What to do, when to do it, how long to do it for:

a normative microscopic approach to the labour leisure tradeoff

Ritwik Kumar Niyogi

Gatsby Computational Neuroscience Unitt
University College London
Alexandra House, 17 Queen Square
London, United Kingdom

THESIS

Submitted for the degree of
Doctor of Philosophy, University College London

2014

I, RITWIK KUMAR NIYOGI, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Dividing limited time between work and leisure is a common, everyday choice. Given the option, humans and other animals elect to distribute their time between work and leisure, rather than choosing all of one and none of the other. Traditional accounts of the allocation of time have characterised behaviour on a macroscopic timescale, reporting and studying average times spent in work or leisure. We develop a novel, normative, microscopic framework in which subjects approximately maximise their expected returns by making momentary commitments to one or other activity. This generic theoretical framework is applied to the work-leisure tradeoff. We determine the microscopic utility of leisure – an animal's innate preference irrespective of all other rewards and costs. We report our analyses of data from our collaboration with experimentalists who use brain stimulation reward (electrically stimulating reward circuits in the brain) – a powerful reward that does not satiate unlike food, and is not secondary, unlike money, on rat subjects. We show that in all subjects, this utility of leisure is non-linear. Subjects either prefer long leisure bouts all at one go, or many short ones, but are not indifferent to the division of leisure durations. We also develop new normative, microscopic models of how fatigue and satiation may impact decision-making, and make predictions about their effect on the temporal distribution of choices.

We then derive macroscopic utilities from microscopic ones and show how macroscopic facets such as imperfect substitutability can arise. We show that by integrating our microscopic choices we can build macroscopic characterisations that are not only equivalent to, but richer than those afforded by previous macroscopic characterisations. We therefore build a superset of traditional macroscopic quantifications. Our normative, microscopic approach sheds new light on the nature of temporally relevant behaviour and may provide a powerful framework for understanding the psychological processes and neural computations underlying real-time cost-benefit decision-making.

Acknowledgments

This thesis would not be possible without the help of several people. I thank them here, begging forgiveness of whomever I have inadvertently left out.

First and foremost, I would like to thank my advisor, Prof Peter Dayan, without whose gentle guidance, superlative supervision, dedicated direction, advice (and sometimes admonishments), encouragement, ingenious ideas, comments and criticisms, no part of this thesis would be at all possible. I thank him for taking me on as his student, expecting me to perform to the high standards that he sets for himself (or, in reality, some fraction thereof), and for believing in and persevering with me throughout the course of my graduate career, and especially during the writing of this thesis. I have learnt from him not only how to be a successful scientist and critical thinker, but also the skills that make for an impactful writer and effective communicator, the art of being a leader, director and visionary, and in general, how a man should carry himself with aplomb. He has been a true inspiration to expend my perspiration. It has been the privilege of a lifetime working with him, learning from him and in some minute way, being part of his legend. Despite my best efforts with all those above, no words can fully express the heartfelt gratitude I have for him; which is why (to borrow his catchphrase), I wish to simply, thank him very much, indeed.

I am extremely grateful to Prof Peter Shizgal and his brilliant laboratory members: Yannick-Andre Breton and Rebecca Solomon and Dr. Kent Conover at Concordia University for very generously sharing the data from their meticulously designed and dedicatedly carried out experiments, upon which a majority of this thesis is based. It is they who introduced me to the field of microeconomics and neuroeconomics, its ideas and its methods. I thank them for taking time out of their busy schedules to dispel any confusion I had, and their patience and persistence, questions and comments. Our multi-continent, multi-timezone collaboration has been a tremendously thrilling experience and I look forward to its upcoming stages whence we see our efforts bear fruit beyond our expectations.

I am much obliged to the faculty at the Gatsby Computational Neuroscience Unit. Firstly, Prof Peter Latham, for permitting me to advance beyond the first slide of my presentation during my interview here, and continue on till now, despite ”firing” and ”rehiring” me innumerable times. As my secondary supervisor, he has helped nurture my talents and chisel away at the numerous rough edges. He embodies how one can be a highly reputed and vastly admired, serious scientist yet be a jovial and genial human-being. He has also been instrumental in improving

the quality of my talks, or, in fact, teaching me how to give potent ones. I thank Prof Maneesh Sahani for the precision, incision and insightfulness of his comments and criticisms on my work and ideas. The material and ideas presented in this thesis would not be as improved and thought-through as in their present form without him. Indeed, Chapter 5 of this thesis was borne out of one of his off-the-cuff questions at one of my research-talks. I am grateful to Dr. Yee-Whye Teh for teaching me many interesting techniques in machine learning, his questions and quiet encouragement. Dr. Arthur Gretton cannot be thanked enough for his encouragement, advice and for listening to my problems. I hope to emulate his mastery of the delicate art of balancing carefree socialising without diminishing the respect that his rank and personality deserves.

I am indebted to those without whom my experiences here would not have been possible. The running of administrative issues has been extremely smooth, bordering sublime. I thank Reign Macmillan for extracting me into the UK, Rachel Howes for taking care of my administrative requirements in my early years at the Gatsby, Jennny Stubbings for her advice on being a successful administrator, and Barry Fong for his quick solutions to myriad issues and general friendship and company. I thank Dr. John Pelan and Faz Sadeghi for maintaining and developing the powerful computational resources without which most of the work in this thesis would not be achievable and their patience and ready help with the technical issues I faced. I also am obliged to Faz for her advice, encouragement and most importantly, appreciation for my tastes and choices and phrasings.

My spiritual ancestor, Yudhishthira, from the great Indian epic the *Mahabharata*, was once asked by *Yama* – the God of Death, whom amongst his brothers he would wish to save. Yudhishthira, the ever-wise, had replied simply, ”The wisest. For it is only in the company of the wise that one may himself be wise”. It is in the company of the wise men and women at the Gatsby that I have learned, improved, grown and prospered. I have the greatest gratitude for them all: Demis Hassabis, Shane Legg, Jeff Beck, Quentin Huys, Ulrik Beierholm, Dilan Gorur, Ricardo Silva, Jakob Macke, Andriy Mnih, Bharath Sriperumbudur, Biljana Petreska, Lars Buesing, Nicholas Hees, Dino Sejdinovic, Srini Turaga, Alexander Lerchner, Katherine Heller, Kai Krueger, Reza Moazzezi, Phillipp Hehrmann, David Barrett, David Schulz, Vinayak Rao, Mandana Ahmadi, Charles Blundell, Ross Williamson and Jan Gasthaus.

Although I am grateful to each of the above for their own special reason, space restrictions permit only a few randomly chosen mentions. I am especially indebted to Kai for the early help in understanding reinforcement learning, and the secrets to keeping up with our advisor's speed of speaking, thinking and writing; Demis for teaching me how to think like a visionary (and goading me to do the same); Srini for his advice (or mostly admonishments) on developing my career; Jeff for his advice on some matters developed and others best not mentioned here, Jakob and Reza for instilling belief in myself, Mandana for convincing me to come to the Gatsby, Biljana for her guidance (and chastisements), Vinayak and David Barrett for their experienced directions, particularly with the writing of the thesis, and Jan for all his computational prowess that has streamlined much of my work. I thank Ross and Charles for pushing me to excel; in particular, I shall remain indebted to my "friend-philosopher-and-guide" Charles for his tutelage on machine learning and his sharp criticisms during the initial stages of my research work. If I ever become better than him, it is because he made me so.

I am grateful to my class-mates Loic Matthey and Lloyd Elliott, who helped me through the initial months at the Gatsby and accompanied and advised me throughout the remaining ones on matters both professional and personal. I especially thank Loic for his help with graphics and computational resources used in this thesis, and Lloyd for telling me what not to do, when not to do it and how long not to do it for.

I thank the current post-docs and graduate students at the Gatsby who have made my last years here a sheer delight: Ben Dongsung Huh, Pedro Goncalves, Kyo Iigaya, Mehdi Keramati, Kevin Lloyd, Mijung Park, Zoltan Szabo, Agnieszka Grabska-Barwinska,Thomas Desautels, Gabi Teodoru, Maria Lomeli, Wittawat Jitkrittum, Carsen Stringer, Gergo Bohner, Laurence Aitchison, Balaji Lakshminarayanan, Heiko Strathmann, Joana Soldado Magraner, Vincent Adam, Federico Mancinelli, Marius Pachitariu and Arthur Guez.

I especially thank Arthur and Marius, my old flat-mates with whom preparing for examinations on a specially purchased whiteboard was a fantastic experience. I learned several issues in reinforcement learning from Arthur during our collaborations and look forward to continuing collaborating with him. Balaji has played a key role in listening to and solving my academic and administrative problems, for which I am deeply grateful. My discussions with Laurence–my "right hand man", has been central to the development of a wealth of the material presented in this thesis; I am much obliged to him for all his aid and assistance. Likewise, I thank Gergo–my "protege"–for the promise of carrying on my mantle when I depart. I especially thank Gabi for being the confidante and tremendous help that he is, and for encouraging me to neither be someone else nor myself, but always, my best self.

Over my years here, I have also learned from the various speakers at the Emotion

Club at the Functional Imaging Lab next door. I thank Prof Ray Dolan, who kindly permitted me to actively participate in these, Tim Behrens who inspired and encouraged me throughout, as well as Dominik Bach, Laurence Hunt, Zeb Kurth-Nelson, Michael Moutoussis, Francesco Rigoli and Robb Rutledge with whom I have collaborated or shared many insightful discussions.

I am grateful to past mentors and collaborators Phil Holmes, Jon Cohen, Jay McClelland, Pat Simen, Angela Yu, Fuat Balci, Yael Niv and Andrew Saxe for their advice and for checking on my progress through my years at the Gatsby. I am especially very thankful to Dr. KongFatt Wong-Lin, my collaborator, mentor and guide for his constant encouragement and nurturing and for persevering with me. I thank Nathaniel Daw for his advice regarding the Gatsby and my career, and Jeremiah Cohen for his understanding in postponing my arrival at his laboratory while I wrote my thesis.

I thank James Cooke, Jen Thompson and the other participants of the RIKEN 2013 course for their encouragement. I am grateful to Kenji Doya and the organisers of the Okinawa Computational Neuroscience Course for a life-changing experience. The professional connections and personal friendships that were developed by the sunsets on the seas of Okinawa run deep. I thank Azriel Ghadooshahy and Sheeraz Ahmad for the exciting experience of collaborating with them. I especially thank Kyo, Eva Naumann and Laureline Logiaco for being there for me (despite myself) in a heartbeat, in the best of times and especially, in the worst of times. I treasure their friendships.

My old and dear friends Eric Dichter, Rohan Sen, Jyotishman Nag, Ritoban Basu Thakur, Chandan Sapkota, Amartya Ganguly, Jitaditya Kumar, Asad Rauf and Stephanie Long have stuck with me through all trials and tribulations – I will always be grateful to them. I particularly thank Eric, Rohan and Jyotishman for putting me up during my annual visits to the US, and providing me a place and incomparable company to return to after interviews and meetings.

My flatmates: Saber, Agathe, Chynna and Barbara have withstood my idiosyncrasies over the years, I am grateful to them for all their advice and support.

My parents, Sipra and Ranjan Kumar Niyogi, have lived my dreams throughout my life, and persevered with me unflinchingly with all their love. Thanking them here is, in my view, many times insufficient than the desire and promise to continue making them proud.

I thank my cousins Ayan Bose and Subhradeep DeSarkar, as well as Debasmita Basak – this thesis got a lot easier with a true economist amongst my friends. I am particularly grateful to Subhradeep, for being my support system on this beloved island of ours, for our many trips abroad, for inducting me to the Porterhouse, and in essence, redefining the word ”leisure”. I dedicate this thesis on work and leisure to him. I thank both my cousins for making sure I focus on writing and submitting my thesis.

I express my heartfelt thanks to Frauke Harms, Maria Niedernhuber, Feryal Mehraban Pour Behbahani, Laetitia Jubin and Joana Soldado Magraner for forcing me to focus on doing nothing but write my thesis, breathe, write my thesis, eat, write my thesis, sleep and write my thesis. They have (ironically) curtailed my distractions and bade me focus.

Finally, I thank Ben, Pedro, Joana, Gabi, Heiko, Kyo, Vincent and Gergo for preserving my sanity when I had all but lost it, getting me on my feet and making me move forward through difficult times. Their friendship, care and support has meant everything to me. I dedicate this thesis to them as well.

Contents

Front matter

List of figures

List of tables

List of algorithms

Chapter 1

Introduction

What to do, when to do it and how long to do it for are fundamental questions for behaviour. Different options across these dimensions of choice yield different costs and benefits. A human or non-human animal deciding across these dimensions thus faces a rich, complex, optimisation problem. For example, consider an individual deciding how to divide limited time between working and enjoying leisure. Work leads to monetary renumeration, but reduces the time available for enjoying the fruits of leisure. For example, take the case of a person waiting at a bus stop, deciding how much longer to wait for a bus: if she waits further, then she may be able to board the bus that takes her conveniently and inexpensively to her destination, although she could also no longer persist and simply take the more expensive tube. Equally, an individual lifting weights at a gym, deciding whether to do a another repetition must contrast the long-run benefits of this against the fatigue or effort it costs. Finally, an animal foraging in the wild, choosing whether to stay in the current patch or leave and forage in another must compare the gains of staying against the costs of leaving.

Most research investigating temporal choices made by humans and other animals have focused on *molar* or *macroscopic* characterisations of behaviour Baum (2002, 2001, 2004, 1995); Baum and Rachlin (1969); Baum (1976) capturing the average times allocated to pursuing a particular activity compared to those spent on another. These provide a coarse, holistic description of behaviour. *Molecular* or *microscopic* analyses characterise the detailed temporal topography of choice, i.e., the fine-scale structure of allocation Ferster and Skinner (1957); Gilbert (1958); Shull et al. (2001); Williams et al. (2009b,a,c), that is lost in molar averages. These offer greater insight into the cost-benefit computations underlying real-time decision-making. In this thesis, we shall characterise behaviour at this microscopic level of detail.

The question of how to allocate time between different activities, each associated with its own benefits and costs, has been studied by economists Frank (2005); Kagel et al. (1995); Battalio et al. (1981); Camerer et al. (1997); Green et al. (1987), behavioural psychologists Skinner (1938, 1981); Herrnstein (1961, 1974); Baum and Rachlin (1969); Baum (1974, 1981); Green and Rachlin (1991); McDowell (1986); Dallery et al. (2000); McDowell (2005), ethologists Haccou and Meelis (1992) and more recently, neuroscientists Conover and Shizgal (2005); Arvanitogiannis and Shizgal (2008); Breton et al. (2009b); Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012); Niv et al. (2007). When attempting to solve a problem such as this we may use the following approach Marr (1982). We may first ask *what* the problem is, or equivalently, what is the goal for a human or other animal trying to solve this problem. A theory that addresses the question at this level is termed a *computational* one, since it seeks to explain what computation is being attempted. Next we may ask *how*, using what algorithms, does an animal simplify and solve the problem. Such an account is *algorithmic*. Finally, we may proceed to the *implementational* level and ask how the problem, the algorithm and solution are implemented in the biological substrate of the brain.

To start at the bottom, at the implementational level, the neuromodulator dopamine has been closely implicated as playing a critical role in biasing action selection towards more beneficial choices. The phasic (short bursts) release of midbrain dopamine neurons have been shown to compute or carry a reward prediction error signal Montague et al. (1996); Schultz et al. (1997); Hollerman et al. (2000); Waelti et al. (2001); McClure et al. (2003); Bayer and Glimcher (2005); Cohen et al. (2012): the difference between the reward received and that expected. This can then be provided to downstream basal ganglia structures to learn the values of stimuli associated with rewards, or update the propensities of executing different actions. The role of tonic (baseline levels) of dopamine in less well understood. This has been proposed to modulate the magnitude of a reward, or signal the rate of receiving rewards Niv et al. (2007); Cools et al. (2011); Dayan (2012) and determine vigour, or carry the costs of physically effortful choices Salamone and Correa (2002). These hypotheses are indistinguishable when behaviour is characterised at a macroscopic level. We shall provide a novel, microscopic framework in this thesis that may help distinguish these.

Algorithmic explanations of choice behaviour have been both macroscopic and microscopic. A macroscopic algorithm, popular in behavioural psychology posits that subjects match their relative times allocated to different choices to the relative payoffs (scalar combinations of reward rates) associated with them Herrnstein (1961, 1974); Baum (1974); McDowell (1986, 2005). This, however, largely does not (attempt to) explain *why* experimental subjects exhibit matching behaviour. From a computational perspective, matching behaviour is suboptimal—a rational subject should devote all its time to the most beneficial action; rather than distribute its time among other alternatives by matching. Alternatively, microscopic algorithmic descriptions, such as continuous time Markov chain (CTMC) models, are popular in the field of ethology Haccou and Meelis (1992). These characterise the entire sequence of observed behaviour using a small set of parametric distributions; with the parameters governing the summary statistics of the durations for which activities occur. They also lack a normative basis.

These algorithmic models are *descriptive*, characterising *what* the animal does, rather than being *normative*: positing *why* it may do so. The microeconomic theory of labour supply Frank (2005) provides a computational, macroscopic framework of studying the allocation of time, particularly focussing on the problem of dividing limited time between work and leisure. An individual is assumed to maximise its utility, the subjective worth it earns, by trading off work and leisure, subject to the constraints of limited time. Microscopic computational accounts have been afforded by the theoretical frameworks of reinforcement learning Sutton and Barto (1998); Puterman (2005). Subjects are posited to maximise their expected returns–the net summed utilities–by maximising benefits accrued and minimising costs incurred. While the flavour of reinforcement learning popularly used in neuroscience considers choices at discrete time points, characterising fine-scale temporal behaviour requires considering choices made more continuously in time. In this thesis, we extensively apply the reinforcement learning framework of average reward semi-Markov decision processes Puterman (2005). In this framework, subjects choose all three of what to do, when and how long for, attempting to maximise the average rate of rewards. While this framework was previously used to provide a normative account of response vigour–how fast to perform actions Niv et al. (2007); Cools et al. (2011); Dayan (2012), in this thesis, we extend it to the case of how long to persist with an action.

We shall thus develop a novel, normative, microscopic approach to characterising behaviour. This seeks to characterise all the choices that an individual makes, but from the perspective of it attempting to maximise its returns. We shall exercise our generic theoretical framework on collaborative experiments investigating free-operant choice. Subjects are largely free to choose which actions to take, when to take them and how to persist with them. Free-operant behaviour is thus the purest expression of choice, minimally encumbered by the experimenter's manipulations. While the normative, microscopic approach is a generic theoretical framework applicable to choices between durations of different activities, throughout the thesis we shall use a simplified instantiation: dividing limited time between non-effortful work and leisure. We shall use behavioural data from experiments in rodents who perform experimenter determined activities in exchange for brain stimulation rewards (BSR): direct electrical stimulation of the reward circuits in the brain Olds and Milner (1954). BSR is a potent, powerful reward and neither satiates (unlike gustatory rewards) nor is a secondary reinforcer (unlike money). This enables the collection of reliable and psychophysically stable–hence repeatedly reproducible data over many months.

1.1 Organisation of the thesis

Since the generic problem of what to do, when to do it and how long to do it for is complicated, throughout this thesis we shall focus on a simplified yet common everyday choice: dividing limited time between work and leisure. In Chapter 2, we shall review the relevant literature on macroscopic and microscopic characterisations of behaviour, both computational and algorithmic, and the schedules of reinforcement used to study them. We shall emphasise those particularly investigating the allocation of time between work and leisure: labour supply theory and generalized matching. We shall further contribute two technical reviews: the literature on BSR and that on average reward reinforcement learning.

We shall introduce our novel, *normative microscopic* theoretical framework in Chapter 3, exercising it to qualitatively capture the key microscopic features of data from rodent experiments choosing between working and engaging in leisure. We shall show that the functional form of the microscopic utility of leisure plays a critical role in capturing the microscopic choice data. The microscopic utility of leisure quantities an individual's innate preference for durations of leisure, irrespective of all other rewards and costs. We shall integrate our microscopic choices in Chapter 4 to not only derive traditional macroscopic characterisations prevalent in both economics and behavioural psychology, but also build a superset of them. We shall show that the assumptions of the traditional macroscopic theories are unnecessary or insufficient when behaviour is characterised from a normative, microscopic perspective. For instance, we shall propose that considering leisure to be beneficial on its own accord, and not because of the recent history of working/rewards received can be used to construct results consistent with those from labour supply theory.

We shall extend our framework in Chapter 5 to the case where leisure can indeed be beneficial because of the recent history of working, e.g. due to fatigue, or rewards e.g. due to satiation. We shall use the latter to explore a curious case where subjects work less when rewards are greater, rather than more. We shall devote Chapter 6 to empirically and quantitatively determining the microscopic utility of leisure, using experimental data from rodent subjects in which other confounding factors, such as satiation, fatigue and effort costs were controlled for.

Finally, in Chapter 7 we detail the contributions of the thesis. We also list several exciting yet plausible general directions for future research.

A list of symbols and their meanings is provided in Table 1.1

Symbol	**Meaning**
$1/\lambda$	mean of exponential effective prior probability density for leisure time
a	matching coefficient
AFR	average foregone reward (also known as the opportunity cost of time)
$\alpha \in [0, 1]$	weight on linear component of microscopic benefit-of-leisure
BC	budget constraint
$\beta \in [0, \infty)$	inverse temperature or degree of stochasticity-determinism parameter
CHT	Cumulative Handling Time (schedule)
$C_L(\cdot)$	microscopic utility of leisure
$C_{L_{max}}$	maximum of sigmoidal microscopic utility of leisure
$C_{L_{shift}}$	shift of sigmoidal microscopic utility of leisure
$\delta(\cdot)$	delta/indicator function
$\mathbb{E}_\pi$	expected value with respect to policy π
f	frequency of pulses in a BSR electrical stimulation train
F_{hm}	frequency at which RI is half its maximal worth
FR	Fixed Ratio (schedule)
FI	Fixed Interval (schedule)
g	slope of the logistic reward growth function
$H(\pi)$	entropy
IC	indifference curve
K_L	marginal utility of linear microscopic utility of leisure
L	leisure
l	cumulative amount of time spent in leisure
macroscopic	coarse, holistic analyses reporting averages over long times
microscopic	fine-scale characterisations of temporal behaviour, but considering commitments to durations of time
miniscopic	average behaviour over time-windows
$\mu_a(\tau_a)$	effective prior probability density of choosing duration τ_a
nanoscopic	finest-scale characterisations of temporal behaviour, choices made at every moment in time
N	total number of rewards accrued
ν	dynamic fatigue variable
P	Price

P_L	price at which $TA = 0.5$, for a maximum subjective reward intensity RI_{max}
PRP	post-reinforcement pause
$\pi([a, \tau_a] \mid \vec{s})$	policy or choice rule: probability of choosing action a, for duration τ_a from state $\vec{s}$
Pav_{shift}	shift parameter of the sigmoidal Pavlovian leisure function
post	post-reward
pre	pre-reward
ψ	dynamic satiation variable
$Q(\vec{s}, [a, \tau_a])$	expected return or (differential) Q-value of taking action a, for duration τ_a from state $\vec{s}$
ρ	reward rate
$\rho\ \tau_a$	average foregone reward (opportunity cost of time) for taking action a for duration τ_a
RI	(subjective) Reward Intensity
RI_{max}	maximum (subjective) Reward Intensity
$R_W = \frac{RI}{P}$	payoff or wage rate
s	degree of substitutability between rewards (or work) and leisure
$\vec{s}$	state
$\sigma(\cdot)$	logistic function
SMDP	semi-Markov decision process
T	trial duration
TA	Time Allocation
τ_L	duration of instrumental leisure
τ_{Pav}	duration of Pavlovian leisure
τ_W	duration of work
ω	cumulative amount of time spent in work
W	work
$w \in [0, P)$	amount of work time so far executed out of the price
$V(\vec{s})$	expected return or value of state $\vec{s}$
VI	Variable Interval (schedule)
U	macroscopic utility

Table 1.1: Table of symbols/glossary

Chapter 2

Literature Review

2.1 Introduction

One common everyday decision is between working (performing an employer-defined task) and engaging in leisure (activities pursued for oneself). Working leads to external rewards such as food and money; whereas leisure is supposed to be intrinsically beneficial (otherwise one would not want to engage in it). Since these activities are usually mutually exclusive, subjects must decide how to allocate time to each. Note that work need not be physically or cognitively demanding, but consumes time; equally leisure need not be limited to rest, and may present physical and/or mental demands.

