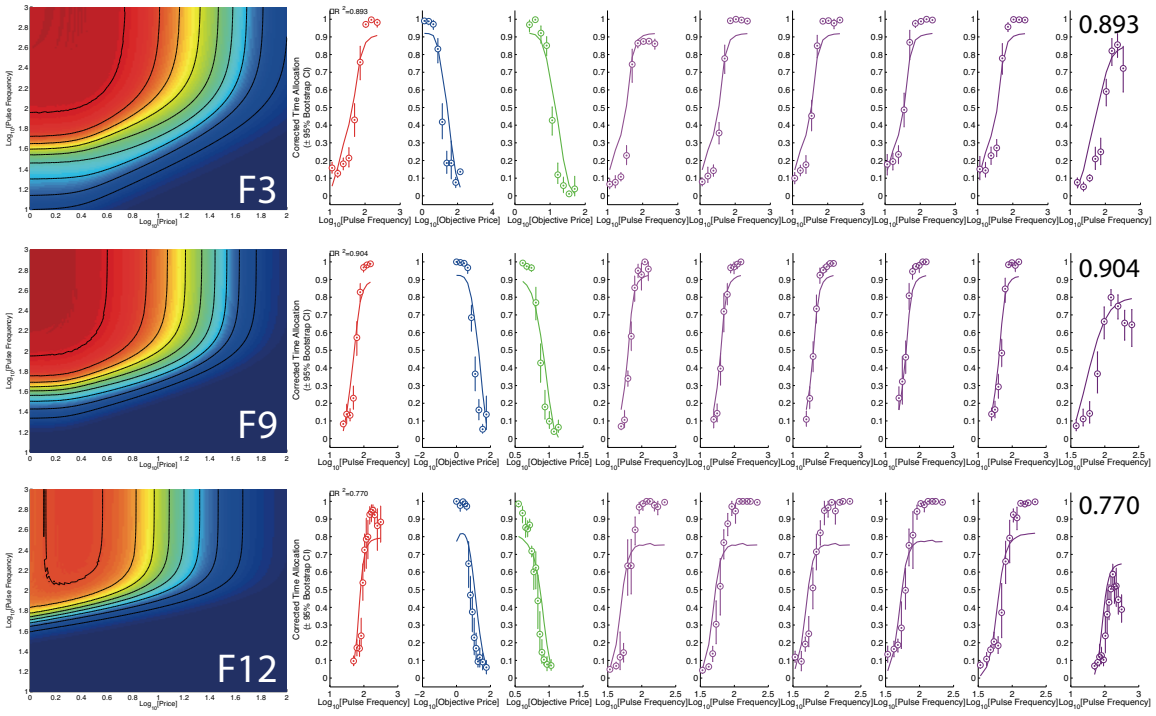

Figure 5.17. Extrapolated mountains and the comparison of molar TA to observed values for subjective-price rats, part 1. Left column depicts time allocation contours extrapolated from the CTMC. Right column compares the time allocation observed throughout the entire trial (open circles) to that predicted by the CTMC (solid lines) for each pseudo-sweep. Numbers in far right panel indicate the proportion of variance in observed time allocation values that can be predicted by extrapolating the CTMC to the entire trial.

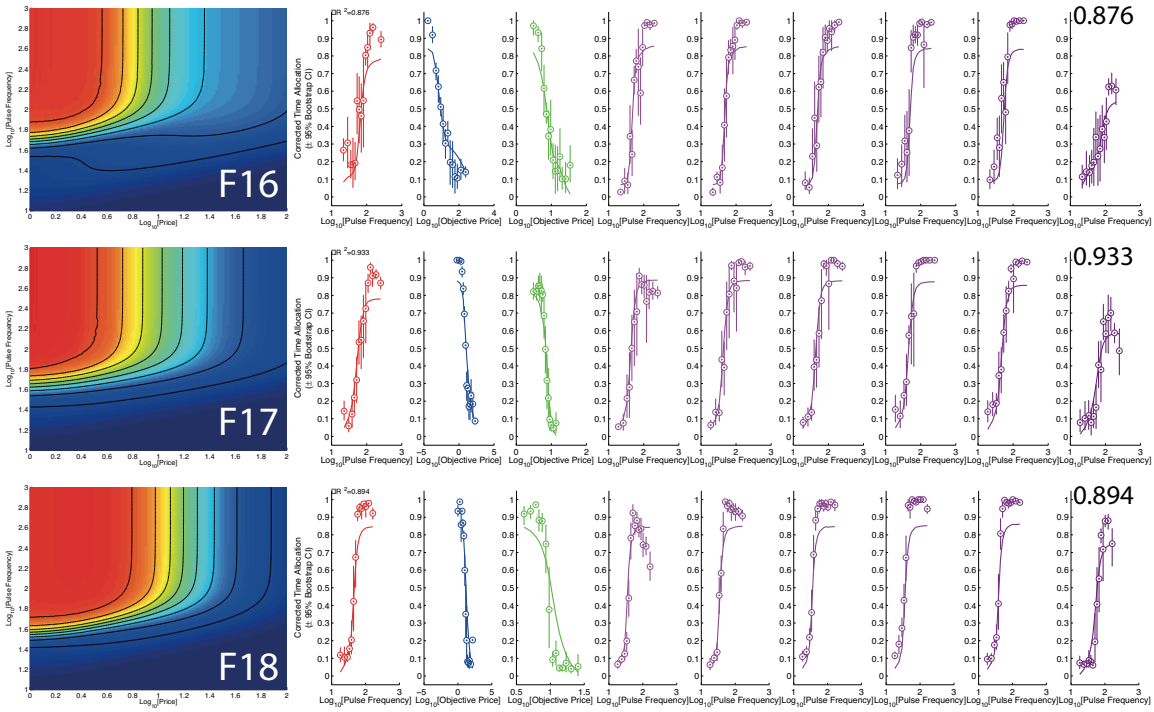

Figure 5.18. Extrapolated mountains and the comparison of molar TA to observed values for subjective-price rats, part 2. Continued from 5.17.

general dependence of molar time allocation on log-price and log-pulse frequency often has a very similar general appearance to that of the Reinforcement Mountain Model. The right panels show a comparison of the time allocation observed throughout the entire trial (open circles, $\pm$ 95% bootstrap-derived confidence intervals) with the time allocation predicted by extrapolating the CTMC to the entire trial (solid lines). There is reasonable agreement between the two, with R^2 values ranging from 0.770 to 0.933, indicated on the far right side of each row.

5.6 Discussion

We have used a continuous-time semi-Markov model to describe performance for brain stimulation reward in the test-trial phase of the randomized-triads design. The model is based on a small number of core principles, simple functions, and an assumption that responses and pauses in various observable activities are independent of the responses and pauses that preceded them.

This is in contrast to reinforcement learning models, which model performance as a punctate choice of what to do at discrete time steps (Montague et al., 1996), and melioration, which models performance in terms of the local rate of reinforcement from each source of reward. The model presented here assumes that the rat does not learn the expected payoff over a number of rewards, as would be proposed by model-free reinforcement learning (Montague et al., 1996); instead, the rat "fills in" the appropriate value for the payoff and adjusts its stay durations accordingly. The model also assumes that the key determinant of stay durations involves an accumulation of time spent lever-pressing over the entire reward encounter, weighted by the probability of reinforcement. If the local rate of reinforcement—that is, the instantaneous rate at which rewards are delivered on a moment to moment basis—were the key determinant, animals would simply stop lever-pressing after the lever is retracted

without reward delivery—making the local rate of reinforcement from lever-pressing 0. The probability of reinforcement appears to be averaged across appropriately-cued trials (the lever for these trials is generally on a different side of the chamber and a flashing cue light provides a discriminative stimulus). In our model, the determinant of the stay durations in each activity (though it is only a weak predictor of hold and tap durations) is the scalar combination of subjective opportunity cost, reward intensity, and reward probability, rather than the local rate of reinforcement. As in the model of Gallistel et al. (2001), which inspired the CTMC model presented here, performance is stochastic, and the dwell time in each state is a function of the payoffs on offer. However, the Gallistel et al. (2001) model differs from ours in that it entails sampling of dwell times from a single, exponential distribution and alternation between options according to their combined payoffs. Model selection based on the AIC shows that more complex distributions and mixtures thereof are required to describe the data adequately. The alternation principle in the Gallistel et al. (2001) model is better suited to dual-operant context than to the present single-operant context.

Our model of performance in real time differs from previous attempts: it is not based on an optimization principle, and it more readily describes performance as a series of non-punctate events. As a result, we shall discuss the overall pattern of responding that is revealed by the algorithm, the response pattern of a typical rat, the model's ability to account for trial-level performance, and its application in identifying the neural substrates that underlie the decision-making process.

5.6.1 Real-time performance

Our modelling suggests that the payoff sets the probability that the rat is in one of a small number of hidden behavioural states, which are only indirectly observable in the stream of holds and releases at the lever. These hidden behavioural states are, by and large, characterized by a constant failure rate (the distribution of

dwell times is roughly exponential) over the entire trial. The first reward delivery (in the randomized-triads design) provides sufficient information to set for the remainder of the trial:

1. The duration of PRP states,

2. the probability that the PRP state terminates at such a high rate that a reward encounter begins, for all intents and purposes, with a hold,

3. the probability that the PRP state terminates at such a low rate that the trial ends before the PRP, and

4. the duration of TLB states when the rat has not given up pursuit of the reward on offer.

In each case, the payoff effectively sets the expected dwell time in states that reflect pursuit of non-self stimulation goals: the expected dwell time in the PRP (points 1 and 2), the expected dwell time in a CTLB (3), and the expected dwell time in TLBs (4). The expected dwell time in holds or taps (Figure 5.5, 5.9 and 5.11) and the probability that a temporary lever release reflects a tap (figure 5.15) bear only a weak relationship with payoff.

What emerges then is a pattern of responding in which the duration of the post-reinforcement pause and the duration of the TLB is set by the expected payoff from self-stimulation in effect during the trial, while the duration of a hold and tapping release is not. Overall, then, our analyses allow us to describe, in real time, how the animal partitions its time among the various activities in which it can engage. Given this description, what does a typical rat do?

5.6.1.1 What does a typical rat do?

The model provides a fairly complete account of what the rat does in real time. The rat extracts a subjective estimate of the total amount of time it worked to earn a reward. It maintains a running average across multiple trials, according to

a discriminative stimulus, of the probability that the reward will be delivered. The rat then computes a payoff by performing a scalar combination of the opportunity cost with the subjective intensity of the rewarding effect and the probability that the reward will be delivered. Following the first reward delivery, performance follows the same chain, with the same probability of transitioning from one hidden state to the next, every time the lever is extended back into the cage.

The payoff sets the duration of one or two PRP states. When the payoff is high, the PRP state reaches a minimum of 0.7s; when it is low, the PRP state approaches a maximum of 18.5s. In the absence of forced motoric side-effects caused by the stimulation, very high payoffs will lead the rat to engage in a 0-duration PRP state, completely forgoing any leisure activity. The PRP states typically lead to one of two types of holds: long-hold (as long as required to obtain the reward) and short-hold (2s). When the typical rat temporarily leaves a hold state, it does so as part of a tap, which consists of two states: "short" taps (0.1s in duration) and "long" taps (0.4s). Longer, leisure-related releases are rarer, comprising a very small subset of all uncensored releases of the lever. Nonetheless, when the TLB activity is begun, it comprises two distinct hidden states with payoff-dependent mean. One of these hidden states is likely an artifact of the arbitrary tapping criterion, as its maximum is very nearly the arbitrary minimum imposed by classifying releases into taps and TLBs; the other likely represents a true leaving process.

This picture of the typical rat's activities is somewhat different than has been described previously. The rat does not appear to continuously alternate between pursuing the goals of work and the goals of leisure, as though pushed from one to the other. Instead, the rat "consumes" its leisure activities, all at once, during the post-reinforcement pause, works continuously until rewarded, or until it gives up, performing the equivalent of a coin toss at each instant of trial time with a coin whose bias depends on the payoff.

Similarly to the model of dual-operant performance proposed by Gallistel et al. (2001), the expected dwell time in each activity that reflects an underlying pursuit of leisure (PRP states and putative CTLB states) is related to the payoff derived from work activities. Unlike the Gallistel et al. (2001) model, we have proposed a non-linear function mapping the payoff from self-stimulation to dwell time. The duration of behavioural states in hold and tap activities is insensitive to the payoff derived from self-stimulation. As a result, we hypothesize that the payoff from pursuing extraneous, non-self stimulation activities would alter the duration of hold and tap activities. This idea remains to be tested, but is empirically verifiable: background stimulation trains have been used in molar studies (Arvanitogiannis, 1997) to validate the Shizgal Mountain Model. If we were to increase the payoff from pursuing non-self stimulation goals by providing electrical brain stimulation while the animal is not lever-pressing, one would expect the termination rate of hold and tap states to increase when background stimulation is available. Conversely, administering foot-shock at random intervals while subjects are not lever-pressing would decrease the payoff derived from pursuing non-self stimulation goals. As a result, one would expect the termination rate of hold and tap states to decrease when background foot shock is possible. In both cases, neither increasing the payoff from other activities by providing background stimulation, nor decreasing it by providing foot-shocks, should alter the mean dwell time in PRP and TLB activities.

In modelling the stay durations in each activity, we can further distinguish, in principle, two types of hidden behavioural states: behavioural states that are hidden by virtue of our inability to detect them without sufficiently sensitive equipment ("hidden to us") and those that may not even be detectable by behavioural means alone ("truly hidden"). As an example of "hidden to us" states, the rat that has quit may have opted against lever-pressing for the remainder of the trial, but may still face an action selection problem that can be modelled by the CTMC, provided we have

appropriate equipment to detect that the animal is grooming, resting, or exploring. As an example of a "truly hidden" states, it may not be possible to visually detect the difference between a "long" hold, lasting on the order of tens of seconds, from a "short" hold, lasting on the order of a few seconds, without fitting a mixture model of the hold durations. The value of the CTMC lies not only in its ability to detect what may be different actions that could be detected with sensitive equipment, but also actions that may look alike to the casual observer. The addition of video camera- and accelerometer-based systems may help make this CTMC model richer, but may not provide all the necessary information. For example, the typical rat has two modes of holding, and the only way to distinguish them is on the basis of their relative means: one appears to be short while the other is long. We find it unlikely that a review of a video record would provide more information about what these two modes of holding represent.

With such a rich, non-normative model of real-time performance, what can be said about its predictions at the level of the whole trial, on the order of minutes to hours? If molar performance is the result of molecular performance, what does the CTMC model at the level of individual holds and pauses say about the whole trial?

5.6.2 Matching is an emergent property of the CTMC

The CTMC, as modelled and as described, makes no explicit assumptions about the matching law, melioration, or reinforcement learning. We used two core principles:

1. that the relevant determinant of the rat's decision to press (the payoff) is an expectation for the scalar combination of subjective opportunity costs accrued over the entire encounter, reward intensities and probabilities of reinforcement, and

2. that this determinant sets what the animal actually does, be it directly observable (activities) or latent (hidden behavioural states).

With only these two core principles, it is possible to re-create performance that is close to the matching law in the six subjects for which there is the most data (F3, F9, F12, F16, F17, and F18). Although no parameters have been added to the CTMC to ensure that the proportion of total time allocated to self-stimulation activities match the suitably-transformed ratio of the payoff from self-stimulation to the sum of suitably-transformed payoffs from each activity, the CTMC often re-captures this relationship. By virtue of the way each piece fits with each other piece—the payoff-dependent probability per unit time of ceasing a PRP state (including those with higher- and lower-than-measurable termination rates), the payoff-dependent probability of long-mean TLB states, and the presumed payoff-independent probability of each hold and tap state—the CTMC reveals that matching on the molar scale is simply an emergent property of the molecular interactions of various hidden exponentially- and gamma-distributed behavioural states that can be quite succinctly summarized in terms of their termination rate or mean-dwell time. Future studies will be required to assess whether this is true in more data-poor cases, and if not, the conditions under which this result obtains. Nonetheless, that performance at the molar level can be reasonably well predicted (at least 77% of the variance in molar time allocation is accounted for) by the multiple interacting pieces of the CTMC is an encouraging first step to understanding how animals partition their time among competing alternatives.

The CTMC model described here provides a potentially powerful tool for investigating what the animal is doing—in real time—as well as what manipulations may do at a level that has heretofore been impossible to assess. The CTMC may even shed light on the answer to where the components of action selection are represented in the brain.

5.6.3 Neural substrates

The CTMC allows us to summarize the behavioural components of real-time performance with a relatively small number of distributional parameters. The summary is that much simpler when we consider that a large proportion (the first bin of the log-shape parameter histograms of figures 5.7, 5.9 and 5.13 provide graphical evidence of this) of the hidden behavioural states extracted by the algorithm are exponentially distributed, requiring only a single parameter that sets the constant rate at which the behavioural state terminates.

An easy implementation of this scheme in neural circuitry involves a winner-take-all competition between populations representing the various possible actions (Neiman and Loewenstein, 2013). Populations of neurons, connected by reciprocal inhibition to each other and recurrent self-excitation, provide a source for which activity the animal chooses to engage in. If neurons representing the pursuit of BSR are active, they will inhibit neurons representing the pursuit of other goals. For example, sub-populations of neurons representing "hold the lever for a while" may equally make reciprocally inhibitory connections with those representing "hold the lever for only a bit" as well as self-excitatory connections. The most active sub-population of neurons will excite itself greatly, inhibit the others, and provide the animal with a temporary goal to pursue. The degree of noise in the representation of competing states, their ability to drown out the currently active sub-population, and the currently active sub-population's ability to excite itself will each determine how long a sub-population maintains control of the animal's behavioural output.

Given that spike rates can be modelled as Poisson processes with a constant termination rate (Werner and Mountcastle, 1963), the fact that most hidden behavioural states can be described with a single termination rate parameter is rather encouraging. The termination rate of a behavioural state and its probability of oc-

currence would certainly be related to the firing rate of the neurons involved in its representation. For example, if a sub-population is involved in representing a hold state of 1s duration, then that population ought only be active while the animal unequivocally emits a short-duration hold. We would expect that sub-population to be active only when the animal is in that behavioural state, and it would be possible to decode, from population activity, the probability that the rat is in that behavioural state using Bayes' rule: the probability of being in state "short-hold," assuming a particular response from a large ensemble of neurons, can be decoded from the neural response when the rat is in various behavioural states, the probability that the rat is in this behavioural state, and probability of the neural response.

The CTMC thus also provides a solid behavioural basis for interpreting ensemble recordings while animals engage in a single-operant task. The most rigorous way to attribute a neural representation to activity within a population of cells is to use Bayes' rule to turn the tuning curve of each neuron for the proposed psychological phenomenon ($\mathcal{P}[activity|state]$), each neuron's baseline firing rate ($\mathcal{P}[activity]$) and the probability of representing the phenomenon ($\mathcal{P}[state]$) into a decoded probability of the phenomenon being represented. For example, some neurons (population A) may be more active than usual when the rat is in the "holding patiently" state; the tuning curve of population A would provide the probability that these neurons fire as a function of the duration of the hold, but not the probability that the rat is engaging in a long hold based on the population's firing rate (the decoded probability). Provided one is recording from multiple units with different holding-related tuning curves, a simple application of Bayes' rule

$$\mathcal{P}[state|activity] = \frac{\mathcal{P}[activity|state] \cdot \mathcal{P}[state]}{\mathcal{P}[activity]}$$

would be sufficient to relate activity in a population of neurons to a putative hidden

behavioural state.

One important population of neurons to consider with the CTMC is in the ventral striatum. The nucleus accumbens shell may be one candidate region that translates the payoffs from different goals into action (Yin et al., 2008), thereby underlying the various behavioural states of the CTMC. Prefrontal cortical regions may provide the information regarding the expected payoff from each goal (Cohen et al., 2002), combining the subjective opportunity costs, reward intensities, and risk involved in acquiring each. Finally, dopamine neurons of the ventral tegmental area may modulate activity within the nucleus accumbens (Goto et al., 2007), making sub-populations representing some behavioural states more resilient to competition from other behavioural states. The roles of each of these regions in the animal's behavioural output could easily be assessed with the CTMC model presented here, and hypotheses concerning how each nucleus contributes to performance would be readily tested empirically.

5.6.4 Concluding remarks

Our model opens a new universe of possibilities for investigating the effect of manipulations to reward-valuation circuitry on hitherto unmeasurable aspects of performance. One interesting question regards the effect of psychostimulants such as cocaine on the pattern of responding. The Shizgal Mountain Model has been useful in identifying the stage of processing at which manipulations act, but cannot on its own determine how it impacts performance on the molecular level. For example, the Mountain Model has identified that cocaine affects the neural circuitry of reward beyond the output of the network that carries out a spatio-temporal integration of the injected signal. In light of these findings, one could argue that the increase in dopamine efflux to the nucleus accumbens produced by cocaine administration scales the translation of injected pulse frequency into subjective reward intensity.

This explanation posits that higher dopamine tone in the nucleus accumbens would makes all rewards more valuable by a constant factor. Others have argued that cocaine alters the animal's proclivity to invest effort into acquiring electrical rewards (Salamone et al., 2005). A final means by which cocaine would alter reward processing beyond the spatio-temporal integration of action potentials elicited at the electrode tip involves decreasing the payoff from alternate activities in the box. The CTMC could provide an answer to how cocaine might affect performance on the molecular level. It may alter only the payoff from self-stimulation activities (either by scaling the reward intensity or making pursuit of brain stimulation more effortful), which would alter the dwell time in activities related to the pursuit of leisure rewards (PRP and TLB state durations, and well as CTLB probabilities). It may alter only the payoff from other activities, which we hypothesize would change the duration of states related to the pursuit of work rewards (holds and taps). A large number of questions regarding how a particular manipulation of neural circuitry—a pharmacological manipulation, a lesion, or a physiological challenge—impacts the patterning of performance in real time can now be readily answered thanks to our CTMC model. The development of new techniques, like optogenetics, will even provide tools to assess the effect of manipulations that occur on the same time scale as the CTMC, that is, in real time and on the order of milliseconds.

Although causal manipulations and lesion studies can easily be conducted in the context of the CTMC model, they are by no means the only way the CTMC model is useful in understanding how the brain evaluates and decides. The real-time nature of the CTMC can just as readily be transported to the study of neural correlates of any behavioural state, presumably resulting from neural activity in neurons that can be identified, to study where the determinants of action selection may be represented individually and where they have been combined. The new methodology presented here for succinctly describing the rich patterning of responses in a single-operant

context provides a springboard for a new era in understanding how the brain evaluates and decides on a moment-to-moment basis.

Chapter 6

General discussion

Before the start of a self-stimulation session, the typical subject appears to be oblivious to its environmental conditions. It often grooms itself as a consequence of having been handled by a large primate, sniffs around the operant chamber, and rests in the corner as the experimenter programs the day's experimental protocol. As soon as the large house light begins to flash, the rat awakens from its apparent stupor, and priming stimulation invigorates its movements. As the lever extends into the operant chamber, the rat leaps to the manipulandum and begins holding it down.

How the animal selects what to do, when to do it, and for how long to do it, has been a burning question in psychology, neuroscience, ecology, and—in the case of the human animal—economics. Answers to the question of action selection inform much more than the question of motivation. In order for an animal to select an action, the animal must learn action-outcome pairs in as simple or complex a sense as necessary (thereby engaging learning systems), convert the outcomes of actions into a common currency (thereby engaging valuation systems), maintain at least an ordinal preference for the actions it can take (thereby engaging action-incentivizing systems), and keep track of the potentially fluctuating payoffs it has obtained from the actions available to it (thereby engaging mnemonic systems). An animal that is unable to learn that a particular patch is bare, or remember that the patch is rife with predators, or convert the benefits from food and those from sex into a common currency, or at least order foraging at two patches in terms of their desirability, would very quickly become extinct.

6.1 The action selection problem

The action selection problem itself has been traditionally considered in the context of instrumental conditioning. The procedure is naturally suited for answering questions about how to choose what to do and how long for which to do it: the animal opts between operant A and operant B, or operant A and everything else, and the action selected directly reveals the animal's preference between the two. It is possible to then study how various manipulations will change what the animal elects to do, either at the molecular level of individual holds and pauses or at the molar level of the entire trial.

What we describe below are three descriptions of action selection. Reinforcement learning models, with their deep roots in artificial intelligence and machine learning, provide a normative account of what an animal ought to do in order to maximize its total net rate of reward. Model-free reinforcement learning implies that rats learn only the net reward at any given point in time, a view consistent with a habit-learning system that is insensitive to the identity of an outcome and maintains a representation of only the net reward that can be expected. In other words, the rat forms only a representation that lever-pressing will lead to a reward, and nothing else. Model-based reinforcement learning implies that rats learn not only the value of an action in a particular state (e.g., pressing the lever is "good"), but also form a representation of the state that will result from that action (e.g., pressing the lever will lead to a delicious banana-flavoured food pellet). In this case, what is learned is a model of the world, a view consistent with a goal-learning system that maintains a representation of both the net reward and the identity of the outcome that can be expected. A different, non-normative description of action selection takes its pedigree from the early days of operant psychology. The matching law (Herrnstein, 1961) is based on the observation that animals will approximately match the relative

rate at which they respond to one operant (and therefore select it) to its relative rate of reinforcement. The law was subsequently extended to the single-operant context by assuming that the choice is between the experimenter-controlled action and everything else the subject can do while in the chamber. Although some (Sakai and Fukai, 2008) have linked the Matching law to reinforcement learning models, this also requires matching behaviour to result from the steady-state of a gradual learning process, a result that (as we shall discuss later) does not obtain.

6.1.1 Model-free reinforcement learning

A highly general description of action selection simply requires the rat to maintain stimulus-response contingencies: responses that lead to desirable stimuli are strengthened, while those that do not are weakened. As a result, no action need be "selected": the action with greatest associative strength with rewarding stimuli is that which the subject performs.

In natural settings, however, the rat does not have an explicit "teacher" for which responses are desirable and which are not. The animal must explore a space of responses, and assign credit for reward to an action it has taken in what may have been a long chain of responses. According to reinforcement learning accounts, the animal solves this assignment of credit problem by way of an internal "critic" that performs an estimate of the total future net reward that can be expected from performing any particular action. The "actor" then evaluates the optimal policy to implement in order to obtain the greatest total net reward. The critic maintains an estimate of the total future net reward that is expected in a particular state, and if it is surprising—that is, the reward magnitude that was delivered is different from what was expected—the critic module will update the expectation as a function of the learning rate. The actor module will then use this updated expectation to tune performance to maximize the total net reward the rat obtains.

For example, suppose a monkey must maintain a key press when it sees a yellow light in order to obtain a juice reward (Apicella et al., 1991). The stimuli are presented at randomly chosen intervals, but (assuming the key is still pressed) the juice reward is delivered two seconds after light onset. If the monkey behaves as a model-free reinforcement learning agent, the first time the juice is delivered, at $t = 2$s, it is a surprising event, with a reward of $R(S_{t=2})$, prompting the critic module to update the value of the state "light came on" (with presumably 0 expected value) at $t = 2$ by some learning factor α of the discrepancy:

$$V(S_{t=2}) \leftarrow 0 + \alpha \times (R(S_{t=2}) - 0).$$

If the learning rate is 1, the new value of the state at $t = 2$ will be the last delivered reward, while a learning rate of 0 will mean the value of the state is never updated. Now let us suppose the light comes on again, following an inter-stimulus interval of random duration. The value of the state "light came on" is no longer 0, because it incorporates both the immediate net reward (0) with the total discounted value of future states. This, of course, differs from the initial estimate of 0, which will drive the critic module to update the value of the state at $t = 0$ according to the discrepancy between the expected reward (0) and the sum of the immediate net reward (0) with the discounted total net future rewards (of which $V(S_{t=2})$ is one such future reward). Over repeated presentations, the value of the state at each point in time converges onto a "true" value, a process which is called "value iteration." In this scenario, the monkey need not learn which states follow which other states; all that needs to be learned is that maintaining the key press will lead to a certain level of reward.

The time required to obtain an accurate estimate of the value of states is inversely related to the learning rate of the process. If α in the expression above is 1, then the expected value of the state "light on" two seconds after the onset of the cue

($V(S_{t=2})$) is immediately equal to the value of the juice reward. In very few trials, the expected value of trial states at any time after the cue light is illuminated will have converged to their "true" values. In natural settings, however, there is variability in the amount and timing of rewards that an animal will receive, and a high learning rate will drive the animal to give an inordinate amount of weight to these deviations. If the learning rate is 1 and the monkey expects that a reward of magnitude 10 will be delivered at $t = 2$, presenting a reward of magnitude 10 at $t = 1.9$ or a reward of magnitude 5 at time $t = 2$ will both drastically alter the value of the state "light came on" at all time points. Indeed, too high a learning rate and an animal navigating its environment would be pulled very strongly by the rare events occurring simply because of random sampling. Too low a learning rate and the animal would require too many trials to form an accurate representation of how much reward to expect from its environment. Without some model of the world, what the animal should or should not do is guided by feed-back of the total net reward that has followed previous actions, until the function that maps values to states and actions at any particular time ($V(S_t)$) is maximized.

6.1.2 Model-based reinforcement learning

In contrast, model-based reinforcement learning requires the rat to learn not only that taking action A at time step t results in a particular total net reward, but also the identity of that reward. In other words, the rat constructs not only a stimulus-response map of the total future reward that can be expected from making a response at a particular time, but also a stimulus-stimulus map of the stimuli that can be expected from making a response at a particular time. Model-based descriptions allow multiple conclusions and representations of what the state of the world will be: if one takes action A at time t, one can expect state S' with total future net reward r. This sort of description, and its contrast with model-free reinforcement learning,

has been particularly useful in formalizing the difference between habit-based and goal-based decision-making.

Habit-based systems of decision-making are insensitive to reinforcement devaluation (Dickinson et al., 1995), a result one would expect if model-free reinforcement learning mechanisms underlie these systems. When a rat is extensively trained to lever-press for food, habit-based systems are at play when the food is subsequently paired with illness but the rat continues to lever-press. Although the reinforcer has been devalued, the rat continues to respond because, in the model-free account, the rat has learned that lever-pressing leads to a desirable total future net reward.

In contrast, goal-based systems of decision-making are sensitive to reinforcer devaluation, a result one would expect model-based reinforcement learning mechanisms underlie these systems. When a rat has not been extensively trained to lever-press for food, goal-based systems are at play when the food is subsequently paired with illness and the rat ceases to respond. The rat ceases to respond because, in the model-based account, it has learned not only that lever-pressing leads to a desirable total future net reward, but also that lever-pressing leads to food. Since that stimulus (food) is no longer desirable, the rat stops lever-pressing.

Model-based reinforcement learning descriptions of action selection are easy to apply when the stimuli resulting from an action are readily observable. There is no reason the stimuli cannot be, in principle, very complex. In this case, a model-based description is still easily applicable. For example, the chess player cannot rely on which next move will allow him to win (though that is certainly helpful). The expert chess player also needs to know which moves will lead to board configurations that are more desirable than others, projected many steps into the future. Thankfully, we do not approach all of life's decisions the same way an expert chess player evaluates their strategy to beat Gary Kasparov, or we would all be lost in thought.

Any problem that can be solved by model-based reinforcement learning can,

in principle, be solved by a biological agent, provided sufficient representational capacity. For example, given infinite time and resources, the rat can form a model of its environment that is sufficiently rich to account for the ripening of food sources, the rise and fall of watering grounds, and the wax and wane of predator populations. However, just as with the above chess example, a combinatorial explosion results from increased representation: there are an estimated 5×10^{52} board configurations (Allis, 1994), and neither human beings nor computer programs have ever been capable of representing them all in the service of a policy. This does not prevent people unlike this author from being very good at chess, but it does suggest that certain problems which can be solved in principle simply cannot be implemented in brains with physical limitations.