The division of time between work and leisure has been extensively studied in both humans and other animals, both in the laboratory and in real-life field studies. Most of these studies have characterised behaviour at the macroscopic timescale, investigating average rates of responding and proportions of time allocated. There has also been some amount of research characterising microscopic choices. Theories have approached the work-leisure tradeoff from computational and algorithmic perspectives. In this Chapter, we shall review the literature on these approaches. We shall also contribute two brief, technical literature reviews that are necessary for the rest of the thesis: on brain stimulation reward and average reward semi-Markov decision processes.

2.2 Microeconomics: Labour supply theory

The prominent theory of how individuals should divide their time between work (labour) and leisure is the macroscopic labour supply theory. This has been largely applied to studies in humans, but also to animals tested in the laboratory. Macroscopic, but microeconomic labour supply theorists Frank (2005) have approached the question of dividing time between work and leisure at a computational level, formulating what problem an agent is solving. They have adopted a normative perspective, formulating what a rational agent *should* do. The optimal allocation of time between work and leisure is of critical economic importance, not only from the perspective of the subject making the choice but also from that of the employer or firm. The latter must set wages, i.e. the rewards of working and work requirements, appropriately so as to maximise profit from production. Typically, the longer a subject works, the more labour it supplies, and so the greater the employer or firms production. The microeconomic theory of labour supply Frank (2005), which we review here, studies the problem from the subject's perspective.

In labour supply theory Frank (2005), subjects are assumed to maximize their *macroscopic* utility by trading off two goods (i) income from working against (ii) leisure. Work, though not necessarily effortful or cognitively demanding, is considered a bad, since it consumes time that a subject could instead invest in leisure. In economic nomenclature, work incurs an *opportunity cost* when leisure is more valuable.

Let m be the *total* income that a subject accumulates over a large time-period (e.g. a day in the case of a person), and l be the *cumulative* amount of time spent in leisure. A *macroscopic utility function* $U(l, m)$ is defined, which increases with both income and leisure, since more of a good is always better than less. Fig.2.1A shows the indifference curves (IC)–contours of equal utility. A subject is indifferent between combinations of these goods along an IC, but combinations on an IC with greater utility are preferred. Preferences are thus transitive. The slope of an IC, negative of which is called the *marginal rate of substitution*, shows how willing a subject is to substitute one good with the other, depending on how much of each it has already consumed. We shall discuss *substitutability* in greater detail in Chapter 4. ICs are typically convex and negatively sloped, subjects must tradeoff consuming an extra unit of one good with reducing some units of the other.

The total time T in which income and leisure can be consumed is fixed (e.g. a day in the case of a person, or an experimental trial or session in the case of an

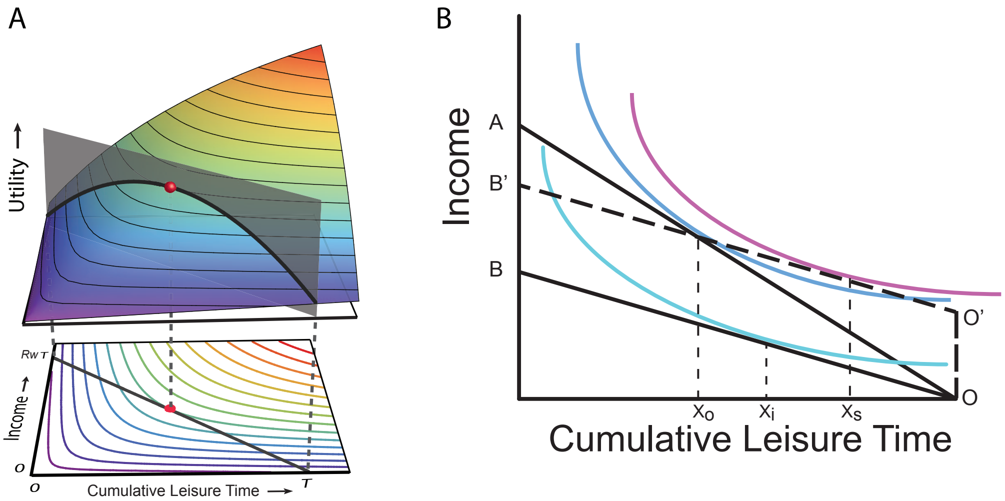


Figure 2.1: Labour supply theory. A) A macroscopic utility function $U(l, m)$ is defined over a combination of two goods: the *total* income that a subject accumulates, and the *cumulative* amount of time spent in leisure. The utility function increases with both income and leisure. Indifference curves (ICs) are contours of equal utility. A subject is indifferent between combinations of these goods along an IC, but combinations on an IC with greater utility are preferred. Black line shows the budget constraint: total duration T is fixed. The optimal combination of income and leisure time is obtained by maximising the utility function subject to the budget constraint, i.e., when the budget constraint is tangent to an IC. Graphic adapted from Kool and Botvinick (2012). B) A wage rate change that is not compensated by any free, prior income can be decomposed into two effects. For example, suppose the wage rate is decreased, so that the budget constraint pivots from OA to OB. The first effect is the substitution effect, which would be due to an imaginary income compensated wage decrease. The free, prior income OO' is provided as compensation for the reduced income possible due to the wage rate decrease. The budget constraint would then shift from OA to O'B', leaving the subject the opportunity to consume the same income-leisure combination (X_o). However, this imaginary budget constraint is tangent to an indifference curve with a greater utility. The optimal allocation is to allocate more time to leisure, and work more (X_s), i.e. substitute leisure for work. Second, the income effect. The decreased wage rate enables the subject to gain less income. The budget constraint is shifted downward from O'B' in parallel, to OB. The new budget constraint OB is tangent to an indifference curve for which the optimal combination of income and leisure (X_i) involves the cumulative leisure time decreasing compared to the imaginary level (X_s) but increasing compared to the original level (X_o). As a consequence of the wage rate decrease, subjects thus work less and engage in leisure more.

animal subject). This enforces a budget constraint, here expressed in terms of the goods income and leisure (Eq.(2.1), but see Eq. (4.17) when it is expressed in terms of time),

$$m = R_W \ \omega = R_W \ (T - l) \tag{2.1}$$

where R_W is the wage rate (the employer or experimenter determined reward per unit time spent working) and $\omega = T - l$ is the *cumulative* amount of time spent working.

Subjects must maximise their macroscopic utilities subject to this budget constraint. The optimal combination of leisure and income is

$$(l^*, m^*) = \mathrm{argmax}_{(l,m)}[U(l, m) - \lambda\,(T - (m/R_W + l))] \tag{2.2}$$

where λ is a Lagrange multiplier to enforce the budget constraint. This is equivalent to the *shadow price* or *marginal utility of wealth* i.e., the extra (macroscopic) utility arising from relaxing the total budget T, (i.e. taking an extra unit of total time for work and/or leisure).

The optimal combination of goods is that at which the budget constraint is tangent to an IC or is at a boundary (Fig.2.1)

Now, as discussed above, macroscopic utilities with respect to both income ($\frac{\partial U}{\partial m}$) and leisure ($\frac{\partial U}{\partial l}$) are positive. Then, since macroscopic utility is constant on an indifference curve, the total derivative with respect to a good (say leisure) is zero:

$$\begin{aligned} \frac{dU}{dl} &= \frac{\partial U}{\partial l} + \frac{\partial U}{\partial m}\frac{dm}{dl} = 0 \\ \Rightarrow \frac{dm}{dl} &= -\frac{\partial U}{\partial l} / \frac{\partial U}{\partial m} < 0 \end{aligned} \tag{2.3}$$

This shows that indifference curves have negative slopes ($\frac{dm}{dl} < 0$).

The optimum (l^*, m^*) associated with the budget constraint occurs when

$$\begin{aligned} \frac{\partial U}{\partial m} - \lambda \frac{1}{R_W} = 0 \\ \frac{\partial U}{\partial l} - \lambda = 0 \\ \Rightarrow -\frac{dm}{dl} = \frac{\partial U}{\partial l} / \frac{\partial U}{\partial m} = R_W \end{aligned} \tag{2.4}$$

At the optimum, negative of the slope of the IC, i.e. the marginal rate of substitution, is equal to the wage rate R_W.

2.2.1 Substitution and income effects: effects of wage rate changes

Studies on the effect of wage rate changes on labour supply come in one of two varieties: those providing a free, prior income to compensate for the change in total income attainable, and those that do not (for an extensive review see Kagel et al. (1995)). A wage rate change that is not compensated by any accompanying income can be decomposed into two effects (Fig.2.1B). The first effect can be considered as an income-compensated wage change that would allow the subject to maintain the same level of income and leisure (with the budget constraint passing through the same income-leisure combination). Although the subject has the opportunity to maintain the same income-leisure combination, the effect of an income compensated wage change is that the budget constraint is now tangent to a different indifference curve with greater utility. Since combinations with greater utility are preferred to those with less, the subject will now substitute more income for leisure (by working more) if the wage rate increases and vice versa. This is called the substitution effect. How much income is substituted for leisure depends on the *substitutability* between the goods — reflecting a subject's willingness to substitute income for leisure based on how much of each it has already consumed.

The second effect reflects the greater (less) income available owing to the increased (decreased) wage rate. The budget constraint shifts upwards (downwards) in parallel due to the wage rate increase (decrease). The optimal income-leisure combination due to the wage rate change occurs when this final budget constraint is tangent to an IC. As long as the substitution effect dominates the income effect, subjects will always work more if a wage rate is increased. We shall discuss the scenario when this is not true in Chapter 5.

2.2.2 Random utility theory

Maximising utility, as above implies deterministic choices. Behavioural data, on the other hand, is inherently variable. This mismatch between theory and data can be alleviated using *random utility theory* McFadden (1984), by adding unobservable noise to the representation of utility and then choosing the consumption bundle with the greatest utility [1]

$$\bar{U}(\omega|T) = U(\omega|T) + \eta(\omega|T). \tag{2.5}$$

The noise $\eta(\omega|T)$ is commonly assumed to be Gumbel distributed (i.e. drawn from an extreme value distribution of type I). Then the probability of choosing the optimal consumption bundle is

$$Pr(\omega|T) = Pr\left(\bar{U}(\omega|T) = \max_{\omega' \in [0,T]} \bar{U}(\omega'|T)\right) = \frac{\exp\left[U(\omega|T)\right]}{\int_0^T d\omega' \ \exp[U(\omega'|T)]} \tag{2.6}$$

which is a *softmax*–rather than a *max*–function of the macroscopic utilities McFadden (1984); Dagsvik et al. (2012). Choices are *consequently* stochastic, rather than deterministic. The softmax function, which we shall use throughout the thesis, is ubiquitous in reinforcement learning as well.

2.3 Miniscopic labour supply theory

A dynamic extension to the static labour supply theory discussed above, is to consider labour supply in time-windows. For example, this can be used to study how interest rates and wage changes affect the supply of labour over the course of a day or even a lifetime Blundell and Macurdy (1999). Considering allocations within time-windows yields a finer characterisation than that from the macroscopic static labour supply theory; but still does not provide a characterisation of the detailed temporal structure of choice. It is thus neither macroscopic nor microscopic. We therefore term this sort of characterisation *miniscopic*.

A utility function over an entire lifetime $\hat{U} = f_u(U_0(l_0, m_0), \ldots, U_t(l_t, m_t), U_{t+1}(l_{t+1}, m_{t+1}), \ldots, U_{t_{end}}(l_{t_{end}}, m_{t_{end}})))$ can be defined, which includes the utilities $U_t(l_t, m_t)$ in each time-period $0, \ldots, t_{end}$, where $f_u(\cdot)$ is some function and t_{end} is the final time-period. The utility function is usually considered to be separable, and more specifically, linearly separable in time, i.e. $f_u(\cdot)$ is a discounted linear sum of utilities

[1] We express the macroscopic utility function in terms of total work time ω, since the total leisure time chosen, $l = T - \omega$, can be computed by subtracting the total work time from the total time T.

$$\hat{U} = \sum_{t=0}^{t_{end}} \gamma^t \; U_t(l_t, m_t) \tag{2.7}$$

where $\gamma \in (0, 1]$ is a discount factor reflecting the subject's rate of time preference. $\gamma = 1$ implies future utilities are not discounted at all.

This utility is maximised subject to a dynamic budget constraint,

$$Y_0 + \sum_{t=0}^{t_{end}} [R_{W_t}(T - l_t)] = \sum_{t=0}^{t_{end}} m_t \tag{2.8}$$

where Y_0 is some initial wealth, l_t, m_t and R_{W_t} are the cumulative leisure times, incomes and wage rates in time-period t. Maximising the lifetime utility subject to this dynamic budget constraint can be done using dynamic programming, by first solving for the optimal income-leisure combination $l^*_{t_{end}}, m^*_{t_{end}}$ in the final time-period subject to the final budget constraint. This is then used to solve for the optimal combination in the previous time-period, which is then used for the preceding time-period, and so on. In essence, this recurses static labour supply models in time-windows. The allocations between income and leisure are still averages for each time-window. Dynamic labour supply is useful for characterising average times spent working in each individual hour over the course of a day, or years over the course of a lifetime. It cannot still account for the fine-scale temporal topography, or the microstructure of work and leisure choices. For these we would need to turn to microscopic theories of behaviour. Testing these on animal, rather than human subjects, and in the laboratory with appropriate schedules of reinforcement, rather than in field studies, enables a cleaner understanding of the processes underlying temporally relevant behaviour, with extraneous confounding factors under closer control.

2.4 Schedules of Reinforcement

Free-operant behaviour, in which animals (typically rats, mice or pigeons) are free (with few restrictions) to choose all three of what to do, when to do it and for how long/fast to do it, i.e., actions and their durations/vigour, is minimally encumbered by experimenter manipulation. Actions usually involve pressing a lever, pulling a chain, pecking a key etc in order to acquire some reinforcement (such as food or water for a hungry or thirsty animal). These experimenter defined actions or tasks performed by an animal in order to receive rewards are defined as work (W); with all other activities that an animal performs (e.g. grooming, resting, exploring), presumably for their intrinsic benefits, defined to constitute leisure (L). Animals adjust their behaviour according to the experimenter determined schedule of reinforcement Ferster and Skinner (1957). Compared to reinforcement schedules where choices are made at discrete time points Sutton and Barto (1998), animals largely set the pace of the task. Free operant behaviour is thus a comparatively pure expression of choice.

2.4.1 Free-Operant schedules

The schedules of reinforcement commonly used are ratio and interval schedules Baum (1993); Domjan (2003). Interval schedules stipulate that the first response after an unsignalled predetermined interval has elapsed is reinforced. The duration of the interval can be fixed (fixed interval (FI) schedule of e.g. 15 seconds is denoted FI15) or randomly drawn from a distribution. If the interval duration distribution is exponential, then it is traditionally called a random interval (RI) schedule; for all other distributions it is called a variable interval (VI) schedule. Intervals are timed from the previous reinforcement, except for the first interval which is timed from the start of an experimental session. In ratio reinforcement schedules, reinforcement is only provided after a predefined number of responses–which can either be fixed (fixed ratio–FR schedules) or randomly drawn from a distribution (variable ratio–VR schedules; random ratio of RR schedules when drawn from a Geometric distribution).

Since animals adjust their behaviour to the schedule of reinforcement, ratio and interval schedules lead to different relationships between the animal's rate of responses and the rate of reinforcement. In ratio schedules response rates are linearly related to reinforcement rates. By contrast, on interval schedules, responses before the termination of the interval is are not reinforced, and the rate of reinforcement is further curtailed by the interval duration. Consequently there is a nonlinear saturating relationship between responses and reinforcements. In general, response rates are slower for longer intervals or larger ratios Herrnstein (1970); Barrett and Stanley (1980); Mazur (1983); Baum (1993); Killeen (1995); Foster et al. (1997), and faster for reinforcers with higher magnitude or quality Bradshaw et al. (1979, 1981a,b).

Behaviour on fixed interval schedules, provides evidence that animals attempt to keep track of time during the interval. Although this is susceptible to timing errors (which scale with the duration of the interval, Gibbon (1977); Gallistel and Gibbon (2000), animals attempt to respond only towards the end of the interval. There is a characteristic pause in responding, called the *post-reinforcement pause*

(PRP) after a reward is obtained. The PRP does not solely reflect time involved in consuming rewards; its duration scales with the mean duration of the interval. The analysis of PRPs shall play a central role in the characterisation of behaviour, and the novel approaches and theories that we develop in explaining them in this thesis. When the above microscopic pattern of behaviour is averaged across animals and intervals, the macroscopic response rates show a scalloping pattern. Response rates are low after the PRP, and accelerate to a higher rate as the end of the interval approaches (see Fig. 2.2).

Responding on fixed ratio schedules is quite regular, except for a paradoxical PRP observed on high ratio schedules Felton and Lyon (1966). It is probable that animals confuse the long inter-reinforcer intervals, which are a side effect of the high ratio requirements, with an interval schedule. This can be deduced from the fact that this PRP is similar in duration to those on interval schedules which are 'yoked', i.e. matched in their interval lengths to the inter-reinforcer intervals on ratio schedules.

2.4.2 Cumulative Handling Time schedule

One disadvantage of conventional schedules of reinforcement such as interval or ratio schedules is that they control either the (average) minimum inter-reward interval or the (average) amount of work required to earn a reward, respectively, but not both. To overcome this Breton et al. (2009b) developed the cumulative handling time (CHT) schedule (Fig. 2.3), which controls both, making it a generalisation of conventional schedules. Subjects choose between working–the facile task of holding down a light lever, and engaging in leisure, i.e., resting, grooming, exploring etc (Figure 2.3). A reward is given after the subject has accumulated work for an experimenter-defined total time-period called the *price* (P). Throughout a task trial, the objective strength of the reward and price are held fixed. The total time a subject could work per trial is fixed at some proportion of the price. For example the trial duration could be 25 times the price (plus extra time for 'consuming' rewards during which the trial clock remains frozen), enabling at most 25 rewards to be harvested. A behaviourally observed work or leisure bout is defined as a temporally continuous act of working or engaging in leisure, respectively. Of course, contiguous short work or leisure bouts are externally indistinguishable from one long bout. Subjects are free to distribute leisure bouts in between individual work bouts. We shall characterise behaviour on the CHT task in detail in Chapters 3 and 6, and use it as an example labour task in Chapters 4 and 5.

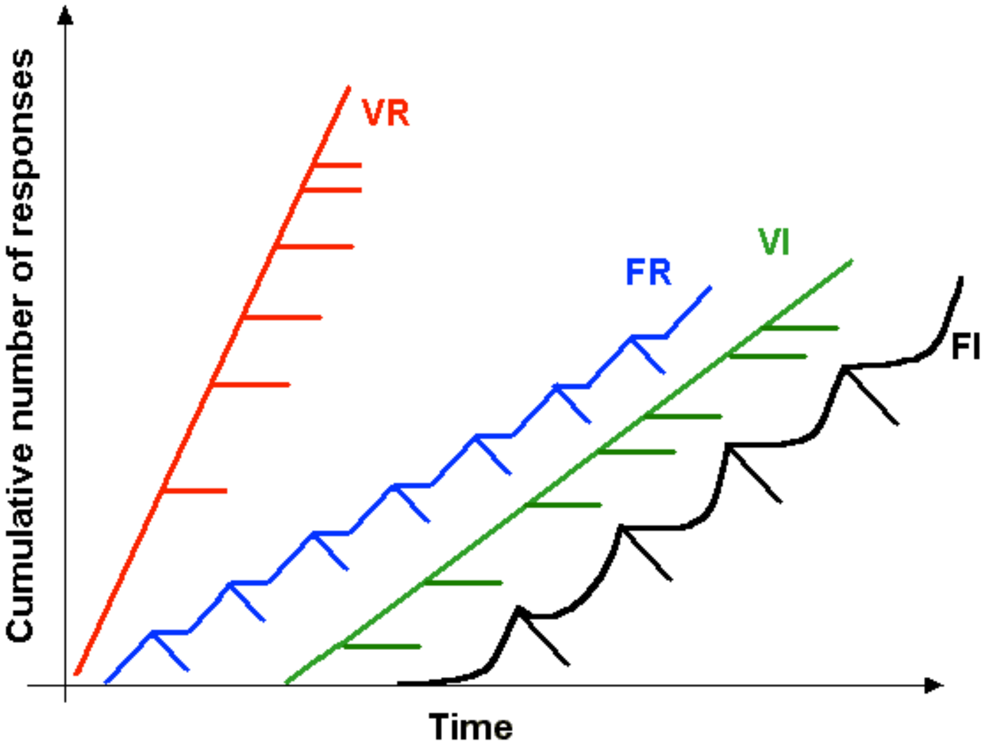


Figure 2.2: Response patterns on free-operant schedules of reinforcement. Cumulative number of responses (y axis) over time (x axis) were marked by a moving pen. The slope of each trace represents the rate of responding. Pen displacements that are large represent rewards. Note the constant response rates on variable interval and ratio schedules, and, in contrast, the scalloping response pattern in fixed interval schedules, and the post-reinforcement pauses on fixed ratio schedules. FR = fixed ratio; VR = variable ratio; FI = fixed interval; VI = variable interval.

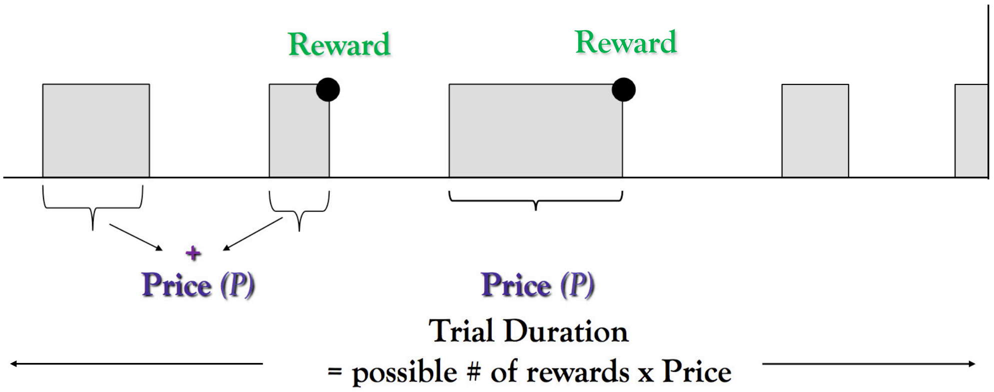


Figure 2.3: Cumulative handling time (CHT) task. Grey bars denote work (depressing a lever), white gaps show leisure. The subject must accumulate work up to an experimenter defined total period of time called the *price* (P) in order to obtain a single reward (black dot) of subjective reward intensity RI. The trial duration is 25 × price (plus 2s each time the price is attained, during which the lever is retracted so it cannot work; not shown). The reward intensity and price are held fixed within a trial.

.

2.4.3 Sweeps and random world

Most experiments collect data by sweeping, i.e. systematically increasing or decreasing from one trial to another, one of the experimenter controlled variables (e.g. food size or rate of reinforcement) while holding the others fixed. However, this could lead to behaviour on a given trial becoming contingent on the procedure of the sweep.