The situation is slightly more difficult to envision when the stimuli are ill-defined and internally-generated. Suppose the action an animal ought to do next is informed not only by a distinct cue, but also by the subjective opportunity cost and reward intensity that has just been in effect. At which point does the spirit of model-based reinforcement learning break down? When the key determinants of decision are themselves stimuli for action selection, model-based descriptions become indistinguishable from feed-forward descriptions that propose that action selection results from the payoffs that can be expected from pursuing each goal through an evaluation of long-term trends in the mean and variability of payoffs in the past.

6.1.3 Matching

Matching refers to the experimental observation that animals will match the relative rate of responding for an option to its relative rate of reinforcement (Herrnstein, 1961). The matching law implies that the relevant determinant of the action to select is the rate at which it provides reinforcement. This is consistent with our view that the key variable in action selection is a scalar combination. The idea that the rel-

evant variables in determining matching behaviour were subjective was incorporated early on (Killeen, 1972). Similarly, payoff involves subjective variables: subjective opportunity cost, effort cost, reward intensity and risk. In the single-operant context, the matching law reduces to a choice between pursuing experimenter-delivered rewards and everything else, so that the rate of responding is related to the relative rate at which experimenter-controlled rewards are delivered compared to that at which other rewards are delivered.

Two mechanisms have been proposed by which matching occurs, neither of which has been entirely successful. The first is melioration (Vaughan, 1981), which describes matching as the result of changing local rates of reinforcement: when the local rate of reinforcement from one option falls below that of its competitor, the animal switches to the competing action. If, for example, the animal responds at a rate of 5 presses per minute on operant A delivering rewards at a rate of 1 per minute, and at a rate of 25 presses per minute on operant B delivering rewards at a rate of 1 per 10 minutes, the local rate of reinforcement from A is 1/5, whereas that from B is 1/250. Since the local rate of reinforcement from A is considerably higher than that from B, the animal will switch to the richer schedule. When the animal is matching, the animal responds at a rate of 10 presses per minute on A (yielding local rate of 1/10) and at a rate of 5 per minute on B (yielding a local rate of 1/10).

A second mechanism is maximizing (Green et al., 1983), which describes matching as the result of optimizing the total rate of reward. The variable interval schedules that typically control the delivery of rewards hold rewards indefinitely after a programmed interval has lapsed. As the animal responds for longer periods of time to one option, the probability that a reward is waiting at the other option increases. In the above example, if the animal simply ignored the leaner operant, it would collect a reward every second. By matching, the animal collects one reward every second and another reward every 10 seconds. There is no other strategy that

will improve the overall rate at which the animal collects rewards, so in the typical scenario, matching is simply a maximizing strategy.

Neither of these two accounts can adequately account for performance in atypical operant contexts. If response rates are an indirect proxy for the amount of time an animal decides to spend pursuing an option, melioration implies that stay durations at each of two operants will be dictated by both the programmed relative rates of reinforcement and unusually long intervals sampled from the variable-interval distributions. If local rates of reinforcement were the source of matching behaviour measured on the molar level, then unusually low rates of reinforcement would alter the local rate of reinforcement and produce changes in performance. Unfortunately for the melioration hypothesis, stay durations were not influenced by unusually long inter-reward intervals (Gallistel et al., 2001).

Herrnstein and Heyman (Herrnstein and Heyman, 1979) tested pigeons on a concurrent VI/VR schedule of reinforcement, under which one pecking key delivered rewards according to a variable interval and another according to variable ratios. Pigeons matched their relative responding to the relative obtained rates of reinforcement. The maximizing strategy would be to respond on the VR schedule to a much greater extent than the equivalent VI. Instead, pigeons had a strong bias toward the VI schedule, rather than the VR schedule predicted by a maximization account. However, Green et al. (1983) argued that the value of extraneous activities, when added to the value of lever pressing for the VR schedule, could account for the bias toward the VI schedule. When responding on a concurrent VR-VR schedule in which responses to one side increases the ratio on one side and responses to the other side increase both ratios, pigeons acted according to a maximizing strategy.

If reinforcement learning models are inadequate for explaining continuous, payoff-dependent stay durations in the randomized-triads design, and neither molecular accounts of matching can provide a satisfying answer to the real-time patterning

of performance, how can we model how the self-stimulating rat selects among the competing activities that are available to it?

6.2 Performance in the randomized-triads design

We have collected data that provide a framework within which to study the action selection problem. In the following section, we shall argue that working for brain stimulation rewards is the product of a series of non-linear mappings, that the self-stimulating rat extracts statistical regularities from the test trials it encounters regarding both dynamic and static periods, and that actions are selected on the basis of the payoffs that may be derived from performing them. Our data also suggest that the way the animal partitions its time between work (lever-pressing) and leisure (everything else) involves consumption of the benefits derived from each in a largely all-or-none manner, rather than through rapid alternation between exertion and rest.

6.2.1 Molar-level

At the molar level, we have presented a further validation of the Shizgal Mountain Model's ability to accurately detect an effect that occurs beyond the spatio-temporal integration of the injected electrical reward signal. The model proposes that the overall proportion of time allocated to self-stimulation is the result of a non-linear behavioural allocation function, which takes as its inputs the payoffs derived from lever-pressing and those derived from performing other activities in the operant chamber. The payoffs themselves involve a scalar combination of subjectively-mapped variables that can, in principle, be controlled experimentally: the subjective opportunity cost, effort cost, reward intensity, and risk. These subjective determinants are often non-linear functions of directly manipulable variables: the price (the required total amount of time the lever must be held to earn a reward), the force required to

hold down the lever, the physical characteristics of the electrical stimulation train, and the probability of reinforcement. Since the mappings are non-linear, changes in objective stimuli do not necessarily result in equivalent changes in their subjective impact. Beyond a certain point, decrements in price are ineffective at lowering the subjective opportunity cost (Solomon et al., 2007). Subjective effort costs should explode when the force required to hold down the lever is beyond the maximum physical work the subject can exert.

Thanks to the non-linearities, however, changes of scale (or, alternately, gain) are separable from changes in threshold (or, alternately, sensitivity). For example, changes in the maximum intensity of the rewarding effect (a change of scale/gain) can be distinguished from changes in the relative impact of each electrical pulse on the intensity of the rewarding effect (a change of threshold/sensitivity). A manipulation that makes all rewards better by a constant factor can be distinguished from one that disproportionately amplifies weak rewards.

The strength of the Mountain Model resides in its capacity to take advantage of these non-linearities. The post-synaptic effect of the stimulation's pulse frequency, for example, grows as a power function at low pulse frequencies and rises to an asymptotic value at high pulse frequencies (Simmons and Gallistel, 1994). The scalar combination of the key determinants of decision (subjective intensity, cost, exertion, risk, etc.) will drive performance according to another non-linear function: the relationship between payoff and performance is a sigmoid, with a maximum at high payoffs, a minimum at low payoffs, and a smooth transition between the two (McDowell, 2005). In each case, the non-linearity allows assignment of three parameters—sensitivity, gain, and slope—where a linear function would allow only two (slope and intercept). When measuring the variables together, changes to the sensitivity of the function mapping pulse frequency to reward intensity (F_{hm}) can be distinguished from changes the sensitivity of the function mapping payoff to performance (P_e).

6.2.1.1 Detecting an effect occurring downstream from the spatio-temporal integrator

We have presented evidence (Breton et al., 2013; Arvanitogiannis and Shizgal, 2008) that the Mountain Model can, indeed, detect changes that are known to affect the post-synaptic integration of the directly-activated substrate for self-stimulation. By increasing the duration of the train, neurons can be stimulated at a lower rate in order to achieve a given (half-maximal) level of subjective reward. By decreasing it, neurons must be stimulated at a higher rate in order to achieve that same level. While the Mountain Model accurately detected changes in the sensitivity of the function mapping pulse frequency to subjective reward intensity in all animals, in a subset of animals, the model also detected changes to the sensitivity of the function mapping payoff to performance. Further modelling demonstrated that these results could be explained if, in those animals, at least two spatio-temporal integration systems had been activated, consistent with a widely-discussed hypothesis that multiple subsystems may be involved in the temporal integration process (Arvanitogiannis et al., 1996; Fulton et al., 2006; Shizgal et al., 2001; Carr and Wolinsky, 1993) This explanation is predicated on the assumption that the model can, indeed, correctly identify the stage of processing for manipulations occurring beyond the peak-detection stage. The Mountain Model has been further validated here, showing that a change in the probability of reinforcement produced large changes in the marker for manipulations occurring beyond the peak detection stage, with only small or time-inconsistent changes in the marker for manipulations occurring prior to peak detection.

6.2.1.2 Quantifying subjective risk

If we assume that no variables other than risk have changed as a result of making the reward probabilistic, it is also possible to evaluate the degree to which probability results in subjective risk via the Mountain Model. The degree to which

the rat must compensate for the lower probability in terms of a constant subjective opportunity cost is, in effect, the subjective impact of probability. If a maximally intense reward requires a price of 30 seconds for the payoff from self-stimulation to be equal to that of non-self stimulation, then one would expect a maximally intense reward delivered with 50% subjective probability to require a price of 15 seconds for the payoff from self-stimulation to be equal to that of non-self stimulation. In other words, assuming that only risk changes, for all four payoffs to be equal, the following two equalities must be met:

$$(I_{max}/SP_{e_{1.0}}) \times \text{no risk} = U_e,$$

and

$$(I_{max}/SP_{e_{0.50}}) \times \text{subjective risk for 50\% probability} = U_e,$$

where I_{max} is the maximal intensity of a reward, U_e is the payoff from non-self stimulation (everything else), and P_{e_i} is the price at which the payoff from a maximally intense brain stimulation reward is equal to that from everything else. Since these two expressions must be equal to each other, the ratio of $P_{e_{1.0}}$ to $P_{e_{0.5}}$ will be equal to the ratio of the risk associated with a reward delivered with probability 1.0 to the risk associated with a reward delivered with probability 0.50.

Our results demonstrate not only the validity of the Mountain Model in correctly identifying an effect occurring downstream from the spatio-temporal integration of the injected signal, complementing previous validation attempts (Breton et al., 2013; Arvanitogiannis and Shizgal, 2008), but also the computational power afforded by the model. By making only one assumption—that risk should not affect the intensity of the rewarding effect of brain stimulation, the subjective effort cost, or the payoff from everything else—our model has also provided some preliminary evidence that the mapping of probability of reinforcement to risk is scalar, or nearly so, over

the range tested.

Overall, rewards delivered with probabilities of 0.75 or 0.50 tend to be devalued to an extent that would be normatively expected. The median change in P_e from a probability of 1 to a probability of 0.75 was found to be 0.13598, corresponding to a subjective risk of 73%. The median change from a probability of 1 to a probability of 0.5 was found to be 0.33089, corresponding to a subjective risk of 47%. These represent modest differences, and they underline the importance of rigorous psychophysical scaling when studying the variables that affect what an animal decides to do and for how long they choose to do it.

The Mountain Model provides a molar description of performance for brain stimulation rewards. It assumes that the rat's allocation decision, at the level of the whole trial, is based on the payoff derived from self-stimulation activities. Our experiments, however, contain structure at a finer temporal scale than the trial. We shall now turn our attention to how the rat makes molecular decisions about which action to take and for how long to take it based on the session's structure, at a different level than can be explained by the Mountain Model.

6.2.2 World model of change

In the randomized-triads design, the rat is presented with a repeating pattern of trial types. Each trial in a session is cued by the flash of a yellow house light during a 10-second inter-trial interval, 8 seconds into which a single train of high-pulse frequency priming stimulation is delivered. All experimentally manipulable variables are constant for the duration of the trial. The pulse frequency, pulse current, pulse duration, train duration, price, force, and probability of reinforcement are all fixed from the time the house light stops flashing to the time it begins again. The trials within a session progress in highly structured manner, according to triads. The first trial (the leading bracket) is characterized by a strong reward (as high as the animal

can tolerate) delivered at a negligible price (1s). At the start of the second trial of a triad (the test trial), the pulse frequency, price, and probability of reinforcement are sampled from a large randomized list without replacement. The third trial (the trailing bracket) is also characterized by a negligible price, but the reward on offer is sufficiently weak (10Hz) that the rat ought never be motivated to work for it. This repeating pattern is presented for hours at a time, for days, until the list is exhausted, at which point the characteristics in effect on test trials are sampled from a new randomized list.

Such a repeating pattern may be difficult to a naive observer to detect. Indeed, were the task on a human observer's part to detect the apparent direction of motion in an array of randomly moving dots against a contrasting background, the same pattern of trial presentations may not be obvious at first glance. Suppose the first trial of a repeating sequence of three presented a large proportion of dots moving in the same direction with high contrast, the second presented a randomly selected proportion and contrast, and the third presented a stimulus in which no dots moved in the same direction against a highly contrasting background. How long would it take for a human observer in these conditions, instructed to identify whether the dots appeared to move in a particular direction and unaware of the triad structure, to act on the basis of this repeating cycle?

Our results imply that rats, indeed, are capable of extracting the statistical regularities inherent in the randomized-triads design. At the very least, they act on the basis of an expectation for payoff that they would not have were it not for some model of the world. Without a simple model of how sessions progress, rats in the randomized-triads design would begin working for brain stimulation rewards as a function of the last trial type encountered. Upon starting a new trial, a rat using model-free reinforcement learning principles will behave as though it expects the next reward to be like the last, leading the animal to produce long pauses following

a trailing bracket trial, short pauses following a leading bracket trial, and an average pause of intermediate duration following test trials. Instead, our results show definitively that the very first pause the rat makes—before it can know anything about the payoff from self-stimulation—depends on the payoff of the trial type to come, in feed-forward fashion, rather than the trial that has just elapsed. Rats spend little or no time pausing on leading bracket trials, slightly more time in this pause on test trials, and rarely ever bother to lever-press on trailing bracket trials. Since there is no other way for them to know what the payoff will be on a particular trial, and every trial is cued the same way, they must form an expectation for what the payoff will be based on a stimulus that is not directly observable.

These data also give hints about the identity of the stimulus that allows rats to form this expectation. In cases where animals make shorter-than-usual pauses at the start of the trailing bracket trial, the preceding test trial presented stimulation that was very cheap and very strong or very weak—not unlike the characteristics of the leading and trailing bracket trials. Rats therefore appear to employ two sets of syllogisms based on the trial that came before. If the last trial was similar to a trailing bracket trial, then the next trial will be a leading bracket; if the last trial was similar to a leading bracket trial, then the next trial will be a test trial. On the test trial, the price and reward intensity may sometimes be sufficiently similar to either of the bracket trials. As one would expect if the rat had maintained a set of three, two-trial sequences, these misleading test trials result in an uncharacteristically short pause on the subsequent trailing bracket trial. At sufficiently low prices and reward intensities, the rat behaves as though it believed the test trial was a trailing bracket, and takes a very short pause on the trailing trial rather than a long one. At sufficiently high intensities and low prices, the rat behaves as though it believed the test trial was a leading bracket, and takes a short pause rather than a long one.

Furthermore, the uncharacteristically short pauses on trailing bracket trials

occur predominantly when the subjective opportunity cost is sufficiently similar, and not necessarily when the objective opportunity cost is similar. When the price is below a minimum value (approximately 3 to 4s), the subjective opportunity cost ceases to change (Solomon et al., 2007). If the rat discriminated among trial types according to the objective price, test trials with very low prices would differ from the bracket trial price of 1s, thereby allowing the rat to discriminate among these trial types and prevent any confusion. If the rat made the discrimination on the basis of the subjective opportunity cost, test trials very low prices would have the same subjective opportunity cost as bracket trials, and thus, mislead the rat to infer that the next trial is not a trailing bracket trial. When the prices presented on test trials have a similar subjective opportunity cost to the price in effect on bracket trials (1s), regardless of its objective value, the rat is misled into taking a shorter-than-characteristic pause on subsequent trailing trials. This indicates that the relevant stimuli that cue what the next trial will bring are subjective.

Finally, when the test trial presents stimulation that is very expensive, and therefore the payoff from self-stimulation is low, rats make their characteristically long pause on the subsequent trailing bracket trial. In other words, the appropriate cue is a compound stimulus involving both the appropriate subjective opportunity cost and subjective reward intensity together, rather than their scalar combination. Moreover, this compound stimulus provides an expectation of not only the next trial's payoff, which would set the duration of the pause to take, but also of the next stimuli the rat is likely to encounter. If the relevant stimulus is a vector of subjective reward intensity and opportunity cost, then the rat must have some mechanism for knowing the subjective reward intensity and opportunity cost that was in effect on trailing bracket trials on which it never worked, and therefore never obtained an estimate. Rats take the same, short pause at the start of leading bracket trials that follow trailing bracket trails on which the animal never pressed, suggesting that the rat

maintains a mapping of the (possibly updated) opportunity cost-intensity compound on one trial to that expected on the next. The pattern of errors imply that, rather than counting to three, the rats have stored a model consisting of three, two-trial sequences:

1) If current trial is high-intensity, low-opportunity cost ("leading-like"), next trial is variable intensity and opportunity cost ("test-like").

2) If current trial is low-intensity, low-opportunity cost ("trailing-like"), next trial is high-intensity, low-opportunity cost ("leading-like").

3) If current-trial is neither high-intensity/low-opportunity cost ("leading-like") nor low-intensity/low-opportunity cost ("trailing-like"), next trial is low-intensity, low-opportunity cost ("trailing-like").

The picture of the self-stimulating rat in the randomized-triads design is now considerably richer: over the course of training, rats form a world model of the progression of trials within a session, the world model provides the rat with an expectation for the subjective opportunity cost, subjective reward intensity, and payoff from self-stimulation to come, and the stimulus the rat uses to identify the next trial type is a vector comprising each subjective determinant of the decision to press. A world model of how trials change within a session, however, leads to the important question of whether rats develop a world model of the stability of the payoff within a trial.

6.2.3 World model of stability

Each trial presented in the randomized-triads design may differ to a varying extent from the trial that preceded it, but from the time the lever first extends into the cage following the flashing house light to the time it retracts and the house light flashes again, conditions within the trial are completely stable. If an animal is capable of developing an inference rule for what the payoff on the next trial is expected to be, then it is natural to ask whether an animal is capable of developing a model for what

the payoff from self-stimulation at any given point in the trial can be expected to be.

The pause made at the start of the test trial will reflect, to a certain extent, the payoff that can be expected. This is because this pause is different from that made at the start of trailing bracket trials, with unquestionably low payoff, but slightly longer than that made at the start of leading bracket trials, with unquestionably high payoff. The test trial, however, has a payoff that has been drawn at random from a list, so the duration of this post-priming pause cannot reflect the true payoff that can be realized throughout the entire trial. If the duration of the pause the rat takes before it begins to lever-press (either at the start of the trial or following lever-retraction) is in any way related to the payoff, two options are possible: either the pause gradually changes over multiple reward deliveries, or it changes abruptly following a sufficient number of exemplars. Moreover, if our molar model of time allocation can explain time allocation from each period of time between lever extension and lever retraction (either because a reward is delivered or because the trial has ended), then the proportion of time allocated to self-stimulation activities should also change as the estimated payoff changes.

Our data suggest that the patterns of post-priming (when the trial begins) and post-reinforcement (when the lever extends back into the chamber) pauses are many orders of magnitude more likely if we assume a step-wise rather than a gradual change. The pause following lever extension changes abruptly, in which case the animal's decision regarding how long to wait before lever-pressing would depend on a single-step update rule following a sufficient number of exemplars. That number is one, or nearly so: in most cases, the maximum-likelihood estimate of the number of reward deliveries necessary before the duration of the first pause switches from what it was at the start of the trial to what it will be at the end of the trial is a single reward.

Similarly, the change in time allocation is greatest between the time the an-

imal knows nothing about the payoff from self-stimulation to the time the payoff is revealed. Following the first reward delivery, time allocation values cease to change systematically. If the proportion of time the rat spends harvesting rewards is controlled by the payoff, and that time allocation ceases to change following the first reward delivery, it is unlikely that the animal continues to make meaningful revisions to its estimate of the payoff from self-stimulation. This is further corroborated by the change in the first pause the rat takes following lever extension: if the post-priming and post-reinforcement pauses are both related to the payoff the rat expects to receive from self-stimulation, the rat ceases to meaningfully update its estimate of that payoff as soon as it has obtained a single reward.

The results not only suggest that the period of time before the first reward delivery is special, but they also suggest that it is unreasonable to treat the well-trained self-stimulating rat as a model-free reinforcement learning agent. Were the rat to require re-learning the value of pressing on every test trial, the process would either produce incremental changes in pause durations and time allocations or the rat would have a high learning rate parameter for tuning those changes based on the experienced record of reward. Tuning the learning rate therefore requires some world model. If a world model of the stability of the trial is required to appropriately tune the learning rate so that it is very nearly one, learning rate tuning is subsumed by a world model that allows a feed-forward update to key decision variables as quickly as the model deems necessary.

It also appears implausible that the rat has "memorized" the large number of potential pairs of prices and reward intensities that it is likely to encounter (which, in the case of some rats, would be 126 combinations) in order to pick out the matching combination and implement the corresponding policy. Instead, the rat appears to use a model of how trials progress and a model of the stability of conditions within a trial to identify, as quickly as possible in principle, the key determinants of the decision to

press.

If the payoff itself is the stimulus that sets the appropriate policy to follow, then this very-liberally defined model-based reinforcement learning model is no different than one in which the payoff directly sets the probability of engaging in activities of varying length. Such a model is discussed below.

6.2.4 Payoff-based action selection

Model-based descriptions of the task involve a table-lookup process in which the total future net reward of a trial state—such as lever-pressing—is found within a list of stimuli. If a cue signals that the total future net reward will be high, the rat presented with the cue can look up the optimal policy to take without re-learning the contingencies for reward. The rat will begin to press with some expectation for what the total future net reward ought to be. Similarly, if a different cue signals that the total future net reward will be low, the rat will implement the policy it has already found to be optimal, rarely if ever sampling the lever for rewards.

The situation is slightly murkier when the stimulus itself is the payoff. If the payoff serves as an internal cue to signal that future rewards will be sufficiently valuable and can be acquired at sufficiently low cost to justify lever-pressing, the rat may still look up in a table the optimal policy to implement for a given payoff. At that point, though, the process of table lookup is no different than a process by which the payoff informs which action to take.

We have modelled (Chapter 5) the molecular-level action selection process as a "filling in" (that is, rapidly-updating) mechanism by which payoff directly provides the animal with what to do and, in so doing, for how long to do it. Rather than assuming the rat has associatively learned a pairing between a particular payoff and the optimal policy to implement, we have assumed that the payoff sets the policy by altering state-termination rates as a function of the payoff from competing activities.

If the payoff is high, it will drive the self-stimulating rat to select a post-priming pause of very short duration before starting to lever-press and will rarely, if ever, release to lever to engage in other activities. If the payoff is low, the rat will elect to do other things it is possible to do in the operant chamber before starting to lever press, and will often leave that lever-pressing state to resume the alternate activities that may be performed. In fact, this pattern often results in a censored true leisure bout: when the payoff is sufficiently low, the rat simply stops lever-pressing altogether and engages exclusively in other activities until the end of the trial.

The overall effect of this scheme is that the payoff sets the effective stay duration in post-reinforcement pauses, true leisure bouts and censored true leisure bouts, while the effective stay duration in holds and taps remains constant. Since animals in our hands engage only very rarely in uncensored true leisure bouts, the data imply that the rat usually consumes the fruits of leisure activities in a single continuous bout, during the post-reinforcement pause.

Furthermore, the fact that the effective stay duration in holds (which may be censored by reward delivery) and taps is payoff invariant implies that the same strategy—a mixture of short and long types of holds—is used no matter what the payoff from self-stimulation will be. When the price is low and intensity is high, all holds will be censored by lever retraction, and the maximum proportion of time allocated will be achieved: the ratio of the price to the sum of the price and the shortest post-reinforcement pause the rat can take. As the payoff decreases, the first factor to make any contribution to changes in time allocation will be the duration of the post-reinforcement pause, because many holds will continue to be censored by lever retraction. At sufficiently low payoff, the post-reinforcement pause will be very long, and if it terminates at all, the duration of the subsequent hold will not yield a reward. As it terminates, the high probability of engaging in a true leisure bout or quitting, rather than releasing the lever briefly as part of a tap, will drive time

allocation to a minimum value.

Conover et al. (2001a) reported a strong relationship between inter-response times and the programmed rate of reinforcement on a traditional, infinite-hold VI schedule of reinforcement. Unlike what they found, we observed that the relationship between tapping-related releases and payoff was weak (accounting for only 11% of the variance in short-release duration). In their experiment, inter-response intervals were composed of two components: just as in our modelling, dwell times in lever-release activities comprise a short- and a long-mean component. Unlike what was found in Chapter 5, the mean of the short-mean component of lever releases was dependent on the VI, which would make it payoff-dependent. The differences between the two procedures used is very likely the reason for our different findings. In an infinite-hold VI schedule, the first response after the lever is armed is rewarded, while in our cumulative handling time schedule, the animal is rewarded as soon as the cumulative time the lever has been depressed reaches the experimenter-set price. As a result, steady responding and steady holding strategies are differentially reinforced. A steadily-holding rat will not obtain many rewards under an infinite-hold VI schedule of reinforcement, as a reward will not be delivered until the lever is released and pressed after being armed. A steadily-tapping rat will obtain rewards at a lower rate under a cumulative handling time schedule of reinforcement, because every lever release increases the time to reward without providing much leisure benefit. In our hands, using a cumulative handling-time schedule of reinforcement, operant tempo varies only very little with payoff. The effect of increasing payoff is to both increase the probability that a lever release will be short, operant-related responding, and to decrease the duration of time for which the rat engages in activities that are unrelated to operant responding.

The many interacting components of the model—time spent in each activity, probability of quitting, probability of releasing the lever as part of a tap compared

to a true leisure bout—work in concert to produce curves that look like the Matching Law would predict on a molar level. In other words, "matching" is an emergent property that results from payoff-dependent sojourns in PRP, TLB, and quit activities and their transition probabilities to and from payoff-independent sojourns in hold and tap activities. The animal matches the relative time invested in an option to the relative rate of reinforcement because the payoff it provides sets the rate at which a particular action will be terminated. The probability-weighted combination of the set of actions determines the effective dwell times in activities that are directly observable. If the effective dwell time in work bouts (holds and taps) depends on the payoff from everything else, it is equivalent to stating that the rate at which work is left (the reciprocal of its expected dwell time) depends on the payoff from everything else, and vice-versa for dwell times in leisure bouts (PRPs, TLBs, and quits). High payoffs from competing activities bias the animal toward selecting actions that terminate at a high rate, while low payoffs from competing activities bias the animal toward selecting actions that terminate at a low rate. Matching occurs because a comparatively high VI from alternative A will result in a high leaving rate from alternative B, and a comparatively low VI from alternative B will result in a low leaving rate from alternative A, resulting in

$$\mathbb{E}[A]/\mathbb{E}[B] = f(\mathbb{E}[U_A])/f(\mathbb{E}[U_B]),$$

where $\mathbb{E}[A]$ is the effective expected dwell time pursuing alternative A, $\mathbb{E}[B]$ is the effective expected dwell time pursuing alternative B, $f(\mathbb{E}[U_A])$ is a function of the payoff expected from alternative A and $f(\mathbb{E}[U_B])$ is a function of the payoff expected from alternative B.

The model presented in Chapter 5 provides a remarkably good (though imperfect) account of single-operant performance in the randomized-triads design for

each test trial, from the time the first reward is delivered onward. However, it can, in principle, apply to leading bracket trials, for which the expected payoff is known to be high as the trial begins, as well as to trailing bracket trials, for which the expected payoff is known to be low as the trial begins. Given the evidence from Chapter 3 that the rat behaves as though it has a world model of the progression of trials in a triad, and the evidence from Chapter 4 that the rat behaves as though it has a world model of the stability of conditions within a trial, how does the rat select which action to do, and how long to do it, in the general framework of the randomized-triads design?

6.3 Putting it all together

Figure 6.1 provides a potential schematic of action selection in the randomized-triads design. The rat may be in one of three different trial types: a leading bracket, test, or trailing bracket trial. Data from Chapter 3 suggest that the rat maintains a representation of the last trial's expected subjective opportunity cost and intensity (which was potentially updated) to infer the next trial's expected subjective opportunity cost and intensity. If both are sufficiently similar to the trailing bracket trial, the rat directly infers it is currently in a leading bracket trial, a process learned through potentially reinforcement-learning mechanisms. Similarly, if both the subjective opportunity cost and reward intensity are sufficiently similar to the leading bracket trial, the rat infers it is currently in a test trial, and will begin to "explore" the mapping of lever-pressing to payoff. Finally, if the last trial's subjective opportunity cost and intensity are not similar to either bracket, the rat infers it is currently in a trailing bracket trial.

Following the exploration stage (in the case of test trials) or trial type inference (in the case of bracket trials), the rat uses the payoff expected from self-stimulation and the payoff expected from leisure activities to determine which activity to perform,

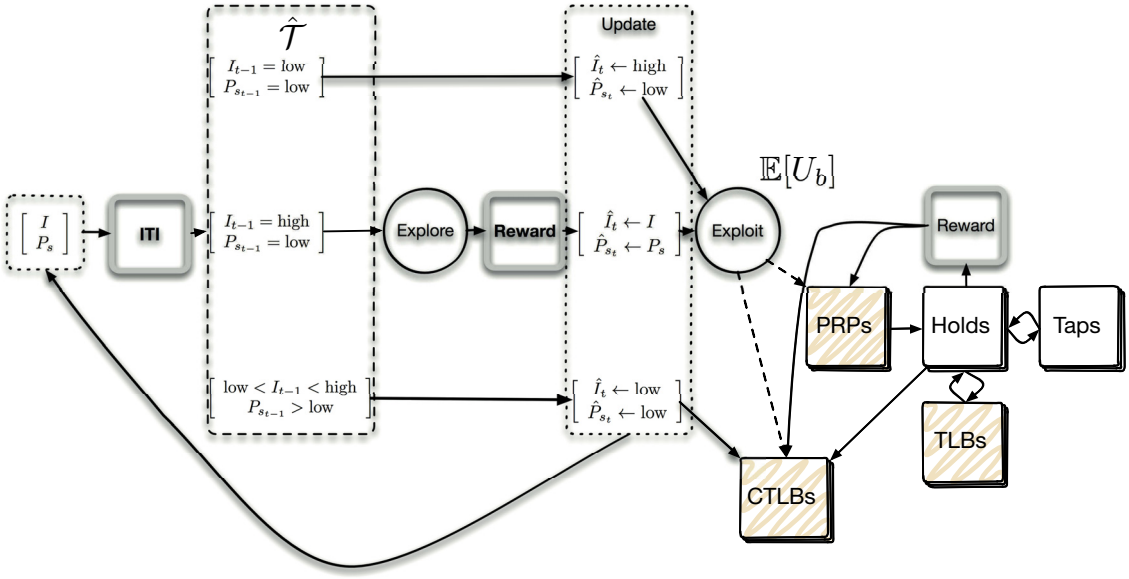

Figure 6.1. Action selection in the randomized-triads design. When an inter-trial interval begins, the trial state—a vector of the cached values of the subjective reward intensity (I) and opportunity cost (P_s)—is used as a signal to infer the subjective reward intensity and opportunity cost in effect on the next trial, according to state transition function $\hat{\mathcal{T}}$. If the trial state on the last trial was consistent with a trailing bracket trial, the rat can update its estimate of the trial state on the current trial to that of leading bracket trials and immediately exploit the rewards of self-stimulation. If the trial state on the last trial was consistent with a leading bracket trial, the rat must explore the mapping between lever-pressing and rewards and consequently update the elements that make up the state vector (subjective opportunity cost and reward intensity). If the trial state on the last trial was inconsistent with either bracket type, the rat can update its estimate of the trial state on the current trial to that of a trailing bracket trial and immediately exploit the rewards of non-self stimulation activities. During the exploitation stage, the payoff from self-stimulation ($\mathbb{E}[U_b]$) sets the balance of time spent pursuing self-stimulation and non-self stimulation activities. Activities that reflect pursuit of brain stimulation rewards are white (holds, taps), and those that reflect pursuit of extraneous rewards are shaded. Experimenter-enforced events, such as the inter-trial interval (ITI) and reward delivery are indicated in grey-outlined boxes, while dashed boxes indicate "cognitive" operations and circles represent trial phases.

and for how long to perform it, by setting the termination rate on these actions. In tandem, the current estimates of the subjective opportunity cost and reward intensity of the electrical reward on offer are updated, if necessary. When the rat never worked—as is the case on most trailing bracket trials—the estimates of the subjective opportunity cost and reward intensity are never updated from the predicted value, and the next trial is inferred on the basis of these estimates, rather than discovered values.