For example, Breton et al. (2009b) conducted an experiment in which rats worked on a VI schedule to receive brain stimulation rewards, while the reward strength or mean duration of the work requirement was swept from trial to trial. The proportion of time allocated to working was greater when the reward strength was swept holding the mean work requirement of the VI schedule fixed at a long duration, than at the predicted point of equivalence when the work requirement was swept. Breton et al. (2009b) concluded that the absence of repeated exposure to the different possible mean durations of the work-requirement led to a reduced *evaluability* of the work-requirement. Furthermore, such sweeps had the potential of inducing *anchoring* effects: the reward strength or duration on one trial would be compared against, i.e., anchored to, that on the previous trial. To avoid such anchoring effects, Breton et al. (2009b) subsequently used a *random world* experimental design, in which the reward strength and work-requirement durations were presented on a trial in random order rather than in sweeps. In order to ensure that subjects could evaluate the reward strength and work-requirement on a given 'test' trials, these trials were sandwiched in between trials with the highest stimulation at the shortest work-requirement (called 'leading' trials) and trials with the lowest stimulation at the shortest work-requirement (called 'trailing' trials). We shall describe these procedures in greater detail in Chapters 3 and 6.

2.5 Macroscopic theories of behaviour

Having reviewed different schedules or reinforcement, we now address how behaviour on them is characterised and understood, starting with traditional macroscopic characterisations.

2.5.1 Matching Law

The most famous macroscopic characterisation of behaviour in behavioural psychology comes from the observation that when two response options are concurrently available (e.g. left and right levers, or keys A and B), subjects match the ratio of their rates of responding on the two options to the ratio of their experienced reward rates. For example, when one lever is reinforced on a RI15 schedule, while the other is on a RI30 schedule, rats will press the latter lever roughly twice as fast as they will on the former. This macroscopic relationship between the ratio of response rates to the ratio of experience reward rates was studied and formalised by Herrnstein Herrnstein (1961, 1970) as the 'Matching Law'. This was later generalised to allow over or under-matching Baum (1974)

$$
\begin{aligned}
\frac{\text{Response Rate}_1}{\text{Reponse Rate}_2} &= \left(\frac{\text{Reinforcement Rate}_1}{\text{Reinforcement Rate}_2}\right)^a \\
\Rightarrow \frac{\text{Reponse Rate}_1}{\text{Reponse Rate}_1 + \text{Reponse Rate}_2} &= \frac{\text{Reinforcement Rate}_1{}^a}{\text{Reinforcement Rate}_1{}^a + \text{Reinforcement Rate}_2{}^a}
\end{aligned}
\tag{2.9}
$$

where a is the matching coefficient; $a < 1$ reflects under-matching, $a > 1$ reflects over-matching; $a = 1$ yields the original matching law. An alternative form of this generalised matching law considers the relative proportions of times allocated to responding

$$\frac{T_1}{T_2} = \left(\frac{\text{Reinforcement Rate}_1}{\text{Reinforcement Rate}_2}\right)^a$$
$$\Rightarrow \frac{T_1}{T_1 + T_2} = \frac{\text{Reinforcement Rate}_1{}^a}{\text{Reinforcement Rate}_1{}^a + \text{Reinforcement Rate}_2{}^a} \tag{2.10}$$

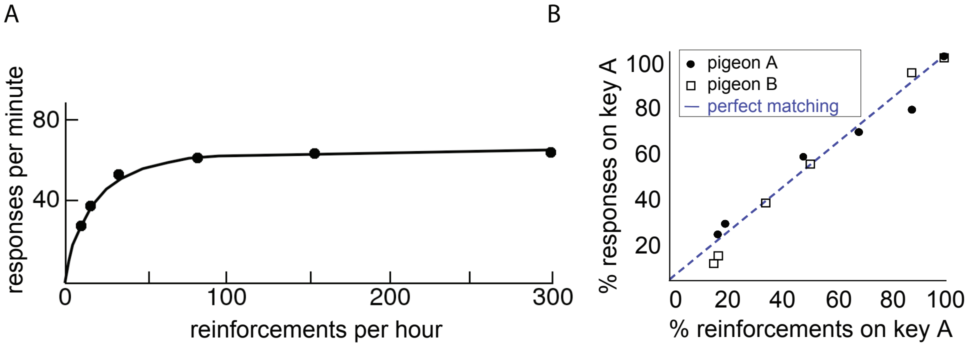


Figure 2.4: Matching Law. A) The relationship between response rate and reinforcement rate on a Variable Interval schedule, is hyperbolic. This can be seen as an instantiation of Herrnstein's Matching Law for one instrumental response. Adapted from Herrnstein (1970). B) Experimentally observed Matching Law behaviour : the proportion of pecks on key A is roughly equal to the proportion of rewards obtained on this key. Adapted from Herrnstein (1961). Note that rates in these cases are measured as overall number of responses in a session McSweeney et al. (1983), not correcting for time involved in e.g. consuming rewards.

When only one experimenter determined instrumental response is available (e.g. a box with only one lever), the rate of responding (Response Rate$_W$) is compared to the (somewhat poorly defined, and largely not considered further) 'rate' of responding (Response Rate$_L$) for all other activities, which are denoted as leisure. If a reward rate from these intrinsically beneficial alternate activities is defined (Reward Rate$_L$) and a fixed total rate of responding for all activities is assumed (Response Rate$_{total}$ = Response Rate$_W$ + Response Rate$_L$), then the rate of responding is related to the rate of reinforcement (Reinforcement Rate$_W$) via the matching law

$$\text{Response Rate}_W = \frac{\text{Response Rate}_{total} \ \cdot \text{Reinforcement Rate}_W{}^a}{\text{Reinforcement Rate}_W{}^a + \text{Reinforcement Rate}_L{}^a} \tag{2.11}$$

This gives a hyperbolic relationship between the rate of responding and the rate of reinforcement (Fig. 2.4A), which was verified experimentally Herrnstein (1970) (Fig. 2.4B). However the matching law is in practice not universally true (e.g.

Wearden and Burgess (1982); Dallery and Soto (2004); Soto et al. (2005, 2006). For example, whether or not response rates match reinforcement rates on concurrent interval schedules is sensitive Pliskoff and Fetterman (1981); Baum (1982); Boelens and Kop (1983) to whether or not a penalty for switching from between options (a change-over delay; COD; Herrnstein (1961, 1970) exists. In the absence of some penalty, whether implicitly due to the need to travel between two distant levers Baum (1982), or explicitly owing to a certain number of actions or a minimal amount of time to pass before the schedule on the switched-to option resumes Shull and Pliskoff (1967); Sugrue et al. (2004), animals often simply alternate rapidly between the two options Herrnstein (1961, 1970).

As we had mentioned in Chapter 1, the matching law is algorithmic, it describes what animals do under these schedules of reinforcement. It does not explain why animals ought to match. Matching is clearly suboptimal, since a normative subject should try to maximise returns, responding exclusively on the option with a greater reinforcement rate, rather than match response rates to reinforcement rates [2].

2.5.2 Curve shift procedure

The subjective worth or utility of a reward may be due to its properties across more than one dimension. For rewards like food pellets this could be the size, calorific value or number of food pellets. As the study of motivated behaviour gained popularity, the macroscopic paradigm used to characterise behaviour was the *curve-shift* method Miliaressis et al. (1986) (Fig. 2.5, right column). By assessing the response rates at various objective reward strengths, varying one dimension while holding others fixed, a pharmacological or lesion manipulation that changes responding overall can be disentangled from one that alters the animals motivation to gain rewarding stimulation. The parameter at which performance is half-maximal, called $M50$ provides a summary of the effect of this parameter on motivated behaviour. For instance, cocaine has been found to reduce the reward strength for half-maximal performance without altering a rat's maximum response rate Hernandez et al. (2008). It can be deduced that cocaine boosts the animals pursuit of non-maximal rewards for a given programmed rate of reinforcement.

The curve-shift method is inadequate in the sense that it cannot distinguish effects on motivation owing to influences that do not concern the reward from those

[2] There exist plenty of attempts to understand matching from a normative perspective, which we shall not review here.

that do. For instance, introducing effort costs (by adding weights to the manipulandum) should not change the utility of a food pellet, or a train of pulses in the case of brain stimulation reward. But since an animal is less motivated to work when there are greater effort costs, a higher reward parameter (e.g. greater number of pellets) must be used to compensate this and achieve the same threshold level of responding. For example, Fouriezos et al. (1990) used the curve shift procedure to asses response rates as a function of reward strength, while introducing effort costs. Increasing the weight on the lever from 0 to 45g reduced the rate of responding at the highest reward strength, but the reward strength necessary for maintaining half-maximal performance had to be increased as compensation for the added weight. The curve-shift procedure cannot distinguish effects downstream of the reward from those concerning the reward only.

2.5.3 Mountain Model

In order to overcome the inadequacies of the curve-shift procedure and disentangle at what the stage of neural processing manipulations may act, Arvanitogiannis and Shizgal (2008) developed a 3-dimensional approach (Fig. 2.5, left column) to characterising behaviour. It characterises macroscopic time allocation as a function the reward objective strength of the reward and the costs of procuring it.

The subjective worth of the reward, called the *reward intensity* (RI), is a *microscopic* utility. The transformation from objective strength to subjective reward intensity has been previously determined Gallistel and Leon (1991); Simmons and Gallistel (1994); Hamilton et al. (1985); Mark and Gallistel (1993); Leon and Gallistel (1992); Sonnenschein et al. (2003)

This is combined in scalar fashion (as in matching law accounts Baum and Rachlin (1969); Killeen (1972)) with the time lost in attaining a reward (P) ; and potential effort costs associated with it (ϵ) to define a *payoff* from working

$$R_W = \frac{RI}{P \cdot \ (1+\epsilon)} \tag{2.12}$$

the +1 in the denominator prevents payoffs from blowing off to infinity when effort costs are negligible. A payoff from alternative activities, or leisure, can be similarly defined $R_L = \frac{RI_{max}}{(1+\epsilon)P_L}$, in relation to that from working. Here, P_L is defined as the price at which, for a maximum subjective reward intensity RI_{max}, the subject allocates half the time to work, and half to leisure. The relative times allocated to working for the reward and leisure are matched to the ratio of their

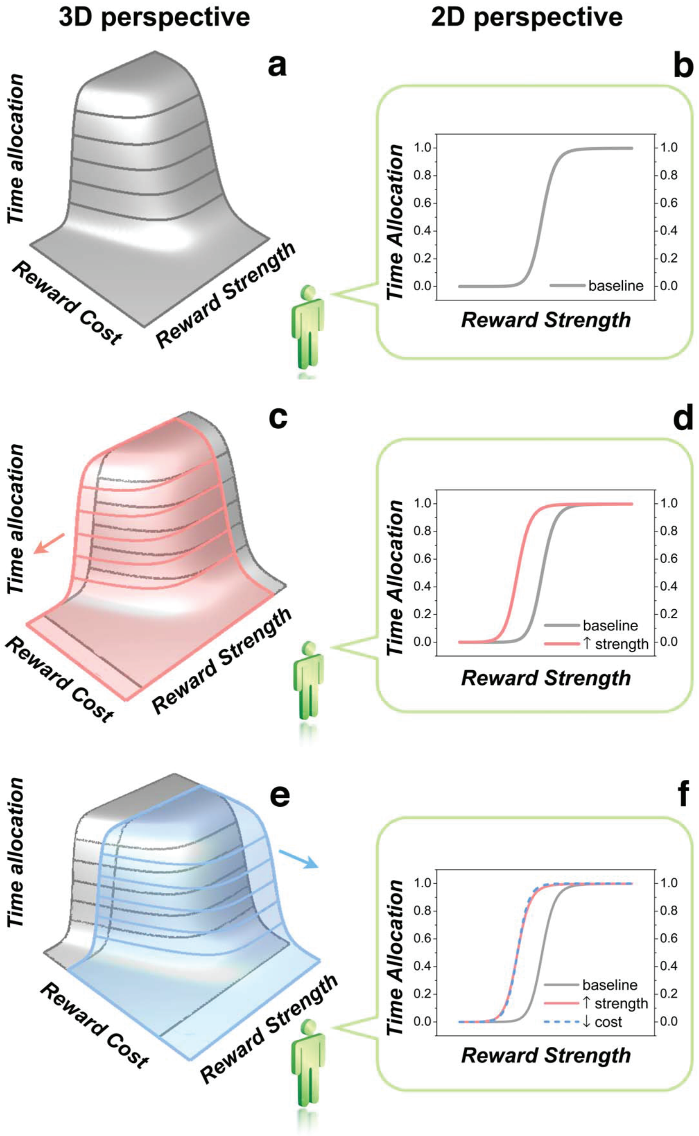


Figure 2.5: Two (curve-shift) and three-dimensional (mountain model) macroscopic approaches to characterising behaviour. The 2-dimensional curve shift procedure characterises time allocation to working as a function of reward strength only, whereas the 3-dimensional mountain model characterises it as a function of both reward strength and the cost (effort or time) of procuring it. Shifts distinguishable in the 3-dimensional mountain model (left column) are ambiguous in the 2-dimensional curve-shift characterisation (right column). The little green figure facing the reward-strength axis perceives the world in 2-dimensions. It cannot see the cost axis. It only sees the 3D structure as a 2D silhouette. Panels b,d,f show the left outlines of the silhouettes perceived by the little green figure. In panel f, the dashed blue outline of the mountain shifted along the cost axis (panel e) is superimposed on the solid pink outline of mountain shifted along the reward-strength axis (panel c). Note that although the pink and blue mountains have been shifted in orthogonal directions and their displacements are readily distinguished in the 3D representations on the left, their 2D outlines (panel f) are virtually identical and could not be distinguished in any real experiment. Adapted from Hernandez et al. (2010). The curve-shift procedure cannot distinguish effects downstream of the reward (e.g. due to costs) from those concerning the reward only, whereas the mountain model can.

respective payoffs. This makes the proportion of time allocated to working, or simply, time allocation (TA)

$$\begin{aligned} TA = &\quad TA_{min} + \left[(TA_{max} - TA_{min}) \frac{R_W{}^a}{R_W{}^a + R_L{}^a} \right] \\ = &\quad TA_{min} + \left[(TA_{max} - TA_{min}) \frac{RI^a}{RI^a + \ (\frac{P}{P_L})^a} \right] \end{aligned} \tag{2.13}$$

where a is the parameter controlling the degree of matching (compare with Eq.(2.10)). TA_{min} and TA_{max} are additional parameters accounting for when the subject works even at high/ long work requirements and when it works less than all of the time at high reward intensities.

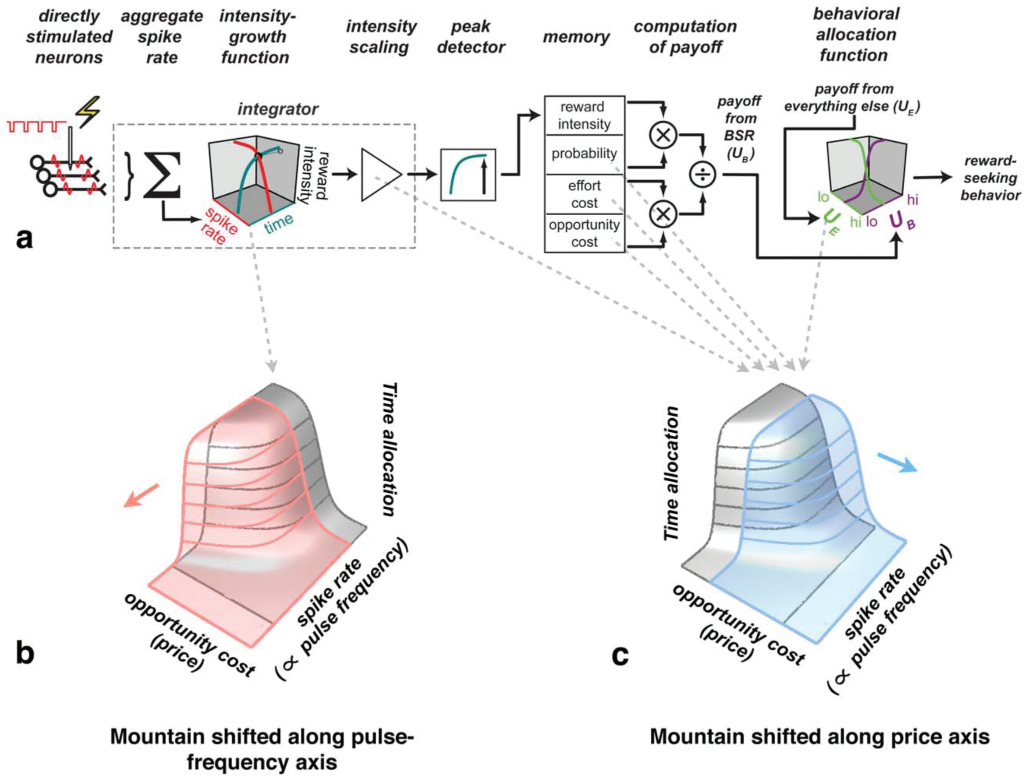
Eq. (2.13) defines a 3-dimensional relationship between the subjective reward intensity, work requirement and time allocation. This 3-dimensional relationship is called the *Mountain Model*. This macroscopic, algorithmic model provided a good fit to data on VI Arvanitogiannis and Shizgal (2008) and CHT schedules Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012); Breton (2013), which manipulated the frequency and the work requirement independently from trial to trial while holding both fixed within a trial. We shall revisit this 3-dimensional approach in Chapter 4.

2.6 Microscopic theories of behaviour

While molecular, particularly nanoscopic, characterisations of behaviour were first proposed early on in behavioural psychology Ferster and Skinner (1957); Gilbert (1958); Shull et al. (2001), they soon lost popularity to the molar approach. However, molecular, particularly microscopic approaches have been more recently used, for example to characterise the post-reinforcement pause durations of spontaneously hyperactive rats and their relationship to the reinforcement rate in VI schedules Williams et al. (2009a,b,c).

2.6.1 Continuous time Markov chains

A class of models that does make predictions at microscopic as well as macroscopic levels involves the continuous time Markov chains (CTMCs) popular in ethology



Mountain shifted along pulse-frequency axis

Mountain shifted along price axis

Figure 2.6: Mountain model. The mountain model expressed in terms of objective reward strength, here the frequency of stimulation of BSR trains. A) In the initial stages of processing, an intensity-growth function transforms the aggregate spike rate induced by the stimulation train in the directly stimulated neurons into a subjective reward-intensity. Following rescaling, the peak reward intensity is transferred to memory. The payoff from working R_W (here called U_B) is computed by taking expected value of the reward-intensity stored in memory (by multiplying the probability that a reward will be delivered when the experimenter-defined work requirement has been fulfilled, in case reward delivery is probabilistic) and scaling it by the effort cost and price (here called 'opportunity cost' by Trujillo-Pisanty et al. (2011)). Time allocation: the proportion of time allocated to working for the reward is matched to the ratio of the payoff from work R_W and the payoff R_L from leisure (here called U_E). P_L (here called P_E) is defined as the price at which, for a maximum subjective reward intensity, the subject allocates half the time to work, and half to leisure. B) Increasing the value of F_{hm} (due to a manipulation that affects the reward), the frequency at which subjective reward intensity is half-maximal, i.e, the location parameter of the intensity-growth function, shifts the 3-dimensional mountain rightward along the frequency axis of the 3D space. C) Reducing the value of the P_L parameter shifts the mountain leftwards along the price axis. Note that this could be due to the maximal reward intensity being rescaled downwards, a reduction in the probability of reward delivery, increased effort costs, or increased benefits of leisure. These are effects downstream of the processing of the reward. While the mountain model can tease apart the stage of neural processing at which a manipulation plays a role, it cannot distinguish between these different computations. Adapted from Hernandez et al. (2010).

Haccou and Meelis (1992). In these models, the entire stream of observed behaviour (every activity and its durations) can be summarized by a small set of parametric distributions, and the effect of experimental variables like reinforcement rate can be assessed with respect to how those parameters change. For example, suppose an animal's activities comprise pressing a lever, grooming and resting. Then a CTMC models how long the animal spends in a particular activity and when it switches to the next one. The probability of switching to one activity is dependent only the previous activity, and independent of the history of past activities, thus obeying the Markov property. The switching between one activity (say activity A) to another (activity B) is characterised by a transition rate parameters ($\lambda_{A\to B}$). In the simplest case, this transition rate is independent of the duration for which the first activity has been performed. If this holds for switches between all activities, then the durations for which an activity is performed are exponentially distributed, with a mean given by the inverse of the transition rate ($1/\lambda_{A\to B}$). However, the probability of switching from one activity to the next may depend on the duration of the first activity, but be independent of the history of past activities and their durations. The stream of observed behaviour is then characterised by a semi-Markov chain.

CTMCs have been used to characterise the behaviour of a variety of species, e.g. the mother-infant interaction of rhesus monkeys Dienske and Metz (1977), the mating behaviour of barbs Putters et al. (1984) as well as the behaviour of mosquitoes Peterson (1980) and juncos Wiley and Hartnett (1980). It has been used by our collaborators to provide a microscopic characterisation of work and leisure choices Breton (2013). CTMC and semi-Markov chain models are descriptive, characterising what the animal does, rather than being normative: positing why it does so. We refer the interested reader to Breton (2013) for further reviews about CTMCs and focus on normative perspectives, which shall be the flavour of theories we shall expand upon in this thesis.

2.6.2 Semi Markov Decision Process models of vigour

A normative, microscopic approach to what to do, when and how *fast* to do it (rather than how long for) was first put forward by Niv et al. (2007). They formulated this as an average-reward semi-Markov Decision Process (SMDP; which we shall review below) model of appetitive vigour in ratio and interval schedules. In this model, a subject jointly chooses both actions and how fast to do them (e.g., how fast to press a lever) depending on how much of the work requirement it has already fulfilled (e.g. the number of lever presses already made). In an average-reward setting, a normative subject attempts to maximise reward rate. This can be achieved by responding faster and enabling oneself greater opportunities to gain rewards. However, faster responses incur a vigour cost, which was assumed to be directly proportional to the vigour of responses. Subjects must thus tradeoff this vigour cost of responding quickly against the opportunities lost by responding slowly.

Given this objective, the optimal latency of actions turns out to be inversely related to the square root of the reward rate and increases with the square root of the cost of vigour. This microscopic latency between actions was then averaged over time-windows (e.g. number of lever presses in 5 mins) to yield a response rate–a macroscopic quantity. The macroscopic hyperbolic relationship between response rate and reinforcement rate in Section 2.5.1 was *derived* from this normative, microscopic relationship between response latency and reward rate. Niv et al. (2007) further proposed that the reward rate was computed or carried by the level of tonic dopamine, a question we return to in Chapter 7. They showed that, in a particular experimental setup Salamone and Correa (2002), although the macroscopic response rates may not increase with reinforcement rate (ostensibly due to the increased time spent consuming more rewards), the microscopic latency between responses could decrease–suggesting the power of the normative, microscopic approach over macroscopic ones.

Niv et al. (2007)'s model of appetitive vigour was extended to the VI schedule and to the aversive domain by Dayan (2012), particularly suggesting that serotonin may signal a negative reward rate Cools et al. (2011). In this thesis, we extend these ideas first proposed by Niv et al. (2007) to a generic, normative microscopic approach in the context of what to do, when to do it and *how to do it long for.*

2.7 Average-reward semi-Markov decision processes

We now briefly review the reinforcement learning framework of infinite-horizon (unichain) Semi-Markov Decision Process (SMDP) Puterman (2005). Unlike finite episodic tasks, which terminate, these infinitely cycle between states. A state $\vec{s}$ contains all the information necessary for making a decision. The subject's next state in the future $\vec{s'}$ depends on its current state $\vec{s}$, the action a, and the duration τ_a of that action, but is independent of all other states, actions and durations in the past. We may further assume that subjects jointly choose both the actions and their durations, as in Niv et al. (2007); Cools et al. (2011); Dayan (2012) [3].