The exploration stage—assuming the rat explores the mapping between lever-pressing and payoff on that trial—appears quantitatively different from the exploitation stage. Once the mapping is known, because the rat has a world model of the structure of the trial which may have developed over the many months it is trained, there is no need for the estimated payoff to be revised again. The post-priming pause occurring at the very beginning of the test trial—when the mapping between lever-pressing and the payoff that will be delivered is as-yet unknown—is only slightly longer than that on leading trials. Since the payoff that can be obtained on test trials is, on average, intermediate, the short post-priming pause at the onset of test trials might reflect a separate process that is usually at odds with self-stimulation but is aligned with exploitation at the start of test trials. When the payoff from self-stimulation is known, the rat may infrequently leave lever-pressing to explore whether the payoff from everything else has changed, or resume lever-pressing from a protracted leisure bout to explore whether the payoff from self-stimulation has changed. In this case, the goal of exploitation—devoting one's time to the option with the better payoff—is at odds with the goal of exploration—investigating whether and in what way the payoff from various activities has changed. However, at the very beginning of a test trial, the rat does not yet know what the mapping between lever-pressing and reward will be. In this special case, the goal of exploitation—devoting one's time to a single option with the presumed best payoff—is aligned with the goal of exploration—investigating

whether and in what way the payoff from each option has changed. Once the payoff from self-stimulation is uncovered, exploration and exploitation are, once again, at odds.

6.3.1 The counter-factual payoff hypothesis

The model presented here is very similar to one proposed by Gallistel et al. (2001). In it, Gallistel et al. propose that the rat extracts the current rate of reinforcement of each option from a small sample of intervals. The scalar combination of reinforcement rate with other key determinants (subjective intensity, for example) provides the income from each side, the ratio of which sets the ratio of the means of the two exponential processes that determine how long to pursue each goal.

$$\mathbb{E}[A]/\mathbb{E}[B] = U_A/U_B$$

They further propose a linear relationship between the sum of the leaving rates from pursuit of each goal and the sum of the experienced incomes that can be derived from each goal.

$$1/\mathbb{E}[A] + 1/\mathbb{E}[B] = b + m(U_A + U_B)$$

Those two constraints—that the ratio of expected sojourns at each option be equal to the ratio of incomes and that the sum of the reciprocal of the expected sojourns at each option be a linear function of the sum of the incomes —result in the prediction that the reciprocal of the expected dwell time in pursuit of an option will be directly proportional to the income from the unchosen option and directly proportional to the ratio of income from the unchosen option to the sum of the incomes from both:

$$1/\mathbb{E}[A] = mU_B + b\frac{U_A}{U_A + U_B}.$$

This description is similar to our view that the rat in the randomized-triads design obtains a single exemplar of what the payoff from self-stimulation will be on a test trial, which then sets the effective expected time the rat will spend in leisure-related activities. As the payoff from non-self stimulation activities does not change, the effective expected time the rat will spend in work-related activities will also not change. Unlike the Gallistellian model, however, the time spent in pursuit of the rewards derived from leisure does not depend on the sum of the incomes from both pursuit of leisure rewards and pursuit of brain stimulation rewards. If that were the case, the effective expected time the rat will spend in work-related activities would also be payoff-dependent. Instead, what is proposed here is that the rat will select a time to perform an action as a function of the payoff that can be derived from the counter-factual (we call this the counter-factual payoff hypothesis). When the next best thing the animal can do is associated with a high payoff, we propose that the animal will select an action that can be quickly completed and terminate that action quickly. When the next best action is associated with a low payoff, we propose that the animal will select an action that will take longer to complete, and engage in that activity at a more leisurely pace. For example, when the payoff from self-stimulation is maximal, the rat may rush to the lever (if it is not already there) as quickly as possible. When the payoff is negligible, the rat may opt to groom, an action that will take considerably longer, and only lever press after it has been grooming for some time.

The hypothesis is testable in principle. If, on a small subset of probe reward encounters, the lever is not retracted and the reward is not delivered, it should be possible to identify how long the rat is willing to work uncensored by lever retraction. If the counter-factual payoff hypothesis is correct, there should be no difference, on these probe reward encounters, in the duration for which rats will hold the lever across all brain stimulation payoff conditions, because the termination rate is maintained

by the payoff from everything else. In contrast, background stimulation, delivered on a variable-time schedule contingent on the lever being released, would increase the payoff from everything else. In this case, while the termination rate of holding on probe reward encounters would be a function of whether or not there was background stimulation, the time spent in post-reinforcement pause activity would not.

6.3.2 Learning the rules

In the randomized-triads design, the animal only needs a single-reward sample to identify the subjective opportunity cost and reward intensity on offer during the test trial phase. On bracket trials, the rat knows these two determinants—as well as their payoff—in principle as soon as the inter-trial interval begins. The temporal dynamics of estimating the subjective opportunity cost are, in our case, moot, as the rat's internalized model of the stability of the trial allow the animal to update this quantity in a single step. The data provided by Gallistel et al. (2001) provide some independent evidence that this process occurs as quickly as an ideal detector would allow. The question, then, is how the world models of stability and change arise through training.

The evolution of the world model that involves a comparison between the last trial's subjective opportunity cost and intensity and those they reliably predict may be a very slow process indeed. Unpublished data demonstrate that the time course over which post-priming pauses become stable—that is, reliably short on leading bracket, intermediate on test, and reliably censored on trailing bracket trials—is on the order of weeks, a slow process that may indeed involve traditional reinforcement learning mechanisms. Given that rats appear to have no problem acting as quickly as an ideal detector of changes in reinforcement rate, it is entirely plausible that rats may have internalized some form of a hidden Markov model (HMM) of how trials lead to each other. In the HMM framework, the subjective opportunity cost and intensity

provide observable signals of the type of trial the animal is in (a hidden variable). The task for the rat is to identify the most likely next hidden variable—the next trial type—on the basis of the observable signals it has observed and the current signals it has encountered. When it is first placed in the randomized-triads design, the rat has no way to know there will be, essentially, three different unsignalled trial types, and no way to know the mapping between one trial type and the next. Some process must occur for the rat to identify the existence of statistical regularities inherent in the observable symbols presented, identify the number of hidden trial types that generated those observable symbols, and identify the mapping of hidden trial types to each other. Similarly, some process must occur for the rat to identify the stationarity of the subjective opportunity cost, reward intensity and probability of reinforcement throughout the trial in the face of a possibly noisy evaluative system.

A natural framework for studying these would be Bayesian (Deneve et al., 1999; Knill and Pouget, 2004; Beck et al., 2008), according to which the animal's model of the world changes in complexity as a function of the quality of the predictions that can be made (thereby revising prior probabilities, numbers of states, etc.) without necessarily referencing a reward prediction error *per-se*. The rat is trained, from the time it is screened for the effectiveness of the self-stimulation electrode, that conditions between the time the cue signalling the start of a trial is presented to the time the cue re-appears will be stable. This must place a high prior on the probability that the subjective opportunity cost and reward intensity will be the same as the last. Following training, rats are presented with the repeating pattern of trial triads for weeks, and testing formally begins when responding is reliably high on leading bracket and reliably low on trailing bracket trials. As a result of this training, it is quite possible that animals have also placed a high prior probability on there being three hidden states of the world, and on the permutation-like mapping of one hidden trial type to the next.

It is now possible to address a question that will be of interest to the general neuroscience community: how can this simplified schematic of how the rat approaches the action selection problem in the randomized-triads design be implemented in neural circuitry? How does the brain solve the action selection problem as outlined here?

6.4 Stages of processing

The rat behaves as though it has a world model of the triad structure of the randomized-triads design, a world model of the stability of trial conditions, and a behavioural allocation strategy that depends on the payoff from self-stimulation (in the case of leisure-related activities) or the payoff from everything else (in the case of work-related activities). These processes must have some basis in neural machinery.

There must already be processes in place to perform the translation of objective determinants (e.g. pulse frequency, price, force, probability) into subjective determinants (e.g. intensity, opportunity cost, effort cost and risk). We hypothesize three stages of processing, above and beyond those necessary for the above psychophysical mappings : a world model-generating process, a process by which the expected outcome of various actions are rapidly updated, and a process by which the updated map of actions to outcomes results in the behavioural policy. A world model generating process is necessary for the rat to use the cued inter-trial interval and unsignalled but available intensity and opportunity cost information in inferring the expected payoff and, potentially, its variability. Such a process is also necessary for the rat to explore the mapping between lever-pressing and payoff only until the first reward has been delivered. That expectation must, in some way, be tied to the saliency or desirability of the actions to be performed in order for it to have any influence on performance. Finally, the updated payoff from self-stimulation must be part of the process by which actions are selected.

One process translates the strength of the stimulation—the directly observable and manipulable "reward strength"—into its subjective impact. Previous work has shown that the activity of directly activated neurons is spatio-temporally integrated by a network, or a system of networks, the peak activity of which is translated into a subjective intensity which endures as an engram to direct future behaviour (Gallistel et al., 1974). Subjective intensity is related to pulse frequency and train duration by logistic and hyperbolic functions, respectively (Sonnenschein et al., 2003; Gallistel, 1978; Simmons and Gallistel, 1994). Similarly, work by Solomon et al. (2007) has shown that the subjective impact of price increases is roughly linear at high prices, and rolls off to a minimum subjective opportunity cost at sufficiently low prices. The data provided in Chapter 2 suggest that the psychophysical translation in evaluating the subjective risk of an option is linear.

As each of these psychophysical mappings provides independent information regarding the payoff that can be expected from lever-pressing, each must inform either directly or indirectly (via their scalar combination) the process involved in mapping the relationship between response and outcome. In the case of bracket trials, for which the animal need not (and does not) engage in an exploration phase, the world model directly provides an expected payoff. As a result, the world model must either maintain a representation of the identity of the last trial and its mapping to the next, as well as the expected payoff from the next trial, or it must maintain representations of the subjective impacts of the last trial's BSR and price and their mapping to the next. According to the former, the vague stimuli "like a leading bracket", "like a trailing bracket" and "unlike leading or bracket" predict the next vague stimulus to come when the inter-trial interval is begun as well as the payoff on that trial, which would require no further updating. According to the latter, the stimulus-vectors "high intensity/low cost", "low intensity/low cost" and "neither high/low nor low/low" predict the next stimulus-vector, which can then be multiplicatively

combined to inform the rat of the payoff. The rat can then use its world model of stability in conjunction with the stored record of reward, opportunity cost, and probability to gauge the degree to which some actions ought to be selected. In other words, the subjective opportunity cost accrued over the reward encounter and the reward intensity delivered upon successful completion, combined multiplicatively with the risk that has been associated with the lever over many trials, provide the animal with the mapping between lever-pressing and payoff. The mapping can thereby influence which actions are selected and which will be neglected, and need not be updated as the animal collects rewards.

When the animal has an expectation of the payoffs from self-stimulation (U_b) and those from extraneous activities (U_e), these expectations must inform circuitry that implements the action selection problem. One simple way is for the payoffs to drive populations of neurons that collectively represent pursuit of different goals. Each population inhibits neighbouring populations, providing an on-centre, off-surround population coding scheme. Such a scheme is presented in figure 6.2. When the payoff from self-stimulation is comparatively high, the increased activity inhibits pursuit of all other goals. When the payoff from self-stimulation is comparatively low, populations involved in grooming, resting and exploring "win" the competition and make it less likely the rat enters into hidden hold and tap behavioural states.

Given the constraints posed by the animal's behaviour, how might the brain implement the strategies we describe here? How may we test whether it does?

6.4.1 World models

Any neural implementation of the action selection problem we have described would need to incorporate the great influence that world models provide. For the rat to have a sense of "trailing is followed by leading", for example, there must be a population that maintains a representation of what the subjective determinants of

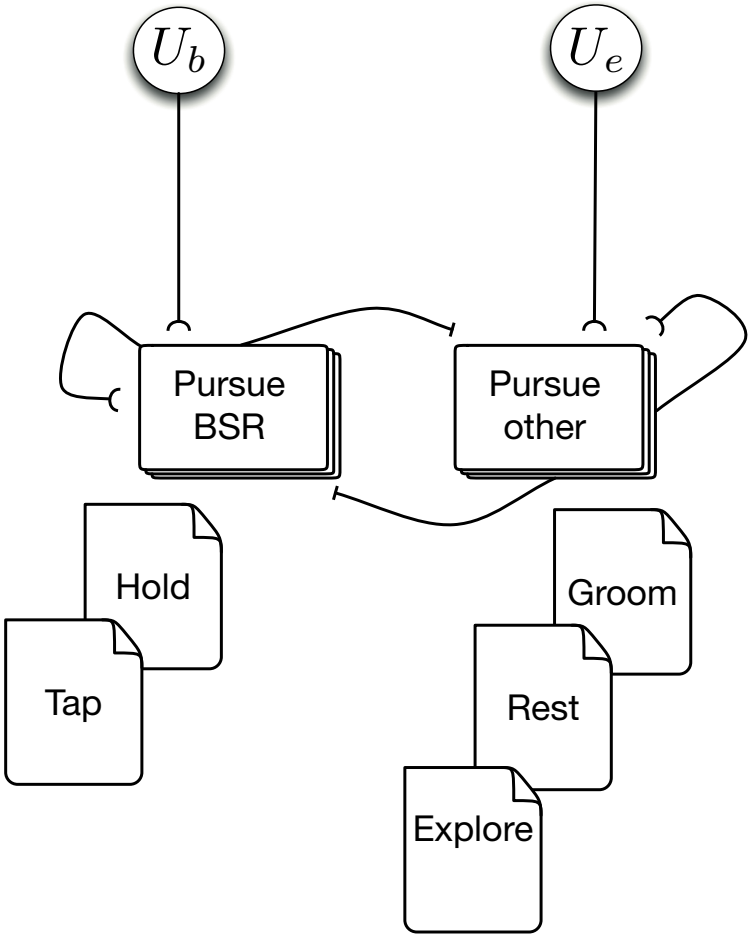

Figure 6.2. Implementation of the action selection problem. To implement the action selection problem, one may envision multiple populations of neurons responsible for pursuit of various goals. Using lateral inhibition and excitatory feedback, each population acts as an on-centre, off-surround unit whereby increases to the payoff of one goal compared to others will increase the probability of engaging in activities related to the higher-payoff goal and decrease the probability of engaging in activities related to other goals. For example, an increase in U_b compared to U_e will shut down, by lateral inhibition, the probability that the rat grooms, rests and explores, increasing the total amount of time spent holding and tapping.

the decision to press were on the last trial, a mapping that provides the expected values of these subjective determinants to those in effect on the previous trial, and a process that updates these values in the face of new information. The inter-trial interval provides a potent cue that conditions will be different; how they will differ will depend on the mapping of stimuli to each other, and their motivational impact will require the scalar combination of the stimuli with each other.

Since the exploration stage during test trials (the period of time before a reward is delivered) appears to differ in terms of the duration of the pause to make, the maximum duration of the responses made, and the overall proportion of time allocated to lever-pressing, the motivational impact of the key determinants of decision may not be the only process at work. The rat may "know," in a sense, that the mapping between responding and payoff will need to be updated on the test trial. Thus, the mapping between subjective opportunity cost and reward intensity from one trial to the next may involve a representation of the variance that can be expected in these variables, as well as their mean.

One region that would be well-suited to representing the world model would be orbitofrontal cortex. Orbitofrontal cortex has been involved in flavour-based unblocking (McDannald et al., 2011), temporal discounting (Roesch and Olson, 2005), reversal learning (Rudebeck and Murray, 2008) and probabilistic learning (Roitman and Roitman, 2010). Each of these is, essentially, a higher-order model of contingencies and rewards: higher-order representations of identity (flavour-based unblocking), of when a reward will come (temporal discounting), of changing task demands (reversal learning), and of relations between stimuli (risk-based discounting). Indeed, it has even been suggested that a fundamental value signal (Padoa-Schioppa and Assad, 2006) is represented in orbitofrontal cortex. There is certainly evidence that human orbitofrontal cortex is differentially activated by reward and non-reward outcomes (Knutson et al., 2001) and that this anticipatory neural activity is correlated with

subsequent behaviour. It is quite possible that the orbitofrontal cortex maintains model-based information specific to the task. The world model may be arbitrarily simple, as is the case when behaving as though one had asked "was the last trial like a trailing bracket, because then the next one will be a leading bracket trial." It may also be highly organized, as is the case when behaving as though one had asked "what is the sequence that guarantees my opponent a checkmate in three moves and what steps can I take to prevent it?"

The hypothesis that model-based information is represented in sensory and association cortical structures and is relayed to sub-cortical evaluative mechanisms may be difficult to directly test, but it is not impossible to test some of the predictions that arise from both the behavioural theory and its proposed neural manifestation. Specifically, any region proposed to provide representations related to the world model for a task would require an evidence-based modulation in firing rate. Acquisition of the world model itself would likely involve a fairly slow-changing, synaptic-weights based system. Acting on the basis of that world model, however, can occur quickly when it is in place.

6.4.2 Expected payoff

The world model will provide essential information about the expected outcome of a lever press: if the current trial follows a trial resembling a trailing bracket, the payoff will be high, if the current trial follows a trial resembling a leading bracket, the payoff is variable and must therefore be uncovered, and if the current trial follows a trial resembling neither, the payoff will be low. The combination of stimuli represented that signal a change in conditions must inform a process that evaluates the expected mapping between each action and the payoff it can be expected to provide. This process would require stimuli—even those internal to the rat, such as subjective opportunity cost and reward intensity—to update the mapping between an action

and its desirability, which can then provide a sort of saliency signal to the process that selects actions.

One means by which the binding of outcomes to actions could occur would involve the ventral striatum, a major input region of the basal ganglia system that receives convergent cortical inputs (Nakano et al., 1999) and dopaminergic inputs (Voorn et al., 1986) from the ventral tegmental area (Beckstead et al., 1979; Ikemoto, 2007). Computational models of the basal ganglia (Chakravarthy et al., 2010) place the ventral striatum at a critical point in the evaluative process: it receives inputs from sensorimotor cortex and projects to different regions of caudate and putamen based on cortical topography (Berendse et al., 1992). The striatum is therefore superbly placed anatomically to bind a variety of actions to the degree to which they ought to be selected—in other words, their salience. Moreover, medium spiny neurons of the ventral striatum have "up" and "down" states (Plenz and Kitai, 1998), regulated in part by dopamine influx, that could serve as a filter to weaken inputs that are already low. Furthermore, local inhibitory connections (Lighthall and Kitai, 1983) and cholinergic tonically-active neurons (Anderson, 1978; Wilson et al., 1990) would also improve the signal-to-noise ratio in ventral striatal representations. One computational hypothesis of striatal activity (Chakravarthy et al., 2010) is that it represents the salience of requests from cortical structures to access the motor system. We propose here that the process that updates the mapping between goals and outcomes may occur at the level of the ventral striatum, using model-based information from cortical pathways, and saliency-based information from tegmental regions.

van der Meer and Redish (2009) recorded from neural ensembles within the hippocampus and the ventral striatum while rats navigated a multiple-T maze. Following two low-cost choice points (concatenated T's), the rat would approach a final choice point with high cost: one arm was baited along the return rail, while the other was not. If the rat made the incorrect choice, it would have to re-start the

maze anew without a food reward. Various hippocampal cells were active while the animal was in a particular place, and thus, one could decode the position the rat was representing on the basis of the population's activity. Similarly, various cells of the ventral striatum were active when the click of the food dispenser was sounded. One could decode, in principle, the reward the rat was representing on the basis of the population's activity. The maze changed configuration every day, and thus, early trials in an experimental session provide the rat with a mapping between actions (left, right) and outcomes (reward, no reward). In these early trials, rats will often pause at the final choice point, seemingly deliberating the correct option. While the rat was immobile at the choice point, early in the day's experimental session, hippocampal cells were reliably activated in a sequence, which when decoded proceeded as if the rat had walked down one arm and the other. Similarly, ventral striatum cells were reliably activated at the final choice point as well as at reward sites, but only while the animal was updating the map of where to go. These findings are certainly consistent with a payoff-updating role in ventral striatum. Were the ventral striatum involved in representing the actual payoff from taking an action, its activity would not be tied to the short (fewer than 10 laps) period of time the mapping is updated.

Lesion studies also provide some indication that the ventral striatum is involved in updating the goal-saliency mapping. In the blocking procedure, a conditioned stimulus (CS) is paired with an unconditioned stimulus (US) until it reliably elicits a conditioned response (CR). Following subsequent pairings of a compound stimulus comprising the CS and a new, blocked stimulus (BS) with the US, presentations of the BS alone do not elicit any CR. This is because the unconditioned stimulus is already predicted by the CS by the time the BS is presented, and therefore, no new learning occurs. In the unblocking procedure, the US paired with the CS-BS compound is different (either in quantity, for value unblocking, or in quality, for identity unblocking) from the original US that was paired with the CS. As a result, the animal

learns that although CS predicts a particular US, that CS when the BS is present predicts a different US. When using a model-free strategy, the rat can learn that the compound predicts more food but not that it predicts a different flavour of food. When using a model-based strategy, the rat can learn both. McDannald et al. (2011) found that NMDA lesions to ventral striatum impaired a rat's ability to unblock value (learning that the compound predicts more food) or to unblock identity (learning that the compound predicts a different flavour), implying an impairment in their use of either a model-free or model-based learning strategy. One would expect that impairment of a system that updates the mapping between goals and their salience would impair both model-free (strictly "amount" learning) and model-based ("amount and identity" learning) strategies.

Human functional neuroimaging studies have shown a role of dorsal striatum in the maintenance of the outcomes of actions so that they may be selected more frequently (O'Doherty et al., 2004). In reward trials of the instrumental task, human subjects had to choose between two stimuli, predicting a high probability of juice reward (60% chance) or low probability of reward (30%). On neutral trials, they had to choose between high and low probabilities of obtaining a neutral solution. In this instrumental condition, one would expect both a process that updates the mapping between actions and outcomes and a process that informs action selection of the updated mapping. On a separate Pavlovian task, stimuli were presented passively along with the reward, in yoked fashion, to individuals who simply had to identify which of the two stimuli the computer had chosen. In this Pavlovian condition, one would expect a process that updates the mapping of outcomes to still operate, but this mapping need not inform action. While activity in ventral striatum was correlated with a "prediction error" derived from reinforcement learning principles in both tasks, the dorsal striatum was correlated with the prediction error only in the instrumental task. Given that the "prediction error" is loosely related (though not identical) to

changes in the payoff from each option, it is interesting that dorsal striatum showed activation only when the updated mapping to outcomes needs to inform which action to take.

Another experiment that supports the idea of a dorsal striatum-based action selector comes from work by van der Meer et al. (2010). As previously described, rats must complete three low-cost sequential choices along three concatenated T-mazes. The final choice point will lead to either a reward or non-reward, at which point the rat returns to the start of the maze. If the rat has chosen incorrectly at the final choice point, it will have to re-navigate the entire maze anew. Unlike reward-responsive neurons of the ventral striatum, whose activity at the final choice point becomes less important as the maze is learned, neurons of the dorsal striatum became more efficient at encoding the sequence of turns in the maze (and at not encoding task-irrelevant portions of the maze) as the rat made more laps. If the dorsal striatum is involved in using gated ventral striatum action-salience signals to select actions on the basis of a noisy competition, those neurons that are most active will necessarily gain behavioural control in the decision-rich portions of the task—that is, during the sequence of turns in the maze—as those sub-populations are slowly updated with respect to the degree to which the animal ought to choose them.

Lesions to the dorsal striatum prevent rats from using putative stimulus-response associations that have been acquired over the course of training in a single T-maze task (Packard and McGaugh, 1996). When navigating a single T maze, the rat may have learned either or both of two contingencies: the reward is east of the choice point (a stimulus-stimulus strategy), or the reward is right of the choice point (a stimulus-response strategy). The rat can use either of the two strategies, which can be assessed simply by rotating the maze 180 degrees. After 8 days of training, rats predominantly use a stimulus-stimulus strategy, implying a model-based, goal-directed process: a delicious food pellet lurks east of the current location. After

16 days, rats predominantly use a stimulus-response strategy, implying a model-free, habit-directed process: turning right is a good thing, and one need not know why. Extensively trained rats (and, thus, rats using a model-free, habit-directed stimulus-response process) with functionally lesioned dorsal striata began to use a stimulus-stimulus, model-based, goal-directed strategy instead. This implies that the goal-directed, stimulus-stimulus strategy is not *lost* over training, but rather, that other, faster model-free pathways wrest behavioural control from slower model-based solutions when feed-forward representations are unnecessary. These data suggest that there may be multiple systems vying for control over behaviour. Indeed, it has been proposed (White and McDonald, 2002) that the dorsal striatum, amygdala and hippocampus subserve interdependent memory systems. Damage to the hippocampal system disrupts performance on the Morris water maze (Morris et al., 1982), while interference with normal dorsal striatum function McDonald and White (1993) disrupts cued radial maze learning, and amygdaloid lesions disrupt the expression of innate fear (Blanchard and Blanchard, 1972) and fear conditioning (Hitchcock and Davis, 1986). It is quite possible that learning how things are related to each other (hippocampus), which responses should be made under different conditions (dorsal striatum) and which stimuli are of any interest (amygdala) would be processed in different regions and would gain access to the motor system by different means.

Optogenetic methods promise to provide an important key to testing whether the ventral striatum is indeed involved in the rapid updating process we describe, and whether the rapid-updating propagates to dorsal striatum sub-populations. These methods involve the insertion of genes encoding a light-sensitive protein into cells that can be targeted with respect to the neurotransmitter released, projection area, or somatic origin. The question of rapid mapping updating is therefore a straightforward one: if one inhibits activity within the nucleus accumbens that is related to updating the payoff from lever-pressing while the first brain stimulation reward of a test trial is

delivered, one would expect to lengthen the exploration phase to two rewards rather than one. In other words, shutting off the area specifically responsible for the updating process while it is in progress should prevent that process from occurring. As a result, the animal would base its actions following the first reward delivery on the same set of saliencies it had before the first reward was delivered. If every reward is accompanied by a selective silencing of the neurons responsible for updating the mapping between action and payoff, although a new value of intensity and opportunity cost can be stored and used to predict the next trial, the animal will not have used that information during the trial to guide its behaviour. On a test trial for which stimulation is too weak to normally support lever pressing, and for which the subjective opportunity cost is also low, continually silencing the process that updates the mapping between response and payoff would presumably maintain the rat in the "exploration" phase rather than allow it to enter the "exploitation" phase. As other processes, like those storing records of the subjective intensity and opportunity cost, would not be affected by the mapping between response and outcome, the next expected trial would be a leading trial (as the intensity and price are low). If the silencing in some way enhances the reward, thereby dramatically reducing the post-reinforcement pause on these low-payoff trials to the duration of the post-priming pause, then the next expected trial would be a test trial. The critical tests would require some very sophisticated technical prowess, but the general behavioural methodology is easily transferable from findings we have already reported in Chapter 3.

6.4.3 Payoff-based selection

Provided payoff sets the probability of selecting an action and the duration for which the animal is engaged in that action, some mechanism must subserve the process. The previous section has provided hints about the location where the saliency of a mapping between response and payoff could be represented and dynamically

updated; below, we discuss what substrates may be involved in arbitrating between competing actions available.

On our view, the process of action selection arises from noisy competition between elements representing actions and the payoffs that may be derived from their pursuit. Neural sub-populations would therefore have to encode what to do, the activity of which would be dependent on the payoff from competing goals. Such a scheme is not necessarily difficult to entrust to sub-populations of neurons with local inhibitory synapses and external diffuse excitatory synapses. The result would be an on-centre, off-surround type of sub-population encoding, whereby the selected action is that which is "on" and those that are not are "off." An alternative view is to assume that action selection emerges from reward delivery driving covariance between reward-related circuitry and choice-related circuitry (Neiman and Loewenstein, 2013). These models can, in fact, account for matching in the traditional infinite-hold variable interval schedule. However, it is unclear whether a covariance-based synaptic plasticity rule can account for the pattern of behaviour seen under the cumulative handling time schedule, given that it requires the rat to accumulate subjective estimates of opportunity cost.

A simple alternative is to assume that the rat maintains an engram corresponding to the subjective intensity of the rewarding stimulation, the opportunity cost, risk, effort, and waiting time involved in acquiring it. We propose that the final decision variable, that which would modulate which actions are salient and which ought to be neglected, is the scalar combination of these key subjective determinants. Under the influence of a high payoff, a sub-population of cortical neurons representing the higher-order goal of "acquiring rewards" would be selectively enhanced, and in so doing, sub-populations representing "groom" and "rest" would be inhibited by local inhibitory connections. The enhanced goal would propagate to the level of individual motor programs like "approach lever" and "hold lever down." Any action unrelated

to the pursuit of the reward would be terminated, because lateral inhibition from BSR-related sub-populations would have terminated it, either directly or upstream.

If actions are selected on the basis of payoff and result from a competition between all motor responses possible, then inducing activity within the sub-population responsible for representing lever-pressing actions will both (1) reduce the duration of alternate responses and (2) increase the duration of time for which the action is selected. In fact, one would expect the behavioural response to be tightly linked to both ongoing activity and the trial phase. During the exploration phase, before the rat knows what the actual payoff will be, the activity of neurons related to lever-pressing activities would be expected to track both the action that is chosen and the decreasing payoff of lever-pressing as the rat holds the lever down. In the exploitation phase, the activity of neurons related to lever-pressing activities would be expected to track only the action that is chosen. It remains to be seen whether or not this is true, but advances in ensemble recording techniques and optogenetic methods make these hypotheses empirically verifiable.

6.4.4 The MFB at the centre of it all

If sensory and association cortices, especially orbitofrontal cortex, are involved in model-based learning, the striatum is involved in updating and maintaining the mapping between the payoff and response, and sensorimotor cortices are involved in the payoff-based selection of which action to take, where does one place the set of neurons excited by the electrode implanted in the medial forebrain bundle?

Electrical stimulation can compete and summate with (Conover and Shizgal, 1994), and substitute for (Green and Rachlin, 1991) natural rewards like sucrose, food, and water. It must therefore carry a signal, either directly or soon thereafter, that is commensurate with all these stimuli. The organizing principle we have used is to assume that this signal provides multimodal information about the underlying

rewarding nature of a stimulus, that a leaky integration of this brief signal over its spatial and temporal extent is conducted downstream, and that the peak activity of the integration network is committed to memory. Animals working for brain stimulation rewards respond consistently to variable and fixed schedules of reinforcement. In order to perform an expectation over the intervals one has seen, those intervals must have been represented somewhere, especially in the cumulative handling time schedule which requires a reasonable estimate of the time spent at the lever to be accumulated. It is likely that a similar representation of the subjective opportunity is committed to memory, and possibly any other key subjective determinant of the decision to press. These representations would have to be combined somewhere—possibly in dorsal striatum, and possibly following an updating process in ventral striatum—in order to inform the rat of the mapping between lever-pressing and its expected payoff on a given trial.