[3] We could have simply assumed that subjects choose actions only, and subjects have no control over their durations; they persist for some duration, following which the subject transitions

A choice rule or *policy* $\pi([a, \tau_a]|\vec{s})$ specifies the subject's probability of taking action a for time τ_a in state $\vec{s}$. Under a given policy, we can define the expected reward rate, or the average reward per unit time

$$\rho^\pi = \lim_{\mathrm{T}\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{\bar{t}=0}^{\mathrm{T}-1} \; r_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'}, \tau_{a_{t'}}])\right]}{\mathrm{T}} \tag{2.14}$$

where $r_{t'}$ and $c_{t'}$ denote the benefits and costs at time points t'. Note that the expected reward rate is independent of the starting state for such unchain, ergodic chains.

The expected return or Q-value of taking action a, for duration τ_a from state $\vec{s}$ is

$$\begin{aligned} Q^\pi(\vec{s}, [a, \tau_a]) &= \mathbb{E}_\pi \left[\sum_{\bar{t}=0}^{\infty} (r_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - \rho^\pi \tau_{a_{t'}}) \; |\vec{s}_t = \vec{s}, a_t = a, \tau_{a_t} = \tau_a\right] \\ &= \hat{r}(\vec{s}, [a, \tau_a]) - \hat{c}(\vec{s}, [a, \tau_a]) - \rho^\pi \tau_a + V^\pi(\vec{s'}) \\ &= \hat{r}(\vec{s}, [a, \tau_a]) - \hat{c}(\vec{s}, [a, \tau_a]) - \rho^\pi \tau_a + \sum_{a'} \int_{\tau_{a'}} \pi([a', \tau_{a'}]|\vec{s'}) \; Q^\pi(\vec{s'}, [a', \tau_{a'}]) \; d\tau_{a'} \end{aligned} \tag{2.15}$$

where $V^\pi(\vec{s}) = \sum_a \int_{\tau_a} \pi([a, \tau_a]|\vec{s}) \; Q^\pi(\vec{s}, [a, \tau_a])$ is the *value* of state $\vec{s}$, averaged across all actions and their times. The Q values in this formulation are approximately equivalent to those obtained using shallow, explicit exponential discounting over an infinite horizon Puterman (2005); Daw and Touretzky (2002).

In RL, ρ^π is considered as the opportunity cost per unit time under policy π. It provides a point of comparison in terms of how lucrative the policy is on average. Adjusting the expected return by the reward rate means that the subject pays an automatic *opportunity cost of time* $\rho^\pi \tau_a$ for taking action a for time τ_a Niv et al. (2007); Dayan (2012); Cools et al. (2011). That is, in an environment with a positive reward rate, actions which take longer cause the subject to forego the opportunity to gain more benefits by performing other actions in that time. The higher the reward rate, the greater the amount of benefits lost due to sloth and therefore, greater the opportunity cost of time. This would be weighed against the benefits of the action. By contrast, in economics, the opportunity cost is defined instead in terms of just the next best action, a quantity that is not very meaningful in our microscopic context.

to another state. However, since this thesis considers how long to perform an action as an integral part of a choice, we consider joint choices over different actions and their durations.

Furthermore, by (approximately) maximising the Q-value, the subject also (approximately) maximises its reward rate (although in general, the converse may not be true).

While simultaneously solving Eqs. (2.14) and (2.15) for the reward rate and the Q-values, we have more unknowns than equations. As conventional, we therefore set the value of a state to 0, and solve for the Q values relative to this baseline. The Q values we consider are therefore *differential* and not the actual ones. We drop differential denotations and simply refer to them as Q-values.

The optimal choice, according to the deterministic *greedy* policy $\pi^*([a, \tau_a] \mid \vec{s})$, is the action-duration that maximises the Q-value from that state. Throughout the thesis, we shall use a stochastic, approximately-optimal policy over action-duration pairs $[a, \tau_a]$ (see Eq.(3.6) in Chapter 3). Subjects will be more likely to choose the action-duration with a greatest Q-value, but have a non-zero probability of choosing a suboptimal action-duration.

Since the policies depend on Q-values, which themselves recursively depend on the policies, except in the case of the optimal policy, we cannot solve for them in closed form. We use policy iteration to find them Sutton and Barto (1998); Puterman (2005). Starting from an initial guess, each iteration involves updating the policy while holding the Q-values fixed, and then updating the Q-values while holding the policy fixed, until they are self-consistent, i.e. policy iteration has converged. Since, to our knowledge, policy iteration for stochastic policies has not been proved to converge to a unique policy, we execute the algorithm from different starting points. All policies reported in this thesis are the *only* dynamic equilibria of policy iteration (irrespective of the starting point, they converge to the same equilibrium). An alternative would be to compute optimal Q-values (for which policy iteration provably converges to a unique equilibrium Singh (1993)) and then make stochastic choices based on them; however, this would result in policies that are inconstant with their Q-values.

2.8 Brain Stimulation Reward

We close our literature review by briefly discussing brain stimulation reward. In general, a reward is a stimulus that attracts a subject towards it, or motivates it to attain it. Brain stimulation reward (BSR), directly *electrically* stimulating the "reward neural circuitry", has historically been one of the best exemplars for studying motivation behaviour. The subject works, even to the point of exhaustion, in order to continue receiving trains of electrical pulses. This phenomenon of brain stimulation reward was discovered by Olds and Milner (1954). They observed that rats would very rapidly return to a location that had been paired with the delivery of electrical stimulation to the septal area. Their seminal discovery paved the way for the study of the neurophysiology of motivation and the neural circuits underlying reinforcements and rewards.

2.8.1 Psychology

If our goal is to understand the pure effects of reward on behaviour, BSR is one of the best candidates. In addition to affording the precise control mentioned above, BSR is a potent reward: the rat seeks out the stimulation and eagerly anticipates its availability, resisting interruption of its contact with the lever. Importantly, BSR is a 'pure' reward, which, unlike gustatory (food and drink) rewards, does not satiate Olds (1958); Olds and Olds (1958). It is also not a secondary reinforcer such as money in human experiments–money primary rewards maybe exchanged for money, but money itself cannot be directly consumed. This enables the collection of psychophysically stable data over many months. Finally, BSR has been shown to compete with Conover and Shizgal (1994b), summate with Conover and Shizgal (1994a); Conover et al. (1994); Conover and Shizgal (1994b), and substitute for Green and Rachlin (1991), gustatory rewards, demonstrating that these two at least partly share a common currency.

2.8.2 Neurobiology

The most prominently studied part of the neural circuit underlying BSR is the medial forebrain bundle (MFB). It is a large tract of axon fibres going ventrally from the olfactory tubercle to the tegmental mesencephalon. Axons from neurons with cell bodies in a variety of areas are sent through the MFB in both ascending and descending directions Nieuwenhuys et al. (1982); Veening et al. (1982). The rewarding effect of the BSR is likely due to the activation of a subset of this large, heterogeneous bundle of axons that maybe stimulated by the electrode. It may depend on the stimulation site, for example, whether it is septal or hypothalamic. Response rates were lower for septal stimulation than posterior hypothalamic stimulation. However, the current required to produce a threshold level of responding was also lower for septal stimulation. A stronger stimulation current was required to drive performance to the same level for posterior hypothalamic stimulation Hodos and Valenstein (1962).

The rewarding effect was initially hypothesised to be due to the direct activation of ascending dopamine neurons Wise (1982). However, the physiological properties of the stimulated MFB neurons involved in the behaviourally characterised rewarding effect are incompatible with that of dopamine neurons. These directly activated neurons have (i) short absolute refractory periods Yeomans and Davis (1975); Yeomans (1979); (ii) they are fine and myelinated and have much faster conduction velocities Shizgal et al. (1980); Bielajew and Shizgal (1982); and (iii) at least a subset of these directly activated neurons project in the anterior-posterior direction Bielajew and Shizgal (1986), i.e. descend rather than ascend. Although there is evidence that ventral vegmental area (VTA) dopamine neurons are eventually activated by owing the stimulation of the MFB Hernandez et al. (2006), this effect is presumably indirect. The rewarding effect of BSR is unlikely to due to the direct activation of dopamine neurons, although this is currently under experimental investigation by our collaborators.

Further, it is also unlikely that the electrical stimulation associated with BSR is directly related to the phasic firing of DA neurons Montague et al. (1996). Phasic DA is known to compute or carry a temporal difference (TD) reward prediction error–the difference between the expected and experienced returns Schultz et al. (1997). This TD error can then be used to updated the values (or expected returns) associated with a particular action or response. However, if the stimulation were to be directly associated with this TD error, then over the course of such trial and error learning, the values of all actions would grow unboundedly to infinity or, under appropriate assumptions, saturate. A subject choosing actions based on these values would not be able to distinguish between actions which lead to large rewards from those that lead to small ones, treating them all as maximal rewards. Empirical findings have shown, however, that rats will respond less to lower stimulation strengths Hodos and Valenstein (1962). These further lend support to the claim that BSR is a reward and not a reward prediction error.

The subjective rewarding effect of BSR can be precisely controlled by manipulating objective properties of the stimulation trains. The spread of the activation is determined by the current and pulse duration, the firing rate induced in the activated neurons is controlled by the frequency of pulses, and the activation period is set by the train duration.

When a train of BSR pulses at a given frequency is harvested by the rat, the cathodal pulses lead to a change in activity of the axon bundles and local somata around the electrode tip. This induces a series of action potentials (spikes) in, among other neurons, the directly stimulated MFB neurons underlying the rewarding effect of BSR. These spikes are integrated over space and time downstream of the directly stimulated neurons, rendering the injected train of pulses into an aggregate rate code. The rat's behaviour reveals that a small number of fibres (owing to a low stimulation current) firing at a high rate (owing to high frequency stimulation) yields a reward of equal magnitude to a larger number of fibres (owing to a high stimulation current) firing at a low rate (low frequency stimulation). BSR exhibits a property called duration-neglect Shizgal and Mathews (1977); Sonnenschein et al. (2003), the duration of the stimulation train is neglected and only the peak activation produced is identified by a downstream peak-detector and committed as a memory engram Gallistel et al. (1974).

If the stimulation current is held constant, the frequency of pulses (f) determines the objective strength of the brain stimulation reward. This has to be transformed to a subjective worth. The transformation from objective strength to subjective reward intensity has been previously determined Gallistel and Leon (1991); Simmons and Gallistel (1994); Hamilton et al. (1985); Mark and Gallistel (1993); Leon and Gallistel (1992); Sonnenschein et al. (2003) as power-function–which can be approximated by a logistic

$$RI = RI_{max} \frac{1}{1 + (F_{hm}/f)^g} \tag{2.16}$$

where RI_{max} is the maximal reward intensity: further increases in frequency do not significantly increase the subjective reward intensity. F_{hm} is the frequency at which RI is half its maximal worth, g is a parameter controlling the slope of the logistic reward growth function.

when BSR is used as a reward, we may define the BSR mountain model

$$\begin{aligned} TA = & \quad TA_{min} + \left[(TA_{max} - TA_{min}) \frac{RI^a}{RI^a + \; (\frac{P}{P_L})^a} \right] \\ = & \quad TA_{min} + \left[(TA_{max} - TA_{min}) \frac{(\frac{1}{1+(F_{hm}/f)^g})^a}{(\frac{1}{1+(F_{hm}/f)^g})^a + \; (\frac{P}{P_L})^a} \right] \end{aligned} \tag{2.17}$$

where TA_{min} and TA_{max} are additional parameters accounting for when the rat works even at long work requirements and when it works less than all of the time at high frequencies. The latter may be due to time involved in 'consuming' the stimulation. We shall revisit this in Chapters 3 and 6.

This model provided a good fit to data on VI Arvanitogiannis and Shizgal (2008) and CHT schedules Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012); Breton (2013) , which manipulated the frequency and the work requirement independently from trial to trial while holding both fixed within a trial.

2.8.3 Pharmacological manipulations

Given the ability to control both the neurobiology and psychophysics of BSR, it became essential, very early on in the study of BSR, to be able to appropriately and adequately characterise its effect on behaviour. A systematic study of the parameters and circuitry underlying the rewarding effect of BSR and its impact on behaviour was undertaken Gallistel et al. (1974); Edmonds et al. (1974); Edmonds and Gallistel (1974), and assessed through the impact of lesions Murray and Shizgal (1991); Waraczynski (2006) and pharmacological manipulations Franklin (1978); Hernandez et al. (2008). Here, we focus on recent ones germane to this thesis, specifically those using the mountain model in Section 2.5.3.

Macroscopic analyses, using the mountain model and BSR rewards, from pharmacological and drugs of addiction studies were used to determine at what stage of neural processing dopamine acts to affect motivated behaviour. These have revealed that an increase in the tonic release of the dopamine shifts the 3-dimensional relationships towards longer prices Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012). This would be the case if tonic dopamine increased the maximal reward intensity RI_{max}; longer prices would be required to compensate this increase. For example, the parameter P_L, the price at which time allocation is half-maximal, for a maximal reward intensity, was shifted towards longer durations by dopamine agonists: cocaine Hernandez et al. (2010) and GBR Hernandez et al. (2012) and towards shorter durations by the dopamine antagonist pimozide Trujillo-Pisanty et al. (2011). The parameter frequency at which reward intensity is half-maximal, F_{hm} remained unchanged as a result of these drug manipulations. This suggested that tonic dopamine does not change the *sensitivity* of the reward growth function–further lending support that the rewarding effect of BSR is not due to the direct stimulation of dopamine neurons. Tonic dopamine acts later in the stage of processing.

Chapter 3

The normative, microscopic approach with applications to the CHT task

3.1 Introduction

Here, we build an approximately normative, reinforcement-learning account of the labour-leisure tradeoff, in which microscopic choices approximately maximize net benefit. Our central intent is to understand the qualitative structure of the molecular behaviour of subjects, providing an account that can generalise to many experimental paradigms. We introduce a novel, generic, normative microscopic framework. We describe an example CHT task and experiment in rodents and use our model to *qualitatively* characterise key microscopic features of the data from those experiments. In Chapter 6, we use our model to quantitatively fit actual behaviour.

3.2 Task and Experiment

In Chapter 2, we briefly described the CHT task Breton et al. (2009b); Hernandez et al. (2010) in which subjects choose between working–the facile task of holding down a light lever, and engaging in leisure, i.e., resting, grooming, exploring etc (Fig. 3.1A). A brain stimulation reward (BSR; Olds and Milner (1954)) is given after the subject has accumulated work for an experimenter-defined total time-period called the *price* (P). BSR does not suffer satiation and allows precise, psychophysically stable data to be collected over many months. We show data initially reported in Breton et al. (2009a) (and subsequently in Breton (2013); Solomon et al.).

The objective strength of the BSR is the frequency of electrical stimulation pulses applied to the medial forebrain bundle. As discussed in Chapter 2, this is assumed to have a subjective worth, or *microscopic utility* called the *reward intensity* (RI, in arbitrary units). The ratio of the reward intensity to the price is called the *payoff*. Leisure is assumed to have an intrinsic subjective worth, which remains to be quantified. Throughout a task trial, the objective strength of the reward and price are held fixed. The total time a subject could work per trial is 25 times the price (plus extra time for 'consuming' rewards), enabling at most 25 rewards to be harvested. A behaviourally observed work or leisure bout is defined as a temporally continuous act of working or engaging in leisure, respectively. Of course, contiguous short work or leisure bouts are externally indistinguishable from one long bout. Subjects are free to distribute leisure bouts in between individual work bouts.

Subjects face triads of trials: 'leading', 'test', then 'trailing' (Fig. 3.2). Leading and trailing trials involve maximal and minimal reward intensities respectively, and the shortest price (we use the qualifiers "short", "long", etc. to emphasise that the price is an experimenter determined *time-period*). We analyze the sandwiched test trials, which span a range of prices and reward intensities. Leading and trailing trials allow calibration, so subjects can stably assess RI and P on test trials. Subjects tend to be at leisure on trailing trials, limiting physical fatigue. Subjects repeatedly experience each test reward intensity and price over many months, and so can readily appreciate them after minimal experience on a given trial without uncertainty, as evidenced by statistically stable performance.

3.3 Molar and molecular analyses of data

The key molar statistic is the Time Allocation TA, namely the proportion of the available time for working in a test trial that the subject spends pressing the lever. Fig. 3.1B shows example TAs for a typical subject. TA increases with the reward intensity and decreases with the price (as predicted by the Mountain model discussed in Chapter 2). Conversely, a molecular analysis, shown in the *ethograms* in (Fig. 3.1C, D), assesses the detailed temporal topography of choice, recording when, and for how long, each act of work or leisure occurred (after the first acquisition of the reward in the trial, i.e., after the 'pink/dark grey' lever

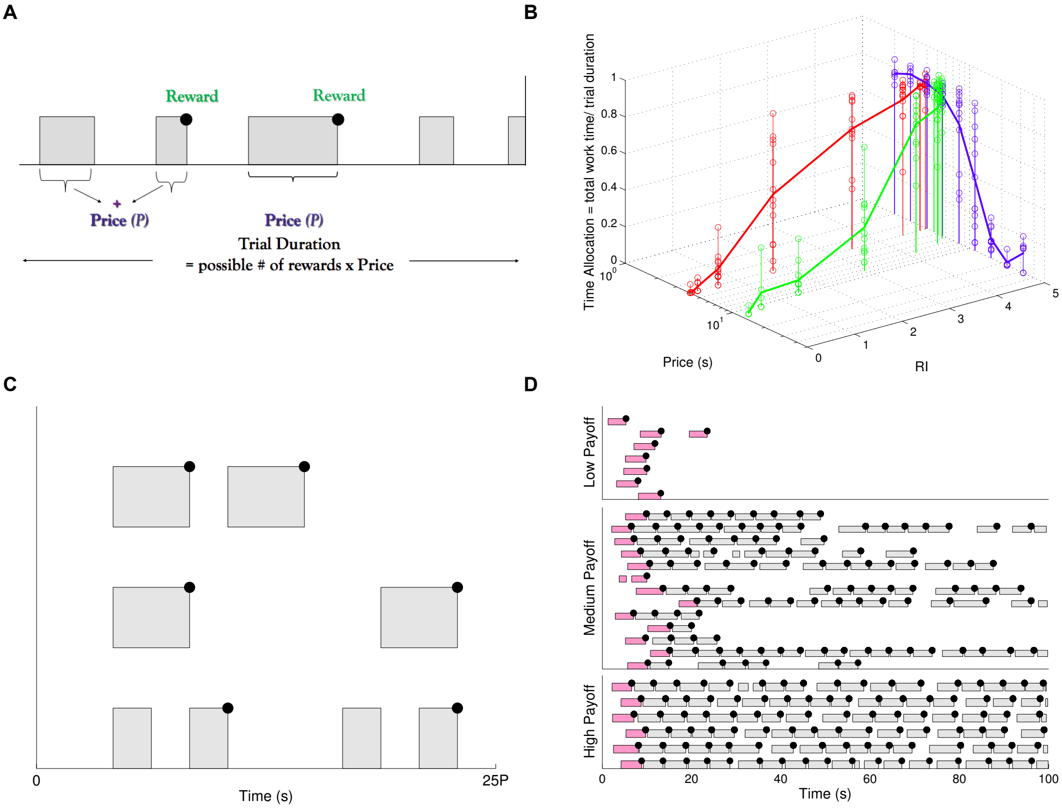


Figure 3.1: Task and key features of the data. A) Cumulative handling time (CHT) task. Grey bars denote work (depressing a lever), white gaps show leisure. The subject must accumulate work up to a total period of time called the *price* (P) in order to obtain a single reward (black dot) of subjective reward intensity RI. The trial duration is 25 × price (plus 2s each time the price is attained, during which the lever is retracted so it cannot work; not shown). The reward intensity and price are held fixed within a trial. B) Molar time allocation (TA) functions of a typical subject as a function of reward intensity and price. Red/grey curves: effect of reward intensity, for a fixed short price; blue/dark grey curves: effect of price, for a fixed high reward intensity; green/light grey curves: joint effect on time allocation of reward intensity and price. C) A molecular analysis may reveal different microstructures of working and engaging in leisure. The three rows show three different hypothetical trials. All three microstructures have the same molar TA, but are clearly distinguishable. D) Molecular *ethogram* showing the detailed temporal topography of working and engaging in leisure for the subject in B). Upper, middle and lower panels show low, medium and high payoffs, respectively, for a fixed, short price. Following previous reports using rat subjects, releases shorter than 1 second are considered part of the previous work bout (since subjects remain at the lever during this period). *Graphically*, this makes some work bouts *appear* longer than others. The subject mostly pre-commits to working continuously for the entire price duration. When the payoff is high, the subject works almost continuously for the entire trial, engaging in very short leisure bouts inbetween work bouts. When the payoff is low, the subject engages in a long leisure bout after receiving a reward. This leisure bout is potentially longer than the trial, whence it would be censored. The part of a trial before the reward and price are certainly known is coloured pink/dark grey and not considered further. Data collected by Yannick-Andre Breton and Rebecca Solomon and initially reported in Breton et al. (2009a).

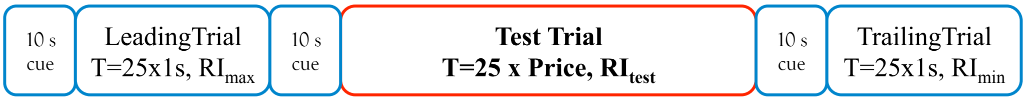


Figure 3.2: Experimental procedure: triads of trials Subjects face triads of trials: 'leading', then 'test', then 'trailing'. Throughout a trial, the reward intensity and price are all held fixed; each trial lasts $T = 25$ times the price, plus a fixed, extra time (2s) on each occasion that the price is attained, during which the lever is retracted so that subjects cannot work. This enables the subject to harvest 25 rewards if it works for the entire trial duration. The leading trial involves maximal reward intensity and the shortest (1s) price; the trailing trial involves minimal reward intensity and the shortest (1s) price. Each trial is separated by a 10s cue during which house-lights are switched on, clearly indicating that a trial has ended and a new trial shall begin. The leading and trailing trials were provided so that subjects could calibrate and adequately evaluate the reward and price on test trials. Engaging in leisure on trailing trials also ensured that the subjects would not be fatigued on test trials.

presses in Fig. 3.1D). The TA can be derived from the molecular ethogram data, but not vice-versa, since many different molecular patterns (Fig. 3.1C) share a single TA.

Qualitative characteristics of the molecular structure of the data (Fig. 3.1D) include: (i) at high payoffs, subjects work almost continuously, engaging in little leisure inbetween work bouts; (ii) at low payoffs, they engage in leisure all at once, in long bouts after working, rather than distributing the same amount of leisure time into multiple short leisure bouts; (iii) subjects work continuously for the entire price duration, as long as the price is not very long (as shown by an analysis conducted by Yannick-Andre Breton, Breton (2013)); (iv) the duration of leisure bouts is variable.

3.4 Micro Semi Markov Decision Process Model

We consider whether key features of the data in Fig. 3.1D might arise from the subject's making stochastic optimal control choices, i.e., ones that at least approximately maximise the expected return arising from all benefits and costs over entire trials. As discussed in Chapter 2, following Niv et al. (2007), we formulate this computational problem using the reinforcement learning framework of infinite horizon (Semi) Markov Decision Processes ((S)MDPs) Sutton and Barto (1998); Puterman (2005) (Fig. 3.3A). Subjects not only choose which action a to take, i.e. to work (W) or engage in leisure (L), but also *the duration of the action* (τ_a). They pay an automatic *opportunity cost of time*: performing an action over a longer period denies the subject the opportunity to take other actions during that period, and thus extirpates any potential benefit from those actions.

Since trials are substantially extended, we assume the subjects do not worry about the time the trial ends, and instead make choices that would (approximately) maximize their average summed microscopic utility per unit time Niv et al. (2007). Nevertheless, for comparison with the data, we still terminate each trial at 25× price, so actions can be *censored* by the end of the trial, preventing their completion. In the *Discussion* section we consider an alternate, Markov rather than semi-Markov, variant which, instead of committing to durations, makes choices between work and leisure at every moment-in-time. The infinite horizon, average-reward formulation is preferred over an episodic version because (a) the subjects in the experimental data that we model work till the end of trials as if they do not worry when the trial ends, i.e., where the horizon occurs (Fig. 3.1D); (b) the choice rules/policies (and hence, predicted behaviour) in our formulation are equivalent to those from a finite horizon version as long as the subject is not close to the horizon; (c) assuming a discount factor as in most RL work would introduce an extra free parameter in our model. In any case, this formulation is equivalent to that using shallow, explicit exponential discounting over an infinite horizon Puterman (2005); Daw and Touretzky (2002). Further, (d) as mentioned in Chapter 2, in RL, the reward-rate is considered as the opportunity cost per unit time under a policy. It provides a point of comparison in terms of how lucrative the policy is on average. The higher the reward rate, the greater the amount of potential benefits from alternate actions lost by persisting longer with a particular action and therefore, greater the opportunity cost of time. This would be weighed against the benefits of the action. By contrast, in economics, the opportunity cost is defined instead in terms of just the next best option. But when dividing time between labour and leisure, the next best option is not necessarily to work rather than engage in leisure (or vice-versa), but a different duration of leisure. The worth of this duration (and also whether or not to switch to the alternate action) is quantified by weighing the benefits of choosing it against its opportunity cost of time. Finally (e), following Niv et al. (2007), the neural representation of the average reward-rate term can be investigated using our formulation.

We discuss components of our microscopic SMDP model in turn: utility, state space, transitions and policy.

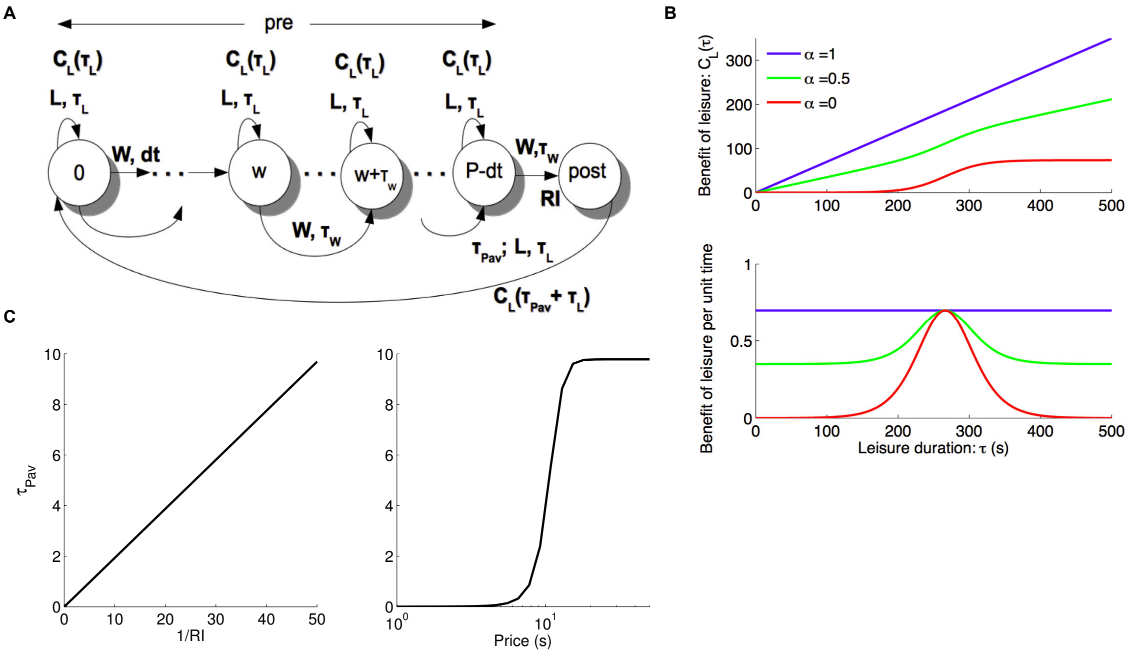


Figure 3.3: Model and leisure functions. A) The infinite horizon micro semi-Markov decision process (SMDP). States are characterised by whether they are pre- or post-reward. Subjects choose not only whether to work or to engage in leisure, but also for how long to do so. Pre-reward states are further defined by the amount of work time w that the subject has so far invested. At a pre-reward state state [pre,w], the subject can choose to work (W) for a duration τ_W or engage in leisure (L) for a duration τ_L. Working for τ_W transitions the subject to a subsequent pre-reward state [pre,$w+\tau_W$] if $w + \tau_W < P$, and to the post-reward state if $w + \tau_W \geq P$. Engaging in leisure for τ_L transitions the subject to the same state. For working, only transitions to the post-reward state are rewarded, with reward intensity RI. Engaging in leisure for τ_L has a benefit $C_L(\tau_L)$. In the post-reward state, the subject is assumed already to have been at leisure for a time τ_{Pav}, which reflects Pavlovian conditioning to the lever. By choosing to engage in instrumental leisure for a duration τ_L, it gains a microscopic benefit-of-leisure $C_L(\tau_{\text{Pav}}+\tau_L)$ and then returns to state [pre,0] at the start of the cycle whence the process repeats. B) Upper panel: canonical microscopic benefit-of-leisure functions $C_L(\cdot)$; lower panel: the net microscopic benefit-of-leisure per unit time spent in leisure. For simplicity we considered linear $C_L(\cdot)$ (blue/dark grey), whose net benefit per unit time is constant, sigmoidal $C_L(\cdot)$ (red/grey), which is initially supra-linear but eventually saturates and so has a unimodal net benefit per unit time; and a weighted sum of these two (green/light grey). See Eq.(3.1) for details. C) Time τ_{Pav} is the Pavlovian component of leisure, reflecting conditioning to the lever. It is decreasing with reward intensity (here, inversely) and increasing with price (here sigmoidally), so that it decreases with payoff.

3.4.1 Utility

The utility of the reward is RI. We assume that pressing the lever requires such minimal force that it does not incur any direct effort cost. We assume leisure to be intrinsically beneficial according to a function $C_L(\tau)$ of its duration (but formally independent of any other rewards or costs). The simplest such function is linear $C_L(\tau) = K_L\tau$ (Fig. 3.3B, upper panel blue/dark grey line), which would imply that the net utility of several short leisure bouts would be the same as a single bout of equal total length (Fig. 3.3B, lower panel, blue/dark grey line).

Alternatively, $C_L(\cdot)$ could be supra-linear (Fig. 3.3B, upper panel, red/grey curve). For this function, a single long leisure bout would be preferred to an equivalent time spent in several short bouts (Fig. 3.3B, lower panel, red/grey curve). If $C_L(\cdot)$ saturates, the marginal utility or benefit of leisure per unit time $\frac{dC_L(\tau)}{d\tau}$ will peak at an optimal bout duration. We represent this class of functions with a sigmoid, although many other non-linearities are possible. Finally, to encompass both extremes, we consider a weighted sum of linear and sigmoid $C_L(\cdot)$, with the same maximal slope (Fig. 3.3B, green/light grey curve. Linear $C_L(\cdot)$ has weight $\alpha = 1$, Eq. (3.1)).

$$C_L(\tau) = \alpha\ K_L\ \tau + (1-\alpha)\ \frac{C_{L_{max}}}{1 + \exp\Big[\ -4\frac{K_L}{C_{L_{max}}}(\tau - C_{L_{shift}})\ \Big]} \tag{3.1}$$

where $C_{L_{max}}$ and $C_{L_{shift}}$ are the maximum and shift of the sigmoidal component and $\alpha \in [0,1]$ is the weight on the linear component (see Fig. 3.3B).

Evidence from related tasks Guitart-Masip et al. (2011); Shidara et al. (1998) suggests that the leisure time will be subject to Pavlovian as well as instrumental influences Breland and Breland (1961); Dayan et al. (2006); Takikawa et al. (2002). Subjects exhibit high error rates and slow reaction times for trials with high net payoffs, even when this is only detrimental. We formalise this with a leisure time as a sum of a mandatory Pavlovian contribution τ_{Pav} (in addition to the extra time for 'consuming' rewards), and an instrumental contribution τ_L, chosen, in the light of τ_{Pav}, to optimize the expected return. The Pavlovian component comprises a mandatory pause, which is curtailed by the subject's reengagement (conditioned-response) with the reward (unconditioned-stimulus)-predicting lever (conditioned-stimulus). As we shall discuss, we postulate a Pavlovian component to account for the detrimental leisure bouts at high payoffs. We assume $\tau_{\text{Pav}} = f_{\text{Pav}}(RI, P)$ decreases with payoff – i.e., increases with price and decreases with reward intensity (Fig. 3.3C). The net microscopic benefit-of-leisure is then $C_L(\tau_L + \tau_{\text{Pav}})$ over a bout of total length $\tau_L + \tau_{\text{Pav}}$.

3.4.2 State space

The state $\vec{s} \in \mathcal{S}$ in the model contains all the information required to make a decision. This comprises a binary component (‘pre’ or ’post’), reporting whether or not the subject has just received a reward; and a real-valued component, indicating if not, how much work $w \in [0, P)$ out of the price P has been performed. Alternatively, $P - w$ is how far the subject is from the price.

3.4.3 Transitions

At state [pre, w], the subject can choose to work (W) for a duration τ_W or engage in leisure (L) for a duration τ_L. If it chooses the latter, it enjoys a benefit-of-leisure $C_L(\tau_L)$ for time τ_L, after which it returns to the same state. If the subject chooses to work up to a time that is less than the price, (i.e. $w + \tau_W < P$), then its next state is $\vec{s'} = [\text{pre}, w + \tau_W]$. However, if $w + \tau_W \geq P$, the subject gains the work reward RI and transitions to the post reward state $\vec{s'} = [\text{post}]$, consuming time $P - w$. Although subjects can *choose* work durations τ_W that go beyond the price, they cannot physically work for longer than this time, since the lever is retracted as the reward is delivered.

In the post-reward state $\vec{s} = [\text{post}]$, the subject can add *instrumental* leisure for time τ_L to the mandatory Pavlovian leisure τ_{Pav} discussed above. It receives utility $C_L(\tau_L + \tau_{\text{Pav}})$ over time $\tau_L + \tau_{\text{Pav}}$, and then transitions to state $\vec{s'} = [\text{pre}, 0]$. The cycle then repeats.

In all cases, the subject’s next state in the future $\vec{s'}$ depends on its current state $\vec{s}$, the action a, and the duration τ_a, but is independent of all other states, actions and durations in the past, making the model an SMDP. The model is molecular, as it generates the topography of lever depressing and releasing. It is microscopic as it commits to particular durations of performing actions. We therefore refer to it as a micro SMDP. In the *Discussion* section we consider an alternate, nanoscopic variant which makes choices at a finer timescale.

3.4.4 Policy evaluation

A (stochastic) policy π determines the probability of each choice of action and duration. It is assumed to be evaluated according to the average reward rate (see Eq. (3.8)).In the SMDP, the state cycles between ’pre’ and ’post’ reward. The average reward rate is the ratio of the expected total microscopic utility accumulated during a cycle to the expected total time that a cycle takes. The expected total microscopic utility comprises RI from the reward and the expected microscopic utilities of leisure (in the post and pre-reward states); the expected total time includes the price P and the expected duration engaged in leisure.

Thus, the average reward rate is (see also Eq. (3.8))

$$\rho^{\pi} = \frac{RI + \mathbb{E}_{\pi([L,\tau_L]|\text{post})}\left[C_L(\tau_{\text{Pav}} + \tau_L)\right] + \int_0^P dw\ Pr(w|h_w)\ \mathbb{E}_{\pi_{w_L}}\left[\sum_{n_{L|[\text{pre},w]}} C_L(\tau_L)\right]}{P + \mathbb{E}_{\pi([L,\tau_L]|\text{post})}[\tau_L] + \tau_{\text{Pav}} + \int_0^P dw\ Pr(w|h_w)\ \mathbb{E}_{\pi_{w_L}}\left[\sum_{n_{L|[\text{pre},w]}} \tau_L\right]} \tag{3.2}$$

Here, $\pi([L,\tau_L]|\text{post})$ and π_{w_L} are the probabilities of engaging in instrumental leisure L for time τ_L in the post-reward and pre-reward state $[\text{pre}, w]$, respectively; $\mathbb{E}_{\pi}$ is the expectation over those probabilities. $Pr(w|h_w)$ is the probability that the subject shall be in pre-reward state $[\text{pre}, w]$ given the history $h_w = \{w', \tau'_W\}$ of pre-reward states previously visited in the cycle and work durations chosen in those states, and $n_{L|[\text{pre},w]}$ is the (random) number of times the subject engages in leisure in this state ($[\text{pre}, w]$). As we shall discuss below, the third terms in both the numerator and denominator of Eq.(3.2) are dominated by the first two terms, so that we may neglect them. It is then possible, in some cases, to solve for ρ^{π} in closed form. Otherwise, we solve for it using policy iteration (see 3.7.1) without directly using Eq.(3.2).

For state $\vec{s}$ = post, the action $a = [L, \tau_L]$ of engaging in leisure for time τ_L has differential value $Q^{\pi}(\text{post}, [L, \tau_L])$ (see Eq. (3.9))that includes three terms: (i) the microscopic utility of the leisure, $C_L(\tau_L + \tau_{\text{Pav}})$; (ii) opportunity cost $-\rho^{\pi}(\tau_L + \tau_{\text{Pav}})$ for the leisure time (the rate of which is determined by the overall average reward rate); and (iii) the long-run value $V^{\pi}([\text{pre}, 0])$ of the *next* state. The value of state $\vec{s}$ is defined as

$$V^{\pi}(\vec{s}) = \sum_a \int_{\tau_a} \pi([a,\tau_a]|\vec{s})\ Q^{\pi}(\vec{s}, [a, \tau_a])$$

averaging over the actions and durations that the policy π specifies at state $\vec{s}$. Thus

$$Q^{\pi}(\text{post}, [L, \tau_L]) = C_L(\tau_L + \tau_{\text{Pav}}) - \rho^{\pi}(\tau_L + \tau_{\text{Pav}}) \ \ + V^{\pi}([\text{pre}, 0]) \tag{3.3}$$

Note the clear distinction between the immediate microscopic benefit-of-leisure $C_L(\tau_L + \tau_{\mathrm{Pav}})$ and the net benefit of leisure, given by the overall Q value.

The value $Q^\pi([\mathrm{pre}, w], [L, \tau_L])$ of engaging in leisure for τ_L in the pre-reward state has the same form, but without the contribution of τ_{Pav}, and with a different subsequent state

$$Q^\pi([\mathrm{pre}, w], [L, \tau_L]) = C_L(\tau_L) - \rho^\pi \tau_L \qquad +V^\pi([\mathrm{pre}, w]) \tag{3.4}$$

Finally, the value $Q^\pi([\mathrm{pre}, w], [W, \tau_W])$ of working for time τ_W in the pre-reward state has two components, depending on whether or not the accumulated work time $w + \tau_W$ is still less than the price (defined using a delta/indicator function as $\delta(w + \tau_W < P)$).

$$\begin{aligned} Q^\pi([\mathrm{pre}, w], [W, \tau_W]) = \delta(w + \tau_W < P)[-\rho^\pi \tau_W \qquad & +V^\pi([\mathrm{pre}, w + \tau_W])] \\ + \delta(w + \tau_W \geq P)[RI - \rho^\pi(P - w) & +V^\pi(\mathrm{post})] \end{aligned} \tag{3.5}$$

3.4.5 Policy

We assume the subject's policy π is stochastic, based on a *softmax* of the (differential) value of each choice, i.e., favouring actions and durations with greater expected returns. Random behavioural lapses make extremely long leisure or work bouts unlikely; we therefore consider a probability density $\mu_a(\tau_a)$ of choosing duration τ_a (potentially depending on the action a), which is combined with the softmax like prior and likelihood (see Subsection 3.7.1). We consider an alternative in the Discussion. For leisure bouts, we assume $\mu_L(\tau_L) = \lambda \exp(-\lambda \tau_L)$ is exponential with mean $1/\lambda = 10P$. The prior $\mu_W(\tau_W)$ for work bouts plays little role, provided its mean is not too short. This makes

$$\pi([a, \tau_a] \; |\vec{s}) = \frac{\exp\left[\beta \; Q^\pi(\vec{s}, [a, \tau_a])\right] \; \mu_a(\tau_a)}{\sum_{a'} \int_{\tau_{a'}} \exp\left[\beta \; Q^\pi(\vec{s}, [a', \tau_{a'}])\right] \; \mu_{a'}(\tau_{a'}) \; d\tau_{a'}} \tag{3.6}$$

Subjects will be more likely to choose the action with a greatest Q-value, but have a non-zero probability of choosing a suboptimal action. The inverse-temperature parameter $\beta \in [0, \infty)$ controls the degree of stochasticity in choices. Choices are completely random if $\beta = 0$, whereas $\beta \to \infty$ signifies optimal choices. We use policy iteration Sutton and Barto (1998); Puterman (2005) in order to compute policies that are self-consistent with their Q-values: these are the dynamic equilibria of policy iteration (see Subsection 3.7.1). An alternative would be to compute optimal Q-values and then make stochastic choices based on them; however, this would lead to policies that are inconsistent with their Q-values. We shall show that stochastic, approximately-optimal self-consistent choices lead to pre-commitment to working continuously for the entire price duration.

3.5 Micro SMDP policies

We first use the micro SMDP to study the issue of stochasticity, then consider the three main regimes of behaviour evident in the data in Fig. 3.1D: when payoffs are high (subjects work almost all the time), low (subjects never work) and medium (when they divide their time). Finally, we discuss the molar consequences of the molecular choices made by the SMDP. All throughout, RI, P are adopted from experimental data, while the parameters governing the benefit-of-leisure are the free parameters of interest.

3.5.1 Stochasticity

To illustrate the issues for the stochasticity of choice, we consider the case of a linear $C_L(\tau_L + \tau_{\text{Pav}}) = K_L(\tau_L + \tau_{\text{Pav}})$, and make two further simplifications: the subject does not engage in leisure in the pre-reward state (thus working for the whole price); and $\lambda = 0$, licensing arbitrarily long leisure durations. Then the Q-value of leisure is linear in τ_L, so the leisure duration distribution is exponential (see Subsection 3.7.2). The expected reward rate and mean leisure duration can be derived analytically (see Subsection 3.7.3).

As long as $RI - K_L P > \frac{1}{\beta}$

$$\begin{aligned}\rho^\pi &= \frac{\beta(RI + K_L\tau_{\text{Pav}}) - 1}{\beta(P + \tau_{\text{Pav}})} \\ \mathbb{E}[\tau_L|\text{post}] &= \frac{P + \tau_{\text{Pav}}}{\beta(RI - K_L P) - 1}\end{aligned} \tag{3.7}$$

Otherwise, if $RI - K_L P < \frac{1}{\beta}$, then $\rho^\pi \to K_L$ (middle panels in Fig. 3.4) and the subject would choose to engage in leisure for the entire trial as $\mathbb{E}[\tau_L|\text{post}] \to \infty$ (upper panels in Fig. 3.4) .

Deterministically optimal behaviour requires $\beta \to \infty$. In that case, provided $RI > K_L P$, the subject would not engage in leisure at all ($\mathbb{E}[\tau_L|\text{post}] = 0$), but would work the entire trial (interspersed by only Pavlovian leisure τ_{Pav})

with optimal reward rate $\rho^* = \frac{(RI+K_L\tau_{\text{Pav}})}{(P+\tau_{\text{Pav}})}$ (Fig. 3.4, upper and middle panels, respectively, dashed black lines). However, if $RI < K_L P$, then it would engage in leisure for the entire trial. Thus time allocation functions would be step-functions of the reward intensity and price, as shown by the dashed black lines in the lower panels of Fig. 3.4.

Of course, as is amply apparent in Fig. 3.1D, actual behaviour shows substantial variability, motivating stochastic choices, with $\beta < \infty$. Since all the other quantities can be scaled, we set $\beta = 1$ without loss of generality. This leads to smoothly changing time allocation functions, expected leisure durations and reward rates, as shown by the solid lines in Fig. 3.4. We now return to the general case ($\lambda \neq 0$, and leisure is possible in the pre-reward state).

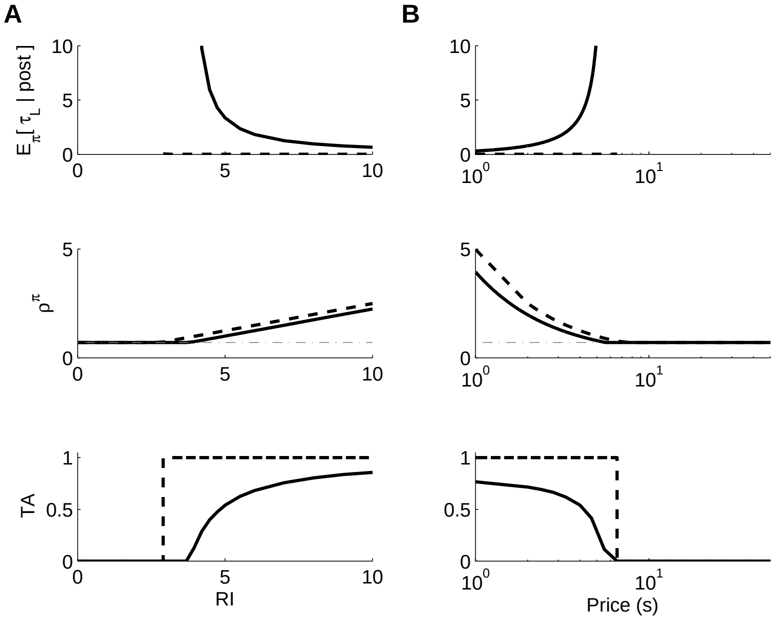


Figure 3.4: Effect of stochasticity. We use a linear microscopic benefit-of-leisure function ($\alpha = 1$) to demonstrate the effect of stochasticity on: upper panels, mean instrumental leisure, post-reward; middle panels, expected reward rate; lower panels, time allocation as a function of A) reward intensity and B) price. Solid and dashed black lines denote stochastic ($\beta = 1$) and deterministic, optimal ($\beta \to \infty$) choices, respectively. Grey dash-dotted line in middle panels are $\rho^\pi = K_L$. Time allocations are step functions under a deterministic, optimal policy but smooth under a stochastic one. Price $P = 4s$ in A, while reward intensity $RI = 4.96$ in B).

3.5.2 High payoffs

The payoff is high when the reward intensity is high, or the price is short, or both. Subjects work as much as possible, making the reward rate in Eq. (3.2)

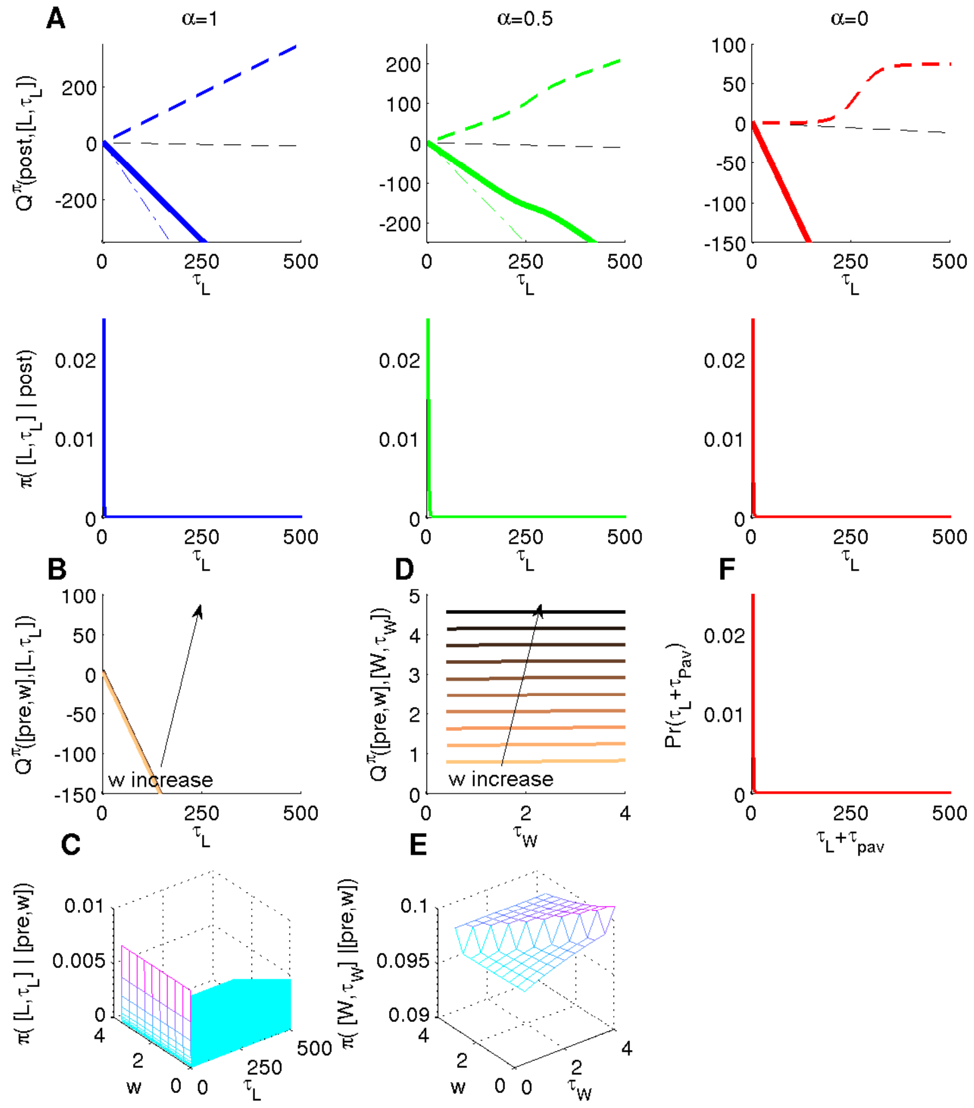


Figure 3.5: Q-values and policies for a high payoff. A) Upper and lower panels show Q-values and policies for engaging in instrumental leisure for time τ_L, respectively, in the post-reward state for three canonical $C_L(\cdot)$. In upper panels, solid bold curves show Q-values; coloured/grey dashed and dash-dotted lines show $C_L(\cdot)$ and the opportunity cost of time, respectively. Black dashed line is the linear component from the effective prior probability density for leisure time $-\lambda\tau_L$. Note the different y-axis scales. B;D) Q-values and C;E) policies for (B;C) engaging in leisure for time τ_L and (D;E) working for time τ_W in a pre-reward state [pre,w]. Light to dark colours shows increasing w, i.e., subject is furthest away from the price for light, and nearest to it for dark. F) Probability of engaging in leisure for net time $\tau_L + \tau_{\text{Pav}}$ in the post-reward state for sigmoid $C_L(\cdot)$ ($\alpha = 0$). This is the same as the lower right panel in A) but shifted by τ_{Pav}. Reward intensity, $RI = 4.96$, price $P = 4$s.

$\rho^\pi \approx \frac{(RI+C_L(\tau_{\text{Pav}}))}{(P+\tau_{\text{Pav}}})$. Since τ_{Pav} is small for high payoffs, $\rho^\pi \approx \frac{RI}{P}$ is just the payoff of the trial. The opportunity cost of leisure time $\rho^\pi(\tau_L + \tau_{\text{Pav}})$ is then linear with a very steep slope (dash-dotted line in Fig. 3.5 A, upper panels; shown here as a negative, i.e. as a cost), which dominates $C_L(\tau_L + \tau_{\text{Pav}})$ (dashed line in Fig. 3.5A upper panel), irrespective of which form it follows. The Q-value of engaging in leisure in the post-reward state then becomes the linear opportunity cost of leisure time, i.e. $Q^\pi(\text{post}, [L, \tau_L]) \to -\rho^\pi(\tau_L + \tau_{\text{Pav}})$ (solid bold line in Fig. 3.5A, upper panels).

From Eq.(3.6), the probability density of engaging in instrumental leisure for time τ_L is $\pi([L, \tau_L]\ |\text{post}) \propto \exp\left[-(\beta\rho^\pi + \lambda)\tau_L\right]$. This is an exponential distribution with very short mean $\frac{1}{\beta\rho^\pi+\lambda}$ (Fig. 3.5A, lower panels). The net post-reward leisure bout, consisting of both Pavlovian and instrumental components has the same distribution, only shifted by τ_{Pav}, i.e., a lagged exponential distribution with mean $\tau_{\text{Pav}} + \frac{1}{\beta\rho^\pi+\lambda}$ (Fig. 3.5F).

The probability of choosing to engage in leisure in a pre-reward state (i.e., after the potential resumption of working) is correspondingly also extremely small. Further, the steep opportunity cost of not working would make the distribution of any pre-reward leisure duration also be approximately a very short mean exponential (but not lagged by τ_{Pav}, Fig. 3.5B,C). Therefore when choosing to work, the duration of the work bout chosen (τ_W) barely matters (as revealed by the identical Q-values and policies for different work bout durations in Fig. 3.5D,E). That is, irrespective of whether the subject performs numerous short work bouts or pre-commits to working the whole price, it enjoys the same expected return. To the experimenter, the subject appears to work without interruption for the entire price. In sum, for high payoffs, the subject works almost continuously, with very short, lagged-exponentially distributed leisure bouts at the end of each work bout (Fig. 3.6, lowest panel). This accounts well for key feature (i) of the data.

3.5.3 Low payoffs

At the other extreme, after discovering that the payoff is very low, subjects barely work (Fig. 3.1D; top panel). Temporarily ignoring leisure consumed in the pre-reward state, the reward rate in Eq. (3.2) becomes

$$\rho^\pi \approx \frac{\mathbb{E}_{\pi([L,\tau_L]|\text{post})}\left[C_L(\tau_{\text{Pav}} + \tau_L)\right]}{P + \mathbb{E}_{\pi([L,\tau_L]|\text{post})}[\tau_L] + \tau_{\text{Pav}}}$$

shown by the dash-dotted line in Fig. 3.7 A (upper panels), and is comparatively small. The opportunity cost of time grows so slowly that the Q-value of leisure

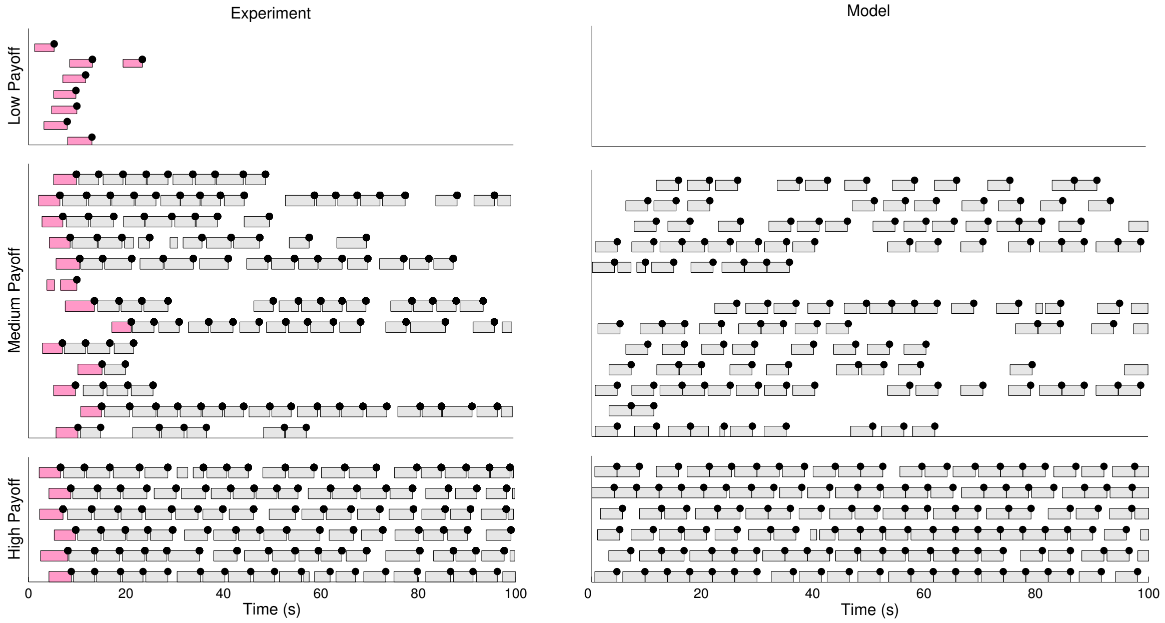


Figure 3.6: Micro SMDP model with stochastic, approximately optimal choices accounts for key features of the molecular data. Ethogram data from left: experiment and right: micro SMDP model. Upper, middle and lower panels show low, medium and high payoffs, respectively. Pink/dark grey bars show work bouts before the subject knows what the reward and price are. These are excluded from all analyses, and so do not appear on the model plot.

is dominated by the microscopic benefit-of-leisure $C_L(\tau_L + \tau_{\text{Pav}})$ (dashed curves in Fig. 3.7A, upper panels).

We showed that for linear $C_L(\cdot)$, the Q-value is linear and the leisure duration distribution is exponential (shown again in Fig. 3.7A, left panel). For initially supra-linear $C_L(\cdot)$, the Q-value becomes a bump (solid bold curve in Fig. 3.7A upper panel, centre and right). The probability of choosing to engage in instrumental leisure for time τ_L is then the exponential of this bump, which yields a unimodal, gamma-like distribution (Fig. 3.7A lower panel, centre and right). Thus for a low payoff, a subject would opt to consume leisure all at one go, if from the mode of this distribution. This accounts for key feature (ii) of the data.

The net duration of leisure in the post-reward state $\tau_L + \tau_{\text{Pav}}$ is then almost the same unimodal gamma-like distribution (Fig. 3.7F). If the Pavlovian component is increased, the instrumental component $\pi(\tau_L|\text{post})$ will decrease leaving identical the distribution of their sum $Pr(\tau_L + \tau_{\text{Pav}}|\text{post})$ (compare Fig. 3.7 A, lower right panel).

The location of the mode of the net leisure bout duration distribution (Fig. 3.7F) is crucial. For shorter prices associated with low net payoffs, this mode lies much beyond the trial duration $T = 25P$. Hence, a leisure bout drawn from this distribution would almost always exceed the trial duration, and so be *censored*, i.e. terminated by the end of the trial. Our model successfully predicts the molecular data in this condition (Fig. 3.6, upper panel). We discuss our model's predictions for long prices later (see Section 3.5.6).

The main effect of changing from partially linear to saturating $C_L(\cdot)$ is to decrease both the mean and the standard deviation of leisure bouts. The tail of the distribution (Fig. 3.7A, centre versus right panel) is shortened, since the Q-values of longer leisure bouts ultimately fail to grow.

Engaging in leisure in post- and pre-reward states are closely related. Thus, if the payoff is too low then the subject will also choose to engage in long leisure bouts in the pre-reward states (Fig. 3.7 B,C). Correspondingly, the subject will be less likely to commit to longer work times and lose the benefits of leisure (Fig. 3.7D,E). If behaviour is too deterministic, then the behavioural cycle from pre- to post-reward can fail to complete (leading to non-ergoditicty). This is not apparent in the behavioural data, so we do not consider it further.

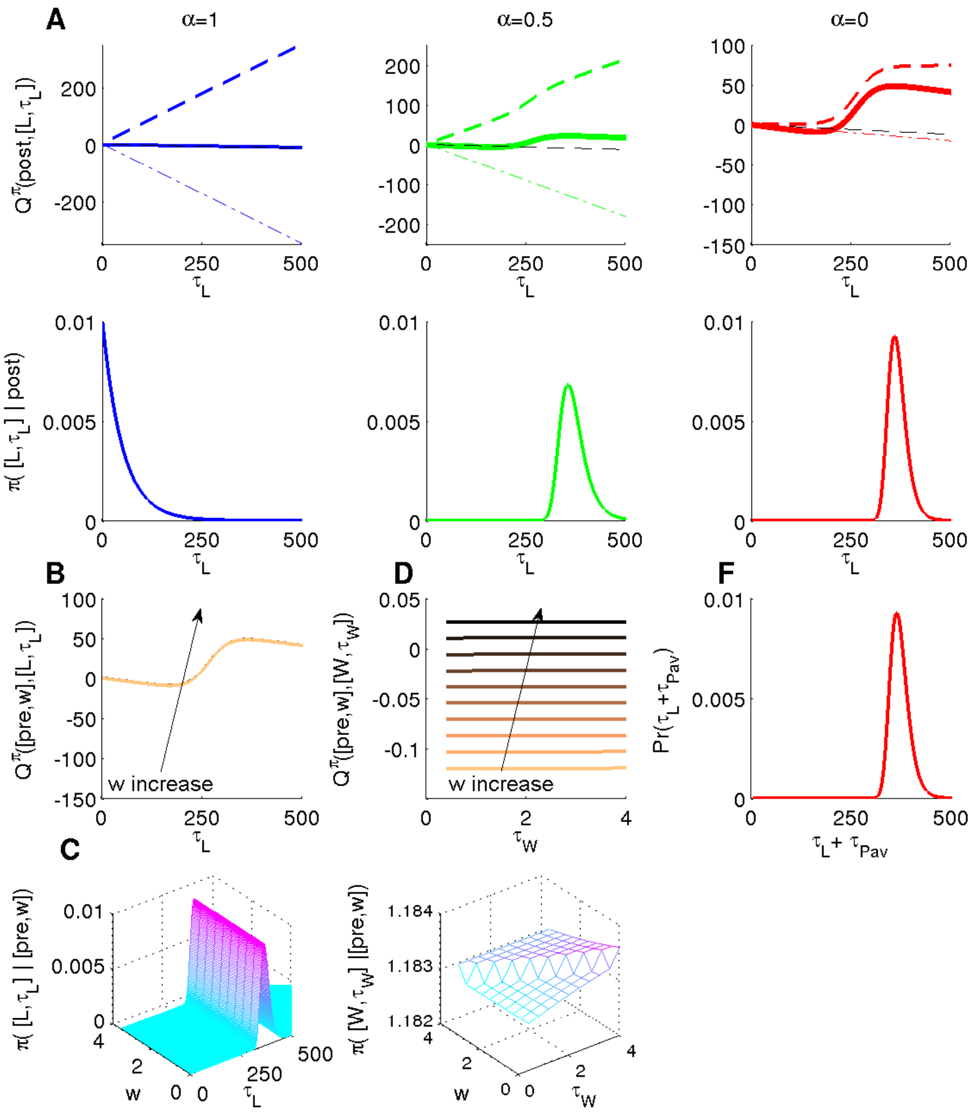


Figure 3.7: ***Q*-values and policies for a low payoff.** Panel positions as in Fig. 3.5. Reward intensity, $RI = 0.04$, price $P = 4$s. Policies in panel C expressed in $10^{-}22$.

3.5.4 Medium payoffs

The opportunity costs of time for intermediate payoffs are also intermediate. Thus the Q-value of leisure (solid bold curves in Fig. 3.8A, upper panels) depends delicately on the balance between the benefit-of-leisure and the opportunity cost (dashed and dashed-dotted lines in Fig. 3.8A, upper panels, respectively). For the sigmoidal $C_L(\cdot)$, the combination of supra- and sub-linearity leads to a bimodal distribution for leisure bouts that is a weighted sum of an exponential and a gamma-like distributions (Fig. 3.8A, lower centre and right panels; F).

Bouts drawn from the exponential component will be short. However, the mode of the gamma-like distribution lies beyond the trial duration (Fig. 3.8F), as in the low payoff case when the price is not long (Fig. 3.7F). Bouts drawn from this will thus be censored. Altogether, this predicts a pattern of several work bouts interrupted by short leisure bouts, followed by a long, censored leisure bout (Fig. 3.6A, middle panel). Occasionally, a long, but uncensored, duration can be drawn from the distribution in Fig. 3.8F. The subject would then engage in a long, uncensored leisure bout before returning to work. Our model thus also accounts well for the details of the molecular data on medium payoffs, including variable leisure bouts (key feature (iv)).

3.5.5 Pre-commitment to working continuously for the entire price duration

The micro SMDP model accounts for feature (iii) of the data, that subjects generally work continuously for the entire price duration. That is, subjects could choose to pre-commit by working for the entire price P, or divide P into multiple contiguous work bouts. In the latter case, even if Q-value of working is greater than that of engaging in leisure, the stochasticity of choice implies that subjects would have some chance of engaging in leisure instead, i.e., the pessimal choice (Fig. 3.8B,C). Pre-committing to working continuously for the entire price avoids this corruption (Fig. 3.8D,E). In Fig. 3.8E, for any given state $[\text{pre}, w]$ the probability of choosing longer work bouts τ_W increases, until the price is reached. Corruption does not occur for a deterministic, optimal policy, so pre-commitment is unnecessary. This case is then similar to that for a high payoff (Fig. 3.5D,E).

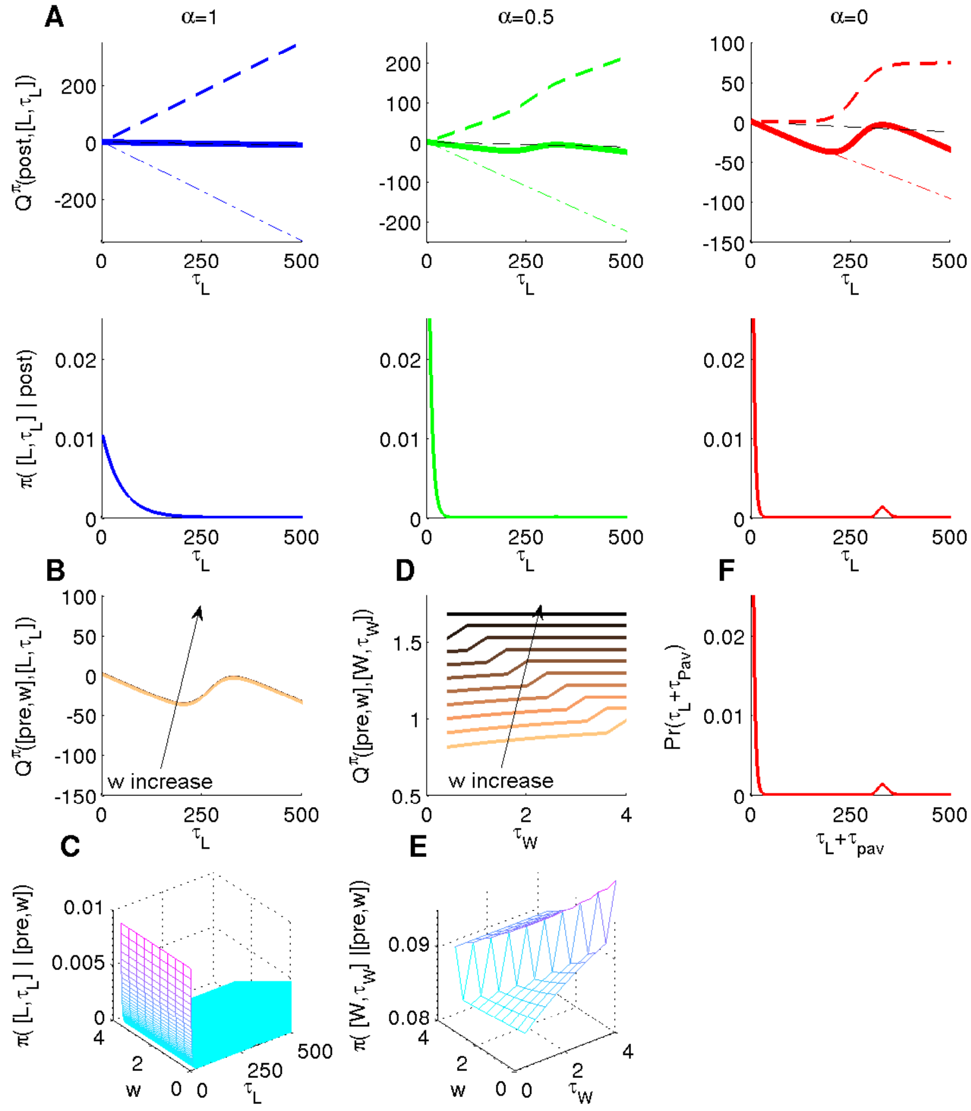


Figure 3.8: ***Q*-values and policies for a medium payoff.** Panel positions as in Fig. 3.5. Reward intensity, $RI = 1.76$, price $P = 4$s.

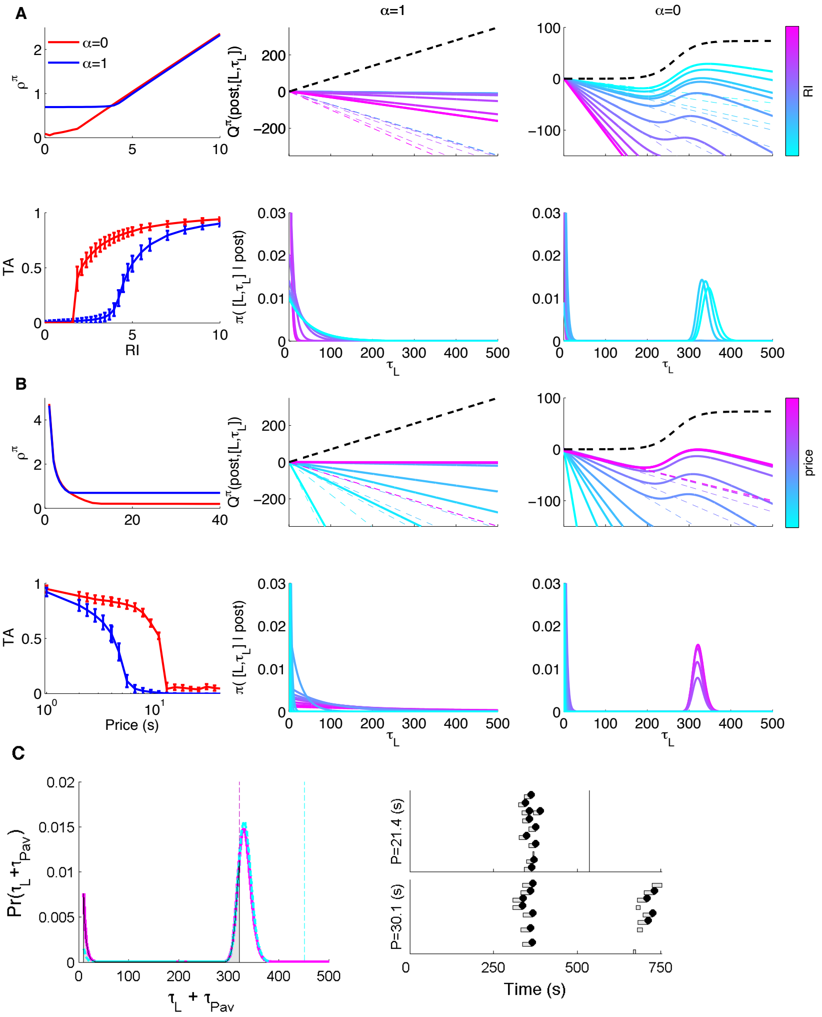


Figure 3.9: Macroscopic characterisations of behaviour. A) Effect of reward intensity for a short price ($P = 4$s). Upper and lower left panels: reward rate ρ^π and time allocation TA, respectively. Blue/dark grey and red/grey curves are for linear ($\alpha = 1$) and sigmoid ($\alpha = 0$) $C_L(\cdot)$ respectively; error bars are standard deviations. Centre and right panels: Q-values and policies for engaging in instrumental leisure for time τ_L in the post-reward state for linear (centre) and sigmoid (right) $C_L(\cdot)$. Black dashed line in upper panel shows $C_L(\cdot)$; dashed and solid bold coloured/grey curves show the opportunity cost of time and Q-values, respectively. Light blue to dark red denotes increasing reward intensity. B) Effect of price for a high reward intensity ($RI = 4.96$). Panel positions as in A). Note that the abscissa in the upper left panel is on a linear scale to demonstrate the hyperbolic relationship between reward rate and price. Light blue to dark red in the centre and right panels denotes lengthening price. C) Left: probability of engaging in leisure for net time $\tau_L + \tau_{\text{Pav}}$ in the post-reward state, and right: ethograms for two long prices (dashed cyan: $P = 30.1$s and solid magenta: $P = 21.4s$). Reward intensity is fixed at $RI = 4.96$. As the price is increased, reward rate asymptotes (B, upper left panel) and hence the mode of this probability distribution does not increase by much. The trial duration, proportional to the price does increase. Therefore more of the probability mass (grey shaded area) is included in each trial. Samples drawn from this distribution for the lower price get censored more often. For a longer price, the subject is more often observed to resume working after a long leisure bout. The effect is an increase in observed time allocation.

3.5.6 Molar behaviour from the micro SMDP

If the micro SMDP model accounts for the molecular data, integrating its output should account for the molar characterizations of behaviour that were the target of most previous modelling. Consider first the case of a fixed short price $P = 4s$, across different reward intensities (Fig. 3.9A). After an initial region in which different $C_L(\cdot)$ affect the outcome, the reward rate ρ^π in Eq. (3.2) increases linearly with the reward intensity (Fig. 3.9A, upper left panel). Consequently, the opportunity cost of time increases linearly too. If $C_L(\cdot)$ is linear, the resultant linear Q-value of leisure in the post-reward state, and hence, the mean of the exponential leisure bout duration distribution decreases (Fig. 3.9A, upper and lower centre panels, respectively). If $C_L(\cdot)$ is sigmoidal, the bump corresponding to the Q-value of leisure shifts leftwards to smaller leisure durations (Fig. 3.9A, upper right panel). Both the mode and the relative weight of the gamma-like distribution decrease as the reward intensity increases (Fig.3.9A, upper right panel). Thus, as the model smoothly transitions from low through medium to high reward intensities, time allocation increases smoothly from zero to one (Fig. 3.9A, lower left panel).

The converse holds if the price is lengthened while holding the reward intensity fixed at a high value, making the time allocation decrease smoothly (Fig. 3.9B, lower panel). The reward rate ρ^π in Eq. (3.2) decreases hyperbolically, eventually reaching an asymptote (at a level depending on $C_L(\cdot)$, Fig. 3.9B, upper left panel). For long prices, the mode of the unimodal distribution does not increase by much as the price becomes longer. However, by design of the experiment, the trial duration increases with the price. When the trial is much shorter than this mode, most long leisure bouts are censored and time allocation is near zero. As the trial duration approaches the mode, long leisure bouts are less likely to get censored (Fig. 3.9C, left panel).

We therefore make the counterintuitive prediction that as the price becomes longer, subjects will eventually be observed to resume working after a long leisure bout. Thus with longer prices, proportionally more work bouts will be observed (Fig. 3.9C, right panel). Consequently, time allocation would be observed to not decrease, and even increase with the price (see the foot of the red/grey curve in Fig. 3.9B lower left panel). Such behaviour would be observed for eventually sub-linear benefits-of-leisure. An increase in time allocation at long prices is not possible for linear $C_L(\cdot)$ (blue/dark grey curve in Fig. 3.9B lower left panel). As the price becomes longer, so does the mean of the resultant exponential leisure bout duration distribution (Fig. 3.9B centre panels) and long leisure bouts will still be censored.

In general, for the same reward intensity and price, less time is spent working for linear than saturating $C_L(\cdot)$ (compare the blue/dark grey and red/grey curves Fig. 3.9A and B, lower left panels), since linear $C_L(\cdot)$ is associated with longer leisure bouts. Thus, larger payoffs are necessary to capture the entire range of time allocation. The effect of different $C_L(\cdot)$ on the reward rate at low payoffs is more subtle (compare blue/dark grey and red/grey curves in Fig. 3.9A and B, upper left panels panels). This depends on the ratio of the expected microscopic benefit-of-leisure ($\mathbb{E}_{\pi([L,\tau_L]|\text{post})}\left[C_L(\tau_{Pav}+\tau_L)\right]$) and the expected leisure duration ($\mathbb{E}_{\pi([L,\tau_L]|\text{post})}[\tau_L]+\tau_{Pav}$) in the reward rate equation, Eq. (3.2). This is constant ($=K_L$) for a linear $C_L(\cdot)$. The latter term can be much greater for a saturating $C_L(\cdot)$, leading to a lower reward rate.

Fig. 3.9 shows that the Pavlovian component of leisure τ_{Pav} will mainly be evident at shorter prices. At high reward intensities, instrumental leisure is negligible and leisure is mainly Pavlovian. That time allocation for real subjects saturates at 1, implies that τ_{Pav} decreases with payoff, as argued.

3.6 Discussion

Real time decision-making involves choices about when and for how long to execute actions as well as well as which to perform. We studied a simplified version of this problem, considering a paradigmatic case with economic, psychological, ethological and biological consequences, namely working for explicit external rewards versus engaging in leisure for its own implicit benefit. We offered a normative, microscopic framework accounting for subjects' temporal choices, showing the rich collection of effects associated with the way that the subjective benefit-of-leisure grows with its duration.

Our microscopic formulation involved an infinite horizon Semi-Markov Decision Process (SMDP) with three key characteristics: approximate optimization of the reward rate, stochastic choices as a function of the values of the options concerned, and an assumption that, a priori, temporal choices would never be infinitely extended (owing to either lapses or the greater uncertainty that accompanies the timing of longer intervals Gibbon (1977)). The metrics associated with this last assumption had little effect on the output of the model. We may have alternately assumed that arbitrarily long durations could be chosen as frequently as short ones, but more noisily executed; we imputed all such noise to the choice rule for simplicity.

We exercised our model by examining a psychophysical paradigm called the cumulative handling time (CHT) schedule involving brain stimulation reward. The CHT controls both the (average) minimum inter-reward interval and the amount of work required to earn a reward. More common schedules of reinforcement such as Fixed Ratio, or Variable Interval control one but not the other. This makes the CHT particularly useful for studying the choice of how long to either work or engage in leisure. Nevertheless, it would be straightforward to adapt our model to treat waiting schedules such as Miyazaki et al. (2011, 2012); Fletcher (1995); Jolly et al. (1999); Ho et al. (1998); Bizot et al. (1988, 1999) or to add other facets. For instance, effort costs would lead to shorter work bouts rather than the pre-commitment to working for the duration of the price observed in the data. Costs of waiting through a delay would also lead subjects to quit waiting earlier than later. Other tasks with other work requirements could also be fitted into the model by changing the state and transition structure of the Markov chain. The main issue the CHT task poses for the model is that it is separated into episodic trials of different types making infinite horizon optimization an approximation. However, the approximation is likely benign, since the relevant trials are extended (each lasts 25 times the price), and the main effect is that work and leisure bouts can sometimes be censored at the ends of trials.

It is straightforward to account for subjects' behaviour in the CHT when payoffs are high (i.e., when the rewards are big and the price is short and the subjects work almost all the time) or low (vice-versa, when the subjects barely work at all). The medium payoff case involves a mixture of working and leisure, and is more challenging. Since the behaviour of the model is driven by relative utilities, the key quantity controlling the allocation of time is the microscopic benefit-of-leisure function. This qualitatively fits the medium payoff case when it is sigmoidal. Then, the predicted leisure duration distribution is a mixture of an exponential and a gamma-like component, with the weight on the longer, gamma-like component decreasing with payoff.

The microscopic benefit-of-leisure function reflects a subject's innate preference for the duration of leisure when only considering leisure. It is independent of the effects of all other rewards and costs. It is not the same as the Q-value of leisure, which is payoff dependent since it includes the opportunity cost of time (see Eq. (3.3)). For intuition about the consequences of different functions, consider the case of choosing between taking a long holiday all at one go, or taking multiple short holidays of the same net duration. Given a linear microscopic benefit-of-leisure function, these would be equally preferred; however, sigmoidal functions (or other functions with initially supra-linear forms) would prefer the former. A possible alternate form for the benefit-of-leisure could involve only its maximum utility or the utility at the end of a bout Diener et al. (2001); however, the systematic temporal distribution of leisure in the data suggests it is its duration which is important.

Stochasticity in choices had a further unexpected effect in tending to make subjects pre-commit to a single long work bout rather than dividing work up into multiple short bouts following on from each other. The more bouts the subject used for a single overall work duration, the more likely stochasticity would lead to a choice in favour of leisure, and thus the lower the overall reward rate. Pre-commitment to a single long duration avoids this. Our model therefore provides a novel reason for pre-commitment to executing a choice to completion: the avoidance of corruption due to stochasticity. If there was also a cost to making a decision – either from the effort expended, or from starting and stopping the action at the beginning and ends of bouts, then this effect would be further enhanced. Such switch costs would mainly influence pre-commitment during working rather than the duration of leisure, since there is exactly one behavioural switch in the latter no matter how long it lasts.

Even at very high payoffs, subjects are observed still to engage in short leisure bouts after receiving a reward – the so-called post-reinforcement pause (PRP). This is apparently not instrumentally appropriate, and so we consider PRPs to be Pavlovian. The PRP may consist of an obligatory initial component, which is curtailed by the subject's Pavlovian response to the lever. This obligatory component could be due to the enjoyment or ”consumption” of the reward. The task was set up so that instrumental rather than Pavlovian components of leisure dominate, so for simplicity we assumed the latter to be a payoff-dependent constant (rather than being a random variable). We can only model PRPs rather crudely, given the paucity of independent data to fit – but our main conclusions are only very weakly sensitive to changes.

By integrating molecular choices we derived molar quantities. A standard molar psychological account assumes that subjects match their time allocation between work and leisure to the ratio of their payoffs as in a form of the generalized matching law Herrnstein (1961, 1974); Baum (1974); McDowell (1986, 2005), see Eq.(2.10) in Chapter 2. This has been used to yield a 3-dimensional relationship known as a mountain (see Eq.(2.13)), which directly relates time allocation to objective reward strength and price Arvanitogiannis and Shizgal (2008); Hernandez et al. (2010). However, the algorithmic mountain models depend on a rather simple assignment of utility to leisure that does not have the parametric flexibility to encompass the issues on which our molecular model has focused. Those issues can nevertheless have molar signatures, as we shall extensively discuss in Chapter 4. For instance, if the microscopic benefit-of-leisure is sigmoidal, then as the price becomes very long, extended leisure bouts are less likely to get censored and so, the subject would then be observed to resume working before the end of the trial. Integrating this, at long prices, time allocation would be observed not to decrease, and even increase with the price, a prediction not made by any existing macroscopic model. Whereas animals have been previously shown to consistently work more when work-requirements are greater (eg. ostensibly owing to sunk costs Kacelnik and Marsh (2002)), the apparent anomaly discussed here only occurs at very long prices, and is unexpected from a macroscopic perspective. Our microscopic model predicts how this anomaly can be resolved. Experimentally testing whether this prediction holds true would shed light on the types of nonlinear microscopic benefit-of-leisure functions and their parameters actually used by subjects. We shall report results from experimental tests in Chapter 6.

Another standard molar (but computational) approach comes from the microeconomic theory of labour supply Frank (2005), discussed in Chapter 2. Subjects are assumed to maximize their *macroscopic* utility over combinations of work and leisure, Conover and Shizgal (2005); Battalio et al. (1981); Green et al. (1987). If work and leisure were imperfect substitutes, so leisure is more valuable given that a certain amount of work has been performed, and/or vice-versa, then perfect maximizers would choose some of each. Such macroscopic utilities do not distinguish whether leisure is more beneficial *because* of recent work e.g. owing to fatigue. We propose a microscopic benefit-of-leisure, which is independent of the recent history of work. We use stochasticity to capture the substantial variability evident at a molecular scale and thus also molar time allocation. We shall expand upon this in Chapter 4.

As we shall discuss in Chapter 5, behavioural economists have investigated real-life time allocation Battalio et al. (1981); Green et al. (1987); Kagel et al. (1995), including making predictions which seemingly contradict those made by labour supply theory accounts Camerer et al. (1997). For instance, Camerer et al. (1997) found that New York City taxi drivers gave up working for the day once they attained a target income, even when customers were in abundance (see Chapter 5). Contrary to this finding, in the experimental data we model, subjects work nearly continuously when the payoff is high rather than giving up early. Income-targeting could be used when the income earned from work can be saved and then spent on essential commodities and leisure activities Dupas and Robinson (2013). Once sufficient quantities of the latter can be guaranteed, there is no need to earn further income from work. In the experimental data we model a reward like BSR cannot be saved for future expenditures, a possible reason why we do not see income-targeting effects.

One class of models that does make predictions at molecular as well as molar levels involves the continuous time Markov chains popular in ethology Haccou and Meelis (1992). In these models, the entire stream of observed behaviour (work and leisure bouts) can be summarized by a small set of parametric distributions, and the effect of variables like payoff can be assessed with respect to how those parameters change. These models are descriptive, characterising what the animal does, rather than being normative: positing why it does so.

Our micro-SMDP model has three revealing variants. One is a nanoscopic MDP, for which choices are made at the finest possible temporal granularity rather than having determinable durations (so a long work bout would turn into a long sequence of 'work-work-work. . . ' choices). This model has a straightforward formal relationship to the micro-SMDP model Sutton et al. (1999). The distinction between these formulations cannot be made behaviourally, but may be possible in terms of their neural implementations. The second, minor alteration, restricts transitions to those between work and leisure, precluding the above long sequences of choices. The third variant is to allow a wider choice of actions, notably a 'quit', which would force the subject to remain at leisure until the end of the trial. This is simpler, and can offer a normative account of behaviour for high and low payoffs. However, in various cases, subjects resume working after long leisure bouts, whereas this should formally not be possible following quitting.

Considered more generally, quitting can be seen as an extreme example of correlation between successive leisure durations – and it is certainly possible that quantitative analyses of the data will reveal subtler dependencies. One source of these could be fatigue (or varying levels of attention or engagement). The CHT procedure (with trailing trials enabling sufficient rest) was optimised to provide stable behavioural performance over long periods. However, fatigue together with the effect of payoff might explain aspects of the microstructure of the data, especially on medium payoff trials, as we shall show in Chapter 6. Fatigue would lead to runs of work bouts interspersed with short leisure bouts, followed by a long leisure bout to reset or diminish the degree of fatigue. Note, however, that fatigue would make the benefit of leisure depend on the recent history of work.

We modelled epochs in a trial after the reward intensity and price are known for sure. The subjects repeatedly experience the reward intensity and price conditions during training over many months, and so would be able to appreciate them after minimal experience on a given trial. However, before this minimal experience, subjects face partial observability, and have to decide whether to explore (by depressing the lever to find out about the benefits of working) or exploit the option of leisure (albeit in ignorance of the price). This leads to a form of optimal stopping problem. However, the experimental regime is chosen broadly so that subjects almost always explore to get at least one sample of the reward and the price (the pink/dark grey shaded bouts in Fig. 3.1D).

Finally, having raised computational and algorithmic issues, we should consider aspects of the neural implementation of the microscopic behaviour. The neuromodulator dopamine is of particular interest. Previous macroscopic analyses from pharmacological and drugs of addiction studies have revealed that an increase in the tonic release of the neuromodulator dopamine shifts the 3-dimensional relationships towards longer prices Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012), as if, for instance, dopamine multiplies the intensity of the reward. Equally, models of instrumental vigour have posited that tonic dopamine signals the average reward rate, thus realizing the opportunity cost of time Niv et al. (2007); Cools et al. (2011); Dayan (2012). This would reduce the propensity to be at leisure. It has also suggested to affect Pavlovian conditioning Berridge (2007); Lyon and Robbins (1975) to the reward-delivering lever. Except at very high payoffs, in our model this by itself would have minimal effect, since instrumental leisure durations would be adjusted accordingly. Finally, it has been suggested as being involved in overcoming the cost of effort Salamone and Correa (2002), a factor that could readily be incorporated into the model.

3.7 Appendix

3.7.1 Supplemental Methods

We formulate our model as a infinite-horizon (unichain) Semi-Markov Decision Process (SMDP) Puterman (2005). For convenience, we repeat some of the material mentioned in Chapter 2. A state $\vec{s}$ contains all the information necessary for making a decision. The subject's next state in the future $\vec{s'}$ depends on its current state $\vec{s}$, the action a, and the duration τ_a of that action, but is independent of all other states, actions and durations in the past. We further assume subjects jointly choose both the actions and their durations, as in Niv et al. (2007); Cools et al. (2011); Dayan (2012).

A choice rule or *policy* $\pi([a, \tau_a]|\vec{s})$ specifies the subject's probability of taking action a for time τ_a in state $\vec{s}$. Under a given policy, we can define the expected reward rate, or the average reward per unit time

$$\rho^\pi = \lim_{\mathrm{T}\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{\bar{t}=0}^{\mathrm{T}-1} r_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'}, \tau_{a_{t'}}])\right]}{\mathrm{T}} \tag{3.8}$$

where $r_{t'}$ and $c_{t'}$ denote the benefits and costs at time points t'. Note that the expected reward rate is independent of the starting state.

Normatively, a subject should try to (approximately) maximise its expected return. The expected return or Q-value of taking action a, for duration τ_a from state $\vec{s}$ is

$$\begin{aligned} Q^\pi(\vec{s}, [a, \tau_a]) &= \mathbb{E}_\pi \left[\sum_{\bar{t}=0}^{\infty} \left(r_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - \rho^\pi \tau_{a_{t'}}\right) \,|s_t = s, a_t = a, \tau_{a_t} = \tau_a\right] \\ &= \hat{r}(\vec{s}, [a, \tau_a]) - \hat{c}(\vec{s}, [a, \tau_a]) - \rho^\pi \tau_a + V^\pi(\vec{s'}) \\ &= \hat{r}(\vec{s}, [a, \tau_a]) - \hat{c}(\vec{s}, [a, \tau_a]) - \rho^\pi \tau_a + \sum_{a'} \int_{\tau_{a'}} \pi([a', \tau_{a'}]|\vec{s'}) \; Q^\pi(\vec{s'}, [a', \tau_{a'}]) d\tau_{a'} \end{aligned} \tag{3.9}$$

where $V^\pi(\vec{s}) = \sum_a \int_{\tau_a} \pi([a, \tau_a]|\vec{s}) \; Q^\pi(\vec{s}, [a, \tau_a])$ is the *value* of state $\vec{s}$, averaged across all actions and their times. The subject pays an automatic *opportunity cost of time* $\rho^\pi \tau_a$ for taking action a for time τ_a Niv et al. (2007); Dayan (2012); Cools et al. (2011). The Q values in this formulation are approximately equivalent to those obtained using shallow, explicit exponential discounting over an infinite horizon Puterman (2005); Daw and Touretzky (2002).

While simultaneously solving Eqs. (3.8) and (3.9) for the reward rate and the Q-values, we have more unknowns than equations. As conventional, we therefore set the value of a state to 0, and solve for the Q values relative to this baseline. The Q values reported here are therefore *differential* and not the actual ones. We drop differential denotations and simply refer to them as Q-values.

We used a stochastic, approximately-optimal softmax policy over action-duration pairs $[a, \tau_a]$ (see Eq.(3.6)). Subjects will be more likely to choose the action-duration with a greatest Q-value, but have a non-zero probability of choosing a suboptimal action-duration. Since arbitrarily long durations should be less likely to be chosen, this was combined with a prior probability density $\mu_a(\tau_a)$ of choosing duration τ_a to yield the net policy π that generates choices. The reward rate ρ^π depends on the policy, and vice-versa (Eqs. (3.4)-(3.6)). Excluding this prior would a priori permit infinitely long leisure durations τ_L to be chosen with the same probability as short ones; these long leisure durations would significantly reduce the reward rate. On the other hand, all work durations τ_W that attain the price ($\tau_W \geq P - w$) would have an identical effect. Since the policy is over all action-durations ($[a, \tau_a]$), irrespective of whether they are of work and leisure, arbitrarily long leisure durations would have a greater effect on the reward rate than work durations. Including a prior that makes longer leisure durations less likely to be chosen normalizes the contributions of durations of work and leisure to the reward rate, affording both an equal role. We therefore employed an exponential prior for leisure $\mu_L(\tau_L)$; the exponential prior for work durations $\mu_W(\tau_W)$ did not matter as long its mean was not so short that it made attaining of the price much unlikely.

Since the policies depend on Q-values, which themselves recursively depend on the policies, except in the case of the optimal policy, we cannot solve for them in closed form. We use policy iteration to find them Sutton and Barto (1998); Puterman (2005). Starting from an initial guess, each iteration involves updating the policy while holding the Q-values fixed, and then updating the Q-values while holding the policy fixed, until they are self-consistent, i.e. policy iteration has converged. Since, to our knowledge, policy iteration for stochastic policies has not been proved to converge to a unique policy, we executed the algorithm from different starting points. All policies reported in the main text are the *only* dynamic equilibria of policy iteration (irrespective of the starting point, they converged to the same equilibrium). An alternative would be to compute optimal Q-values (for which policy iteration provably converges to a unique equilibrium Singh (1993)) and then make stochastic choices based on them; however, this would result in policies that are inconstant with their Q-values.

3.7.2 Linear benefit-of-leisure yields exponential instrumental leisure duration distributions

If $C_L(\tau_L + \tau_{\text{Pav}})$ is linear in duration τ_L, then, according to Eq. (3.4), the total Q-value of engaging in instrumental leisure in the post-reward state is also linear, $Q^\pi(\text{post}, [L, \tau_L]) = (K_L - \rho^\pi)(\tau_L + \tau_{\text{Pav}}) + V^\pi([\text{pre}, 0])$. Then, according to the softmax policy, the probability of choosing to engage in instrumental leisure for time τ_L in the post-reward state is proportional to the exponential of the Q-value (minus the $\lambda\tau_L$ contributed by the effective prior probability density, see Eq. (3.6)). This probability is $\pi([L, \tau_L] \,|\text{post}) \propto \exp\left[-\{\beta(\rho^\pi - K_L) + \lambda\}\tau_L\right]$, which is an exponential distribution with mean $\mathbb{E}[\tau_L|\text{post}] = \frac{1}{\beta(\rho^\pi - K_L)+\lambda}$. Thus, for linear $C_L(\cdot)$, instrumental leisure bout durations are always exponentially distributed with a mean which depends on the reward rate. The greater the reward rate, the shorter is the mean leisure bout.

When $C_L(\tau_L + \tau_{\text{Pav}})$ is nonlinear, it is typically not possible to derive the optimal policy analytically. We therefore report numerical results.

3.7.3 Derivation of Eq. (3.7)

We derive the result in Eq.(3.7). We consider a linear $C_L(\tau_L + \tau_{\text{Pav}}) = K_L(\tau_L + \tau_{\text{Pav}})$, and make two further simplifications: (i) the subject does not engage in leisure in the pre-reward state (and so works for the whole price when it works); and (ii) *a priori*, arbitrarily long leisure durations are possible ($\lambda = 0$). Then the reward rate in Eq. (3.2) becomes

$$\rho^\pi = \frac{RI + K_L\{\ \mathbb{E}[\tau_L|\text{post}] + \tau_{\text{Pav}}\}}{P + \mathbb{E}[\tau_L|\text{post}] + \tau_{\text{Pav}}} \tag{3.10}$$

As discussed in the *Results* section, the probability of engaging in instrumental leisure in the post-reward state is $\pi([L, \tau_L] \,|\text{post}) = \exp\left[-\{\beta(\rho^\pi - K_L)\}\tau_L\right]$, which is an exponential distribution with mean

$$\mathbb{E}[\tau_L|\text{post}] = \frac{1}{\beta(\rho^\pi - K_L)} \tag{3.11}$$

Re-arranging terms of this equation,

$$\rho^\pi = \frac{1}{\beta\ \mathbb{E}[\tau_L|\text{post}]} + K_L \tag{3.12}$$

Equating Eqs. (3.10) and (3.12) and solving for the mean instrumental leisure duration $\mathbb{E}[\tau_L|\text{post}]$, we derive

$$\mathbb{E}[\tau_L|\text{post}] = \frac{P + \tau_{\text{Pav}}}{\beta(RI - K_L P) - 1} \tag{3.13}$$

which is the second line of Eq.(3.7). This is the mean instrumental leisure duration as long as $RI - K_L P > 1/\beta$, and $\mathbb{E}[\tau_L|\text{post}] \to \infty$ otherwise. When the former condition holds, we may substitute Eq. (3.13) into Eq. (3.10) and solve for ρ^π

$$\begin{aligned} \rho^\pi &= \frac{(RI - K_L P)\ [\beta(RI + K_L \tau_{\text{Pav}}) - 1]}{(RI - K_L P)\ \beta(P + \tau_{\text{Pav}})} \\ &= \frac{\beta(RI + K_L \tau_{\text{Pav}}) - 1}{\beta(P + \tau_{\text{Pav}})} \end{aligned} \tag{3.14}$$

which is the first line of Eq. (3.7).

Chapter 4

Some work and some play: macroscopic and microscopic approaches to the labour-leisure tradeoff

4.1 Introduction

We introduced the labour-leisure tradeoff in Chapter 2. When suitably free, humans and other animals are observed to divide their limited time between work, i.e., performing employer-defined tasks remunerated by rewards such as money or food, and leisure, i.e., activities pursued for themselves that appear to confer intrinsic benefit. The division of time provides insights into these quantities and their interaction, and has been addressed by both microeconomics and behavioural psychology.

4.1.1 Task and Experiment

We consider the Cumulative Handling Time (CHT) task discussed in Chapter 2 and 3 as an example labour task (Fig. 2.3).

As we mentioned in Chapter 3, in an SMDP account, reward and leisure are both assumed to enjoy a subjective worth. We call these *microscopic utilities* to distinguish them from the *macroscopic utilities* used by traditional theories. The microscopic utility of the former is called the (subjective) *reward intensity*

(RI, in arbitrary units); the ratio of this to the price is called the payoff (or in economic nomenclature, wage rate) $R_W = \frac{RI}{P}$. For simplicity, we consider the objective price, recognising that its subjective value may differ. We explore different functional forms for the presumed microscopic utility of leisure.

4.1.2 Macroscopic and microscopic analyses

The key macroscopic statistic is the Time Allocation (*TA*): the proportion of trial time that the subject spends working Baum and Rachlin (1969). Fig.3.1B shows example *TA*s for a typical subject. As expected, the *TA* increases with reward intensity and decreases with price. A microscopic analysis, as shown by *ethograms* in (Fig.3.1C), considers the detailed temporal topography of choice, recording when and for how long each act of work or leisure occurred. Note that at intermediate payoffs, when partial allocation is most noticeable, subjects consume almost all leisure immediately after getting a reward, and then work continuously for each entire price Breton (2013).

4.2 Traditional macroscopic accounts

4.2.1 Microeconomics: Labor supply theory

We reviewed labour supply theory in Chapter 2. Here we briefly delineate it once again, but specifically focus on the issue of *substitutability* between work and leisure.

In labour supply theory Frank (2005), subjects are assumed to maximize their *macroscopic* utility by trading (i) income from working (worth RI per reward), against (ii) leisure (worth, in the simplest case, a marginal utility of K_L per unit time). Let N be the *total* number of rewards that a subject accumulates, and l be the *cumulative* amount of time spent in leisure. A commonly assumed form of *macroscopic utility function* is Arrow et al. (1961); Conover and Shizgal (2005).

$$U(l, N) = (K_L\ l^s + RI\ N^s)^{1/s} \tag{4.1}$$

where $s \in (-\infty, 1]$ is a dimensionless number representing the degree of *substitutability*, the willingness to replace rewards (or work) with leisure. Fig.4.1 shows the indifference curves (IC)–contours of equal utility. A subject is indifferent between combinations of these goods along an IC, but combinations on an IC with greater utility are preferred. The slope of an IC shows how willing a subject is to substitute one good with the other, depending on how much of each it has already consumed. Given a fixed total trial time (a budget constraint; BC Eq.(4.17)), subjects must maximise their macroscopic utilities; this occurs for the combination of goods at which the BC is tangent to an IC or is at a boundary.

Work and leisure are perfect substitutes ($s = 1$ in Eq. (4.1)) for subjects who are willing to substitute work for leisure at the same rate, irrespective of the amount of either already consumed. The ICs become (negatively sloped) straight lines. The optimum allocation is then at the boundary with all work (if returns from work exceed those from leisure, i.e. $RI > K_L\ P$) or all leisure (otherwise). This would make TA a step-function of the relative returns from work and leisure (black curves in Fig.4.1A), an outcome that is not observed empirically.

However, if work and leisure are imperfect substitutes ($-\infty < s < 1$ in Eq. (4.1)), then leisure is preferred more if the subject has worked more, and vice versa even for deterministic subjects. The slope of the IC decreases as additional amounts of leisure are consumed. The optimal combination includes both rewards (work) and leisure, making TA a smooth function of the relative returns from work and leisure (blue curves in Fig.4.2, Eq.(4.18)), as is observed empirically.

Of critical psychological importance is the relationship between the macroscopic marginal utility of leisure ($\frac{\partial U}{\partial l}$) and the amount of work so far done. For imperfect substitutability associated with the utility function of Eq.(4.1), the former depends on the latter. By contrast, we show in both deterministic and stochastic settings that this is not necessary to achieve partial allocation. The possibilities of non-determinism, which is experimentally ubiquitous, can be treated in various ways, including traditional random utility models McFadden (1984); Dagsvik et al. (2012) discussed in Chapter 2.

4.3 Normative microscopic approach: Micro SMDP model

Labor supply theory averages over the temporal topography shown in Fig.3.1C). By contrast, we follow Niv et al. (2007); Dayan (2012); Niyogi et al. (2013) in formulating a so-called micro Semi-Markov Decision Process (SMDP) Sutton and Barto (1998); Puterman (2005) (Fig. 4.3A) with actions, states, and utilities, for which policies (i.e., the stochastic choices of actions at states) are quantified by the average reward per unit time accrued over the long run. We formulated the general normative, microscopic theoretical framework in Chapter 3. Here we delineate a simplified model pertinent to the partial allocation problem.

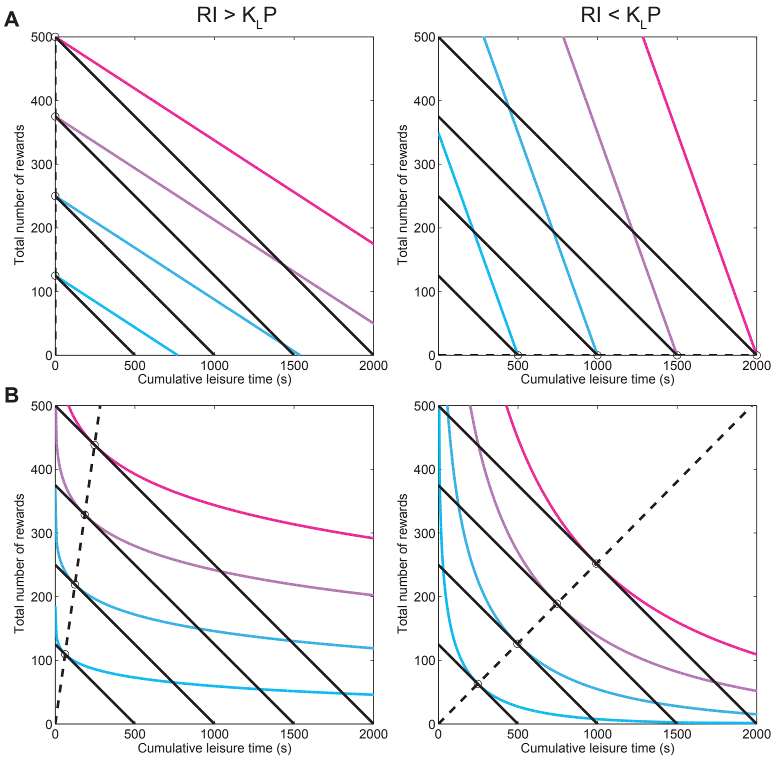


Figure 4.1: Indifference curves (ICs) of the labor supply theory model in Eq.(4.1). Left: Returns from work exceed those from leisure ($RI > K_L\ P$) and right: vice versa ($RI < K_L\ P$). Solid black lines show the budget constraint (BC): trial duration T is constant. Open circles show optimal combination of rewards and leisure for which macroscopic utility is maximised subject to BC. Dashed black lines denote the path through theoretically predicted optimal leisure and reward combinations as T is increased. A) perfect substitutability between rewards (work) and leisure ($s = 1$). Optimal combination is when the subject works all the time and claims all rewards if $RI > K_L P$, and engage in leisure all the time otherwise. B) imperfect substitutability (e.g. $s = 0.25$). Optimal combination comprises non-zero amounts of work and leisure.

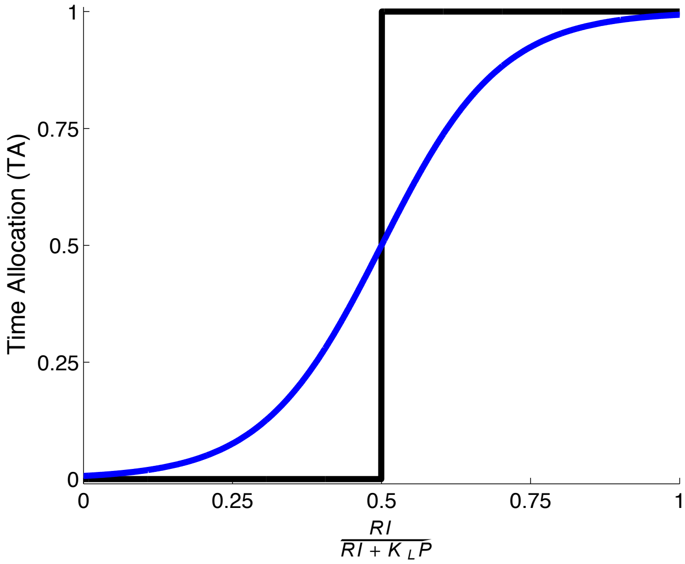


Figure 4.2: Time allocation from labour supply theory. *TA* as a function of the relative returns from work and leisure predicted by labor supply theory model in Eq. (4.1). Black and blue curves show the cases of perfect ($s = 1$) and imperfect substitutability ($s < 1$), respectively.

Actions and States: subjects choose what action (a) to do, and for how long (τ_a). The longer the duration, the more the forgone opportunity to collect rewards for other actions they could instead have been doing during that time. We simplify the task to $s =$ post- and $s =$ pre-reward states. In the former, the subject consumes leisure ($a = L$) for a freely chosen duration τ_L; then the state becomes pre-reward. If $s =$ pre, the subject works ($a = W$) for the entire price $\tau_W = P$, collects a reward and transitions to the post-reward state. The cycle then repeats. Though simply *assumed* here, working for the entire price is evident in the data (Fig.3.1D)) and can arise from optimisation in the face of stochasticity as we showed in Chapter 3.

Utilities: The *microscopic* utility of the external reward is the subjective reward intensity RI. The microscopic utility of leisure $C_L(\cdot)$ is innate and assumed to depend on its duration, but not any other reward or cost, or the amount of work performed. Based on findings in the case of discrete choices Caplin and Dean (2008); Rutledge et al. (2010); Hart et al. (2014), we expect aspects of these utilities to be discernable through neuroscience experiments; one of our main intents is to construct a framework in which such inferences are precise.

Critically, the assumptions of our microscopic utility function are different from that of the macroscopic utility function, from labor supply theory, in Eq.(4.1), which assumes that when work and leisure are imperfect substitutes, the macroscopic marginal utility of leisure ($\frac{\partial U}{\partial l}$) depends on the amount of work performed or the number of rewards received. In particular, we leave to later work considerations of fatigue or satiation, both of which can couple the microscopic utilities for working and engaging in leisure. Note, however, that this dependence is for the macroscopic utility function in Eq.(4.1); other macroscopic utility functions exist in labor supply that do not necessitate this interaction. In general, labor supply theory is concerned with the dependence in the marginal rate of substitution when work and leisure are imperfect substitutes, rather than the macroscopic marginal utilities themselves.

The simplest form for $C_L(\tau) = K_L\tau$ is linear (Fig. 4.3B, left panel blue line). This makes the total microscopic utility of several short leisure bouts the same as that of a single bout of equal total length (Fig.4.3B, right panel, blue line), and so, just by itself, implies indifference to the division of the duration of a leisure bout. Alternatively, although we had not considered this in Chapter 3, $C_L(\tau)$ could be concave (e.g., logarithmic, as in Fig. 4.3B, left panel, red curve). The marginal microscopic utility of leisure would then always decrease as more leisure is consumed (Fig. 4.3B, right panel, red curve). Subjects should then prefer several short leisure bouts to one long leisure bout. Other non-linear forms are also possible (sigmoidal, quasi concave, see Chapter 3).

As we described in Chapter 3, in an average reward SMDP model, a subject's policy (choice-rule) π is evaluated according to the average reward rate, which can be shown to be the ratio of the expected total microscopic utility accumulated during a cycle to the expected total time a cycle takes,

$$\rho^{\pi} = \frac{RI + \mathbb{E}_{\pi}\left[C_L(\tau_L)\right]}{P + \mathbb{E}_{\pi}[\tau_L]}, \tag{4.2}$$

$\mathbb{E}_{\pi}$ denotes the expected value under the distribution of leisure durations $\pi(\tau_L)$ in the post reward-state. The reward-rate increases mostly linearly with reward intensity and decreases mostly hyperbolically with price (Fig.4.8A).

The terminology in reinforcement learning (RL) Niv et al. (2007); Puterman (2005); Daw and Touretzky (2002) and optimal foraging Charnov (1976); Stephens and Krebs (1986) concerning the average reward rate differs from that in economics. As we mentioned in Chapter 2, in RL, ρ^{π} is considered as the opportunity cost per unit time under policy π. It provides a point of comparison in terms of how lucrative the policy is on average. Committing to performing an action for duration τ implies forgoing a mean total reward of $\rho^{\pi}\tau$. This would be weighed against the benefits of the action. By contrast, in economics, the opportunity cost is defined instead in terms of just the next best action, a quantity

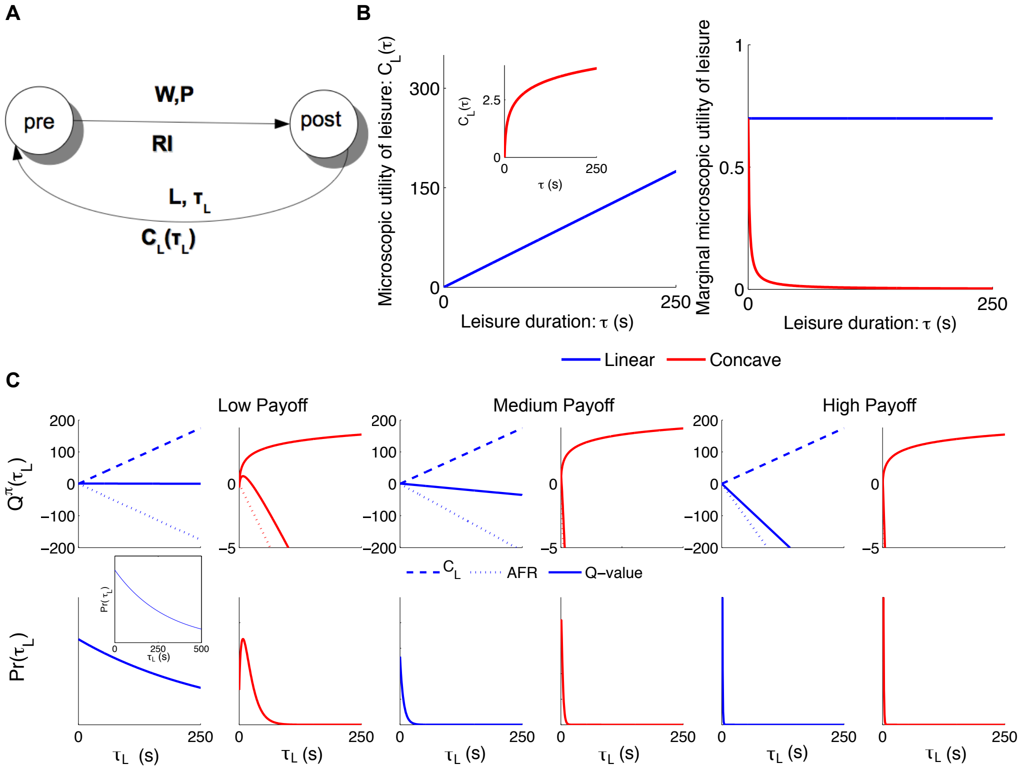


Figure 4.3: Micro SMDP model, microscopic utilities of leisure and policies. A) The infinite horizon Micro semi-Markov decision process (Micro-SMDP). States are characterised by whether they are pre- or post-reward. Subjects choose not only whether to work or to engage in leisure, but also for how long to do so. For simplicity, we assume that a subject pre-commits to working for the entire price duration when it works. Then it receives a reward of reward intensity RI and transitions to the post-reward state. In the post-reward state, by choosing to engage in leisure for a duration τ_L, it gains a microscopic benefit of leisure $C_L(\tau_L)$ and then returns to pre-reward state; this cycle repeats. B) Left: canonical microscopic utility of leisure functions $C_L(\cdot)$, right: the marginal microscopic utility of leisure. For simplicity we considered linear $C_L(\cdot)$ (blue); whose marginal utility is constant and concave (here logarithmic) $C_L(\cdot)$ (red) whose marginal utility is always decreasing. C) Q-values and policies for engaging in leisure for low, medium and high payoffs. In upper panels, dashed, dotted and solid curves show: $C_L(\cdot)$, AFR and Q-values, respectively.

that is not very meaningful in our microscopic context. To avoid confusion, in this Chapter, we refer to $\rho^\pi \tau$ as the average foregone reward (AFR) over period τ.

The (differential) Q-value (see Eq.(2.15))is defined as the expected return of taking action a for time τ_a from state s, including the immediate microscopic utility, the AFR and the differential value of the next state to which the subject transitions. For engaging in leisure for duration τ_L in the post-reward state (using simplified notation [1]), this is

$$Q^\pi(\tau_L) = C_L(\tau_L) - \rho^\pi \tau_L + V^\pi(\text{pre}) \tag{4.3}$$

which makes clear the distinction between the immediate, innate microscopic utility of leisure $C_L(\tau_L)$ and the net excess return from leisure $Q^\pi(\tau_L)$. The Q-value of working in the pre-reward state can be similarly computed. The Q-value of working in the pre-reward state then comprises: (i) the reward of reward intensity RI, (ii) an AFR $\rho^\pi P$, and (iii) the value of the post-reward state

$$Q^\pi(\text{pre}, [W, P]) = RI - \rho^\pi P + V^\pi(\text{post}). \tag{4.4}$$

Finally, we assume that the Q-values generate a policy/ choice rule for choosing leisure duration τ_L based on a stochastic *softmax*

$$\pi(\tau_L) = \frac{\exp\left[\beta\ Q^\pi(\tau_L)\right]}{\int_{\tau_{L'}\in\Gamma} \exp\left[\beta\ Q^\pi(\tau_{L'})\right]\ d\tau_{L'}} \tag{4.5}$$

where Γ is a suitable range of possible leisure durations. Durations with greater Q-values will be more likely to be chosen. The inverse-temperature parameter $\beta \in [0, \infty)$ controls the degree of stochasticity in choices: $\beta \to \infty$ signifies deterministic, optimal choices, while $\beta = 0$ leads to complete uniformity. Here we ignore the prior over durations $\mu(\tau_L)$ mentioned in Chapter 3.

4.3.1 Model policies

As discussed in Chapter 3, we can distinguish various policy regimes. If the payoff is high, then so is the reward rate; thus the AFR $\rho^\pi \tau_L$ tends to dominate the benefit of leisure $C_L(\tau_L)$ in Eq.(4.3), no matter what form the latter takes

[1] Since we define leisure to be possible in the post-state only, we simplify notation by dropping the "post" and $[L, \tau_L]$ denotations in $\pi([L, \tau_L]|\text{post})$ and $Q(\text{post}, [L, \tau_L])$, and simply use $\pi(\tau_L)$ and $Q(\tau_L)$.

(Fig. 4.3C, right panels). The probability of duration τ_L implied by the softmax policy (Eq.(4.5)) is then the exponential of a nearly linear function with a steep slope – therefore, an exponential distribution with a short mean (see Sec. 4.7.2). Thus, the subject would work almost continuously, with very short, yet stochastic, exponentially distributed leisure bouts in between work bouts.

At the other extreme, when the payoff is low, the reward rate is small. Consequently, the AFR has a very shallow slope (Fig. 4.3C, left panels). The Q-value of leisure then becomes dominated by the microscopic utility of leisure $C_L(\cdot)$. For a linear $C_L(\cdot)$, the Q-value is still linear, but with a very shallow slope, and the resulting exponential distribution has a long mean (Fig. 4.3C, left panel, blue curves). For an eventually sub-linear $C_L(\cdot)$, the Q-value becomes a unimodal bump. The exponential of this bump yields a unimodal gamma(-like) distribution. If $C_L(\cdot)$ is concave and its marginal microscopic utility does not decrease slowly, the exponential of this bump yields a unimodal gamma(-like) leisure duration distribution with a long tail (Fig. 4.3C, left panels, red curves). The leisure durations are actually gamma distributed for logarithmic $C_L(\cdot)$ (see Sec 4.7.3). The leisure durations for some other eventually sub-linear $C_L(\cdot)$ (e.g. sigmoidal) are discussed in Chapter 3.

For intermediate payoffs, the AFR has a slope that is neither too steep nor too shallow (Fig. 4.3C, middle panels). The Q-value of leisure depends delicately on the balance between the microscopic utility of leisure and this intermediate AFR.

4.4 Partial allocation with independent marginal utilities

4.4.1 Softmax policy is equivalent to maximising expected returns subject to an entropy gain

Although we use it largely for convenience, the softmax policy can be rationalised as being approximately normative. Suppose we wish to find a stochastic policy $\pi(\tau_L)$ for which the expected returns $\mathbb{E}_\pi[Q(\tau_L)]$ is maximised. Stochasticity arises optimally if it is awarded a utility. The natural way to quantify this utility is via the entropy $H(\pi) = -\mathbb{E}_\pi[\log(\pi(\tau_L))]$, making the problem one of finding

$$\begin{aligned}\pi^*(\tau_L) &= \operatorname{argmax}_\pi \left[\mathbb{E}_\pi[Q(\tau_L)] + \frac{1}{\beta}H(\pi)\right] \\ &= \operatorname{argmax}_\pi \int_{\tau_L} d\tau_L \pi(\tau_L)\left[Q(\tau_L) - \frac{1}{\beta}\log(\pi(\tau_L))\right] \qquad (4.6)\end{aligned}$$

where $1/\beta$ is a temperature parameter that trades off value for entropy. The optimum can be found by computing functional derivatives with respect to π and solving

$$\begin{aligned}\frac{\delta}{\delta\pi}\int_{\tau_L} d\tau_L \pi(\tau_L)\left[Q(\tau_L) - \frac{1}{\beta}\log(\pi(\tau_L))\right] = 0& \\ \Rightarrow \pi^*(\tau_L) \quad \propto \quad \exp\left[\beta Q(\tau_L)\right]& \qquad (4.7)\end{aligned}$$

yielding the softmax policy. Thus, making (possibly stochastic) choices according to a softmax policy is equivalent to maximising expected returns subject to an entropy gain.

4.4.2 Macroscopic utility derived from microscopic utility

To compare our account with that of labor supply theory, we construct a *macroscopic utility* function that is consistent with the microscopic choices *on average.* Consider the case that the subject works for time ω, thus completing ω/P reward and leisure cycles (we allow these to be fractional