The medial forebrain bundle is therefore at the very heart of this action selection system. Although stimulation via macroelectrodes does not have the specificity necessary for determining which of the many dozens (Nieuwenhuys et al., 1982) of fibre tracts coursing past the electrode tip are responsible for reward, much progress has been made in deriving their characteristics. The properties of these neurons have been behaviourally derived: the direction of conduction of at least a subset of the neurons is likely anterior-posterior (Shizgal et al., 1980), their absolute refractory period is short (Yeomans, 1979), and can follow pulse frequencies of up to roughly 400Hz (Solomon et al., 2010). If one could identify the neurons responsible for the rewarding effect, it would greatly understand where each subsequent stage of processing occurs and how that information is processed, including where world models may be represented and how they evolve, where action-outcome mappings are updated and maintained, where and how those mappings influence action selection. Techniques that improve the spatial and temporal selectivity of a causal manipulation will likely

provide the tools necessary for elucidating these pathways, while new protocols and analysis methods, such as those presented in Chapters 2 through 5, could provide a strong quantitative basis for evaluating theories of action selection.

6.5 Conclusions

Psychophysical methods have tremendous power to uncover the processes and biophysical mechanisms at work in a large number of applications. For example, painstaking work by Hecht et al. (1942) demonstrated that the retina was capable of detecting on the order of 5 to 8 quanta of light. Were it not for carefully crafted experiments, using the most sophisticated equipment available, it would have been impossible to assess the incredible degree of sensitivity in the visual system. These psychophysical data inspired a great deal of subsequent molecular and physiological work regarding how vision is implemented in the brain. Their findings narrowed the potential chemicals responsible for light transduction, and the biochemical cascade that allows such minute quantities of light to result in visual perceptions. Psychophysical methods provide the crucial information to direct and inspire molecular, cellular, and systems neuroscience.

The same can be said of action selection. Crude methods, like response-rates and the degree of a consummatory response, have been fairly good at detecting the effect of manipulations on a gross level. Indeed, it was the percentage of time spent responding that was originally used in demonstrating the rewarding effect of septal stimulation (Olds and Milner, 1954). Without even an insensitive measure of choice, it would not have been possible to assess where stimulation was effective at reinforcing an action and where it was not. However, as the various processes that govern choice are parsed, the methods for measuring them must also be refined. Seminal work by Hodos and Valenstein (1962) showed that response rates alone were incapable

of discriminating between highly rewarding stimulation that produced motoric side effects that competing with lever-pressing and weakly rewarding stimulation. By measuring the threshold stimulation strength at which animals would lever-press at some criterion rate, it was possible to determine whether a manipulation had altered motivation, in some way, or whether it had altered the motor capacity to respond. The curve-shift method (Miliaressis et al., 1986) has since become the dominant paradigm in assessing the effects of manipulations to rewarding brain stimulation, but it, too, is incapable of a distinction: those manipulations that change the effectiveness of the stimulation to induce reward from those that change other key determinants of decision-making. This thesis presents a way to disambiguate between the two, by assuming that intensity is evaluated separately from opportunity cost, via the Shizgal Reinforcement Mountain model. This molar model of choice is, despite its great usefulness, not a model of individual action selection. This thesis presents a molecular model of choice, based on the idea that animals do not simply make stimulus-response associations that must be re-learned when conditions change. The organism stands between stimulus and response, and has likely extracted statistical properties of its environment to better act upon regularities that are apparent to it. The great challenge is to accurately describe what the animal is doing, in real time, along with how it accomplishes it.

The methods used to infer the processes that underlie action selection will necessarily need to grow in complexity as our understanding of the decision-making process becomes more complex. New methodologies will always allow us to probe deeper and more insightfully into the neural organization of choice. This thesis demonstrates that along with the increased power of new technologies, the behavioural and statistical methods we use to relate neural findings to behaviour can—and will have to—complement the neurobiological sophistication of emerging genetic and recording methods.

References

Ainslie, G. and Herrnstein, R. J. (1981). Preference reversal and delayed reinforcement. *Learning & Behavior*, 9(4):476–482.

Akaike, H. (1976). An Information Criterion. *Math Sci*, 14(153):5–9.

Allais, M. (1953). L'extension des theories de l'equilibre economique general et du rendement social au cas du risque. *Econometrica, Journal of the Econometric Society*, pages 269–290.

Allis, V. L. (1994). *Searching for solutions in games and artificial intelligence*. PhD thesis, Rijksuniversiteit Limburg te Maastricht.

Allison, J. (1983). *Behavioral economics*. Praeger New York.

Anderson, M. E. (1978). Discharge patterns of basal ganglia neurons during active maintenance of postural stability and adjustment to chair tilt. *Brain Research*, 143(2):325–338.

Apicella, P., Ljungberg, T., Scarnati, E., and Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. *Experimental Brain Research*, 85(3):491–500.

Arvanitogiannis, A. (1997). *Mapping the substrate for brain stimulation reward: New approaches to an old problem*. PhD thesis, Concordia University.

Arvanitogiannis, A. and Shizgal, P. (2008). The reinforcement mountain: allocation of behavior as a function of the rate and intensity of rewarding brain stimulation. *Behavioral Neuroscience*, 122(5):1126–1138.

Arvanitogiannis, A., Waraczynski, M. A., and Shizgal, P. (1996). Effects of excitotoxic lesions of the basal forebrain on MFB self-stimulation. *Physiology and Behavior*, 59(4-5):795–806.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the experimental analysis of behavior*, 32(2):269–281.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152.

Beckstead, R. M., Domesick, V. B., and Nauta, W. J. (1979). Efferent connections of the substantia nigra and ventral tegmental area in the rat. *Brain Research*, 175(2):191–217.

Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W., and Everitt, B. J. (2009). Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. *Behavioural Brain Research*, 199(1):89–102.

Bellman, R. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

Berendse, H. W., Galis-de Graaf, Y., and Groenewegen, H. J. (1992). Topographical organization and relationship with ventral striatal compartments of prefrontal corticostriatal projections in the rat. *The Journal of Comparative Neurology*, 316(3):314–347.

Beyene, M., Carelli, R. M., and Wightman, R. M. (2010). Cue-evoked dopamine release in the nucleus accumbens shell tracks reinforcer magnitude during intracranial self-stimulation. *Neuroscience*, 169(4):1682–1688.

Bielajew, C. and Shizgal, P. (1982). Behaviorally derived measures of conduction velocity in the substrate for rewarding medial forebrain bundle stimulation. *Brain Research*, 237(1):107–119.

Bielajew, C. and Shizgal, P. (1986). Evidence implicating descending fibers in self-stimulation of the medial forebrain bundle. *The Journal of neuroscience*, 6(4):919–929.

Bindra, D. and Mendelson, J. (1962). Interaction of habit strength and drug effects. *Journal of Comparative and Physiological Psychology*, 55:217–219.

Blanchard, D. C. and Blanchard, R. J. (1972). Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of Comparative and Physiological Psychology*, 81(2):281–290.

Boring, E. G. (1950). *A History of Experimental Psychology*. Appleton-Century-Crofts, New York, 2nd ed. edition.

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., and Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, 8(9):1263–1268.

Brady, J. V., Boren, J. J., Conrad, D., and Sidman, M. (1957). The effect of food and water deprivation upon intracranial self-stimulation. *Journal of Comparative and Physiological Psychology*, 50(2):134–137.

Breland, K. and Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(11):681.

Breton, Y.-A., Marcus, J. C., and Shizgal, P. (2009). Rattus Psychologicus: construction of preferences by self-stimulating rats. *Behavioural Brain Research*, 202(1):77–91.

Breton, Y.-A., Mullett, A., Conover, K., and Shizgal, P. (2013). Validation and extension of the reward-mountain model. *Frontiers in Behavioral Neuroscience*, 7:125.

Burke, C. J. and Tobler, P. N. (2011). Coding of reward probability and risk by single neurons in animals. *Frontiers in Neuroscience*, 5:121.

Cardinal, R. N. and Howes, N. J. (2005). Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. *BMC Neuroscience*, 6(1):37.

Carr, K. D. and Wolinsky, T. D. (1993). Chronic food restriction and weight loss produce opioid facilitation of perifornical hypothalamic self-stimulation. *Brain Research*, 607(1-2):141–148.

Chakravarthy, V. S., Joseph, D., and Bapi, R. S. (2010). What do the basal ganglia do? A modeling perspective. *Biological cybernetics*, 103(3):237–253.

Cohen, J. D., Braver, T. S., and Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, 12(2):223–229.

Conover, K., Fulton, S., and Shizgal, P. (2001a). Operant tempo varies with reinforcement rate: implications for measurement of reward efficacy. *Behavioural Processes*, 56(2):85–101.

Conover, K., Fulton, S., and Shizgal, P. (2001b). Operant tempo varies with reinforcement rate: implications for measurement of reward efficacy. *Behavioural Processes*, 56(2):85–101.

Conover, K. and Shizgal, P. (1994). Competition and summation between rewarding effects of sucrose and lateral hypothalamic stimulation in the rat. *Behavioral Neuroscience*, 108(3):537–548.

Conover, K. and Shizgal, P. (2005). Employing labor-supply theory to measure the re-

ward value of electrical brain stimulation. *Games and Economic Behavior*, 52:283–304.

Conover, K., Woodside, B., and Shizgal, P. (1994). Effects of sodium depletion on competition and summation between rewarding effects of salt and lateral hypothalamic stimulation in the rat. *Behavioral Neuroscience*, 108(3):549–558.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 1st edition.

de Villiers, P. and Herrnstein, R. J. (1976). Toward a law of response strength. *Psychological Bulletin*, 83(6):1131.

Deneve, S., Duhamel, J.-R., and Pouget, A. (2007). Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of Kalman filters. *Journal of Neuroscience*, 27(21):5744–5756.

Deneve, S., Latham, P. E., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., and Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, 23(2):197–206.

Eblen, F. and Graybiel, A. M. (1995). Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *The Journal of neuroscience*, 15(9):5999–6013.

Edmonds, D. E. and Gallistel, C. R. (1974). Parametric analysis of brain stimulation reward in the rat: III. Effect of performance variables on the reward summation function. *Journal of Comparative and Physiological Psychology*, 87(5):876–883.

Edmonds, D. E., Stellar, J. R., and Gallistel, C. R. (1974). Parametric analysis of brain stimulation reward in the rat: II. Temporal summation in the reward system. *Journal of Comparative and Physiological Psychology*, 87(5):860–869.

Estes, W. K. (1943). Discriminative conditioning. I. A discriminative property of conditioned anticipation. *Journal of Experimental Psychology*, 32(2):150.

FitzGerald, T. H. B., Seymour, B., and Dolan, R. J. (2009). The role of human orbitofrontal cortex in value comparison for incommensurable objects. *Journal of Neuroscience*, 29(26):8388–8395.

Fouriezos, G., Bielajew, C., and Pagotto, W. (1990). Task difficulty increases thresholds of rewarding brain stimulation. *Behavioural Brain Research*, 37(1):1–7.

Franklin, K. B. (1978). Catecholamines and self-stimulation: reward and performances effects dissociated. *Pharmacology, Biochemistry, and Behavior*, 9(6):813–820.

Fulton, S., Woodside, B., and Shizgal, P. (2006). Potentiation of brain stimulation reward by weight loss: evidence for functional heterogeneity in brain reward circuitry. *Behavioural Brain Research*, 174(1):56–63.

Gallistel, C. R. (1978). Self-stimulation in the rat: quantitative characteristics of the reward pathway. *Journal of Comparative and Physiological Psychology*, 92(6):977–998.

Gallistel, C. R., Mark, T. A., King, A. P., and Latham, P. E. (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4):354–372.

Gallistel, C. R., Stellar, J. R., and Bubis, E. (1974). Parametric analysis of brain stimulation reward in the rat: I. The transient process and the memory-containing process. *Journal of Comparative and Physiological Psychology*, 87(5):848–859.

Gilbert, R. J., Mitchell, M. R., Simon, N. W., Bañuelos, C., Setlow, B., and Bizon, J. L. (2011). Risk, reward, and decision-making in a rodent model of cognitive aging. *Frontiers in Neuroscience*, 5:144.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166.

Goto, Y., Otani, S., and Grace, A. A. (2007). The Yin and Yang of dopamine release: a new perspective. *Neuropharmacology*, 53(5):583–587.

Green, L. and Rachlin, H. (1991). Economic substitutability of electrical brain stimulation, food, and water. *Journal of the experimental analysis of behavior*, 55(2):133–143.

Green, L., Rachlin, H., and Hanson, J. (1983). Matching and maximizing with concurrent ratio-interval schedules. *Journal of the experimental analysis of behavior*, 40(3):217–224.

Hamilton, A. L. and Stellar, J. R. (1985). Reward, performance, and the response strength method in self-stimulating rats: Validation and neuroleptics. *Physiology and Behavior*.

Han, X. (2012). In vivo application of optogenetics for neural circuit analysis. *ACS Chemical Neuroscience*, 3(8):577–584.

Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of general physiology*, 25(6):819–840.

Hernandez, G., Breton, Y.-A., Conover, K., and Shizgal, P. (2010). At what stage of neural processing does cocaine act to boost pursuit of rewards? *PLoS ONE*, 5(11):e15081.

Hernandez, G., Haines, E., and Shizgal, P. (2008). Potentiation of intracranial self-stimulation during prolonged subcutaneous infusion of cocaine. *Journal of Neuroscience Methods*, 175(1):79–87.

Hernandez, G., Hamdani, S., Rajabi, H., Conover, K., Stewart, J., Arvanitogiannis, A., and Shizgal, P. (2006). Prolonged rewarding stimulation of the rat medial forebrain bundle: neurochemical and behavioral consequences. *Behavioral Neuroscience*, 120(4):888–904.

Hernandez, G., Trujillo-Pisanty, I., Cossette, M.-P., Conover, K., and Shizgal, P. (2012). Role of dopamine tone in the pursuit of brain stimulation reward. *Journal of Neuroscience*, 32(32):11032–11041.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, 4:267–272.

Herrnstein, R. J. (1974). Formal properties of the matching law. *Journal of the experimental analysis of behavior*, 21:159–164.

Herrnstein, R. J. and Heyman, G. M. (1979). Is matching compatible with reinforcement maximization on concurrent variable interval variable ratio? *Journal of the experimental analysis of behavior*, 31(2):209–223.

Herrnstein, R. J. and Prelec, D. (1991). Melioration: A theory of distributed choice. *The Journal of Economic Perspectives*, 5(3):137–156.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539.

Heyman, G. M. and Luce, R. D. (1979). Operant matching is not a logical consequence of maximizing reinforcement rate. *Animal Learning & Behavior*, 7(2):133–140.

Hitchcock, J. and Davis, M. (1986). Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience*, 100(1):11.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*, volume 3. Wiley New York.

Hodos, W. and Valenstein, E. (1960). Motivational variables affecting the rate of behavior maintained by intracranial stimulation. *Journal of Comparative and Physiological Psychology*.

Hodos, W. and Valenstein, E. (1962). An evaluation of response rate as a measure of rewarding intracranial stimulation. *Journal of Comparative and Physiological Psychology*, 55:80–84.

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. *Brain Research: Brain Research Reviews*, 56(1):27–78.

Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: A neurobiological theory. *Neuroscience and Biobehavioral Reviews*.

Jeffreys, S. H. (1998). *The Theory of Probability*. Oxford University Press.

Johnson, A. and Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189.

Johnson, A., van der Meer, M. A. A., and Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, 17(6):692–697.

Kacelnik, A. and Bateson, M. (1996). Risky theories—the effects of variance on foraging decisions. *American Zoologist*, 36(4):402–434.

Kagel, J. H., Battalio, R. C., and Green, L. (1995). *Economic choice theory: An experimental analysis of animal behavior*. Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Strauss and Giroux, 2nd ed. edition.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, Journal of the Econometric Society*, pages 263–291.

Kalenscher, T. and van Wingerden, M. (2011). Why we should use animals to study economic decision making – a perspective. *Frontiers in Neuroscience*.

Kelley, A. E. and Domesick, V. B. (1982). The distribution of the projection from the hippocampal formation to the nucleus accumbens in the rat: an anterograde- and retrograde-horseradish peroxidase study. *NSC*, 7(10):2321–2335.

Kelley, A. E., Domesick, V. B., and Nauta, W. J. (1982). The amygdalostriatal projection in the rat–an anatomical study by anterograde and retrograde tracing methods. *NSC*, 7(3):615–630.

Kelly, J. B., Cooke, J. E., Gilbride, Mitchell, C., and Zhang, H. (2006). Behavioral limits of auditory temporal resolution in the rat: amplitude modulation and duration discrimination. *Journal of Comparative Psychology*, 120(2):98.

Kennerley, S. W., Dahmubed, A. F., Lara, A. H., and Wallis, J. D. (2009). Neurons in

the frontal lobe encode the value of multiple decision variables. *Journal of cognitive neuroscience*, 21(6):1162–1178.

Kepecs, A., Uchida, N., Zariwala, H. A., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231.

Killeen, P. (1972). The matching law. *Journal of the experimental analysis of behavior*, 17(3):489–495.

Kirby, K. N. and Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological Science*, 6(2):83–89.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., and Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport*, 12(17):3683–3687.

Kringelbach, M. L., O'Doherty, J. P., Rolls, E. T., and Andrews, C. (2003). Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral Cortex*, 13(10):1064–1071.

Kurth-Nelson, Z., Bickel, W., and Redish, A. D. (2012). A theoretical account of cognitive effects in delay discounting. *The European Journal of Neuroscience*, 35(7):1052–1064.

Leon, M. I. and Gallistel, C. R. (1998). Self-stimulating rats combine subjective reward magnitude and subjective reward rate multiplicatively. *Journal of Experimental Psychology: Animal Behavior Processes*, 24(3):265–277.

Lighthall, J. W. and Kitai, S. T. (1983). A short duration GABAergic inhibition in identified neostriatal medium spiny neurons: In vitro slice study. *Brain Research Bulletin*, 11(1):103–110.

Logan, F. A. and Spanier, D. (1970). Chaining and nonchaining delay of reinforcement. *Journal of Comparative and Physiological Psychology*, 72(1):98.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

MacDonald, D., Kagel, J., and Battalio, R. (1991). Animals' choices over uncertain outcomes: Further experimental results. *The Economic Journal*, 101(408):1067–1084.

Mark, T. A. and Gallistel, C. R. (1994). Kinetics of matching. *Journal of Experimental Psychology: Animal Behavior Processes*, 20(1):79–95.

Mazur, J. E. (1995a). Conditioned reinforcement and choice with delayed and uncertain primary reinforcers. *Journal of the experimental analysis of behavior*, 63(2):139–150.

Mazur, J. E. (1995b). Development of preference and spontaneous recovery in choice behavior with concurrent variable-interval schedules. *Learning & Behavior*, 23(1):93–103.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., and Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, 31(7):2700–2705.

McDannald, M. A., Takahashi, Y. K., Lopatina, N., Pietras, B. W., Jones, J. L., and Schoenbaum, G. (2012). Model-based learning and the contribution of the or-

bitofrontal cortex to the model-free world. *The European Journal of Neuroscience*, 35(7):991–996.

McDonald, R. J. and White, N. M. (1993). A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience*, 107(1):3–22.

McDowell, J. J. (2005). On the classic and modern theories of matching. *Journal of the experimental analysis of behavior*, 84:111–127.

Miliaressis, E., Rompre, P.-P., Laviolette, P., Philippe, L., and Coulombe, D. (1986). The curve-shift paradigm in self-stimulation. *Physiology and Behavior*, 37(1):85–91.

Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377(6551):725–728.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5):1936–1947.

Morris, R. G., Garrud, P., Rawlins, J. N., and O'keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683.

Murray, B. and Shizgal, P. (1991). Anterolateral lesions of the medial forebrain bundle increase the frequency threshold for self-stimulation of the lateral hypothalamus and ventral tegmental area in the rat. *Psychobiology*.

Nakano, K., Kayahara, T., and Chiba, T. (1999). Afferent connections to the ventral striatum from the medial prefrontal cortex (area 25) and the thalamic nuclei in the macaque monkey. *Annals of the New York Academy of Sciences*, 877:667–670.

Neiman, T. and Loewenstein, Y. (2013). Covariance-based synaptic plasticity in an attractor network model accounts for fast adaptation in free operant learning. *Journal of Neuroscience*, 33(4):1521–1534.

Nieuwenhuys, R., Geeraedts, L. M., and Veening, J. G. (1982). The medial forebrain bundle of the rat. I. General introduction. *The Journal of Comparative Neurology*, 206(1):49–81.

Niv, Y. (2008). *The effects of motivation on habitual instrumental behavior*. PhD thesis, The Hebrew University.

Niv, Y., Daw, N. D., and Dayan, P. (2006). How fast to work: Response vigor, motivation and tonic dopamine. *Advances in neural information processing systems*, 18:1019.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacologia*, 191(3):507–520.

Noonan, M. P., Walton, M. E., Behrens, T. E. J., Sallet, J., Buckley, M. J., and Rushworth, M. F. S. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, 107(47):20547–20552.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K. J., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.

Olds, J. (1956). A preliminary mapping of electrical reinforcing effects in the rat brain. *Journal of Comparative and Physiological Psychology*, 49(3):281–285.

Olds, J. (1958). Satiation effects in self-stimulation of the brain. *Journal of Comparative and Physiological Psychology*, 51(6):675–678.

Olds, J. and Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.

Packard, M. G. and McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology Of Learning And Memory*, 65(1):65–72.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226.

Pai, S., Erlich, J., Kopec, C., and Brody, C. (2011). Minimal impairment in a rat model of duration discrimination following excitotoxic lesions of primary auditory and prefrontal cortices. *Frontiers in systems neuroscience*, 5.

Pennartz, C. M. A., van Wingerden, M., and Vinck, M. (2011). Population coding and neural rhythmicity in the orbitofrontal cortex. *Annals of the New York Academy of Sciences*, 1239:149–161.

Phillips, A. and LePiane, F. G. (1986). Effects of pimozide on positive and negative incentive contrast with rewarding brain stimulation. *Pharmacology, Biochemistry, and Behavior*, 24(6):1577–1582.

Plenz, D. and Kitai, S. T. (1998). Up and down states in striatal medium spiny neurons simultaneously recorded with spontaneous activity in fast-spiking interneurons studied in cortex-striatum-substantia nigra organotypic cultures. *The Journal of neuroscience*, 18(1):266–283.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132.

Prévost, C., Pessiglione, M., Météreau, E., Cléry-Melin, M.-L., and Dreher, J.-C. (2010). Separate valuation subsystems for delay and effort decision costs. *Journal of Neuroscience*, 30(42):14080–14090.

Pyke, G., Pulliam, H., and Charnov, E. (1977). Optimal foraging: a selective review of theory and tests. *The Quarterly Review of Biology*, 52(2):137–154.

Rachlin, H., Raineri, A., and Cross, D. (1991). Subjective probability and delay. *Journal of the experimental analysis of behavior*, 55(2):233–244.

Rescorla, R. A. and Wagner, A. R. (1972). Associative Learning and Conditioning Theory: Human and Non-Human Applications. In Black, A. and Prokasy, W., editors, *Classical conditioning II: current research and theory*, pages 64–99. Appleton-Century-Crofts, New York.

Roesch, M. R. and Olson, C. R. (2005). Neuronal activity in primate orbitofrontal cortex reflects the value of time. *Journal of Neurophysiology*, 94(4):2457–2471.

Roitman, J. and Roitman, M. F. (2010). Risk-preference differentiates orbitofrontal cortex responses to freely chosen reward outcomes. *The European Journal of Neuroscience*, 31(8):1492–1500.

Rolls, E. T. (2005). Taste, olfactory, and food texture processing in the brain, and the control of food intake. *Physiology and Behavior*, 85(1):45–56.

Rothemund, Y., Preuschhof, C., Bohner, G., Bauknecht, H.-C., Klingebiel, R., Flor, H., and Klapp, B. F. (2007). Differential activation of the dorsal striatum by high-calorie visual food stimuli in obese individuals. *NeuroImage*, 37(2):410–421.

Rudebeck, P. H. and Murray, E. A. (2008). Amygdala and orbitofrontal cortex lesions differentially influence choices during object reversal learning. *Journal of Neuroscience*, 28(33):8338–8343.

Sakai, Y. and Fukai, T. (2008). The actor-critic learning is behind the matching law: matching versus optimal behaviors. *Neural computation*, 20(1):227–251.

Salamone, J. D., Correa, M., Mingote, S., and Weber, S. M. (2005). Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology*, 5(1):34–41.

Shizgal, P., Bielajew, C., Corbett, D., Skelton, R., and Yeomans, J. (1980). Behavioral methods for inferring anatomical linkage between rewarding brain stimulation sites. *Journal of Comparative and Physiological Psychology*, 94(2):227–237.

Shizgal, P., Fulton, S., and Woodside, B. (2001). Brain reward circuitry and the regulation of energy balance. *International journal of obesity (2005)*, 25 Suppl 5:S17–21.

Simmons, J. M. and Gallistel, C. R. (1994). Saturation of subjective reward magnitude as a function of current and pulse frequency. *Behavioral Neuroscience*, 108(1):151–160.

Skinner, B. F., Ferster, C. B., and Ferster, C. B. (1997). *Schedules of reinforcement*. Copley Publishing Group.

Solomon, R. B., Conover, K., and Shizgal, P. (2007). Estimation of subjective opportunity cost in rats working for rewarding brain stimulation: further progress . In *Society for Neuroscience Annual Meeting*, pages 1–2, San Diego, CA.

Solomon, R. B., Trujillo-Pisanty, I., and Shizgal, P. (2010). The maximum firing frequency of the neurons subserving brain stimulation reward - Solomon. In *Society for Neuroscience Annual Meeting*, pages 1–4, San Diego.

Sonnenschein, B., Conover, K., and Shizgal, P. (2003). Growth of brain stimulation

reward as a function of duration and stimulation strength. *Behavioral Neuroscience*, 117(5):978–994.

St Onge, J. R., Abhari, H., and Floresco, S. B. (2011). Dissociable contributions by prefrontal D1 and D2 receptors to risk-based decision making. *Journal of Neuroscience*, 31(23):8625–8633.

St Onge, J. R., Chiu, Y. C., and Floresco, S. B. (2010). Differential effects of dopaminergic manipulations on risky choice. *Psychopharmacology*.

Staddon, J. E. R. (1992). Rationality, melioration, and law-of-effect models for choice. *Psychological Science*, 3(2):136–141.

Sutton, R. S. and Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88(2):135–170.

Taha, S. A., Nicola, S. M., and Fields, H. L. (2007). Cue-evoked encoding of movement planning and execution in the rat nucleus accumbens. *The Journal of physiology*, 584(Pt 3):801–818.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied*, 2(4).

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208.

Trujillo-Pisanty, I., Conover, K., and Shizgal, P. (2012). Dopamine receptor antagonism reduces the opportunity cost at which rats maintain operant performance for rewarding brain stimulation. In *Society for Neuroscience Annual Meeting*, pages 1–2, New Orleans, LA.

Trujillo-Pisanty, I., Hernandez, G., Moreau-Debord, I., Cossette, M.-P., Conover, K., Cheer, J. F., and Shizgal, P. (2011). Cannabinoid receptor blockade reduces the

opportunity cost at which rats maintain operant performance for rewarding brain stimulation. *Journal of Neuroscience*, 31(14):5426–5435.

van der Meer, M. A. A., Johnson, A., Schmitzer-Torbert, N. C., and Redish, A. D. (2010). Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron*, 67(1):25–32.

van der Meer, M. A. A. and Redish, A. D. (2009). Covert Expectation-of-Reward in Rat Ventral Striatum at Decision Points. *Frontiers in integrative neuroscience*, 3:1.

van der Meer, M. A. A. and Redish, A. D. (2011). Ventral striatum: a critical look at models of learning and evaluation. *Current Opinion in Neurobiology*.

van Duuren, E., Lankelma, J., and Pennartz, C. M. A. (2008). Population coding of reward magnitude in the orbitofrontal cortex of the rat. *Journal of Neuroscience*, 28(34):8590–8603.

van Duuren, E., van der Plasse, G., Lankelma, J., Joosten, R. N. J. M. A., Feenstra, M. G. P., and Pennartz, C. M. A. (2009). Single-cell and population coding of expected reward probability in the orbitofrontal cortex of the rat. *Journal of Neuroscience*, 29(28):8965–8976.

Vaughan, W. (1981). Melioration, matching, and maximization. *Journal of the experimental analysis of behavior*, 36(2):141–149.

Veening, J. G., Swanson, L. W., Cowan, W. M., Nieuwenhuys, R., and Geeraedts, L. M. (1982). The medial forebrain bundle of the rat. II. An autoradiographic study of the topography of the major descending and ascending components. *Journal of Comparative Neurology*, 206(1):82–108.

Voorn, P., Jorritsma Byham, B., Van Dijk, C., and Buijs, R. M. (1986). The dopaminergic innervation of the ventral striatum in the rat: A light- and electron-microscopical study with antibodies against dopamine. *Journal of Comparative Neurology*, 251(1):84–99.

Waraczynski, M. A. (2006). The central extended amygdala network as a proposed circuit underlying reward valuation. *Neuroscience and Biobehavioral Reviews*, 30(4):472–496.

Ward, H. P. (1959). Stimulus factors in septal self-stimulation. *The American journal of physiology*, 196(4):779–782.

Werner, G. and Mountcastle, V. B. (1963). The variability of central neural activity in a sensory system, and its implications for the central reflection of sensory events. *Journal of Neurophysiology*, 26:958–977.

White, N. M. and McDonald, R. J. (2002). Multiple parallel memory systems in the brain of the rat. *Neurobiology Of Learning And Memory*, 77(2):125–184.

Williams, B. A. and Royalty, P. (1989). A test of the melioration theory of matching. *Journal of Experimental Psychology: Animal Behavior Processes*, 15(2):99–113.

Wilson, C. J., Chang, H. T., and Kitai, S. T. (1990). Firing patterns and synaptic potentials of identified giant aspiny interneurons in the rat neostriatum. *The Journal of neuroscience*, 10(2):508–519.

Wise, R. A. (1982). Neuroleptics and operant behavior: The anhedonia hypothesis. *Behavioral and Brain Sciences*, 5(01):39–53.

Yacubian, J., Sommer, T., Schroeder, K., Gläscher, J., Braus, D. F., and Büchel, C. (2007). Subregions of the ventral striatum show preferential coding of reward magnitude and probability. *NeuroImage*, 38(3):557–563.

Yeomans, J. S. (1979). The absolute refractory periods of self-stimulation neurons. *Physiology and Behavior*, 22(5):911–919.

Yeomans, J. S. and Davis, J. K. (1975). Behavioral measurement of the post-stimulation excitability of neurons mediating self-stimulation by varying the voltage of paired pulses. *Behavioral biology*, 15(4):435–447.

Yin, H. H., Ostlund, S. B., and Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *The European Journal of Neuroscience*, 28(8):1437–1448.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *The European Journal of Neuroscience*, 22(2):513–523.

Zeeb, F. D. and Winstanley, C. A. (2011). Lesions of the basolateral amygdala and orbitofrontal cortex differentially affect acquisition and performance of a rodent gambling task. *Journal of Neuroscience*, 31(6):2197–2204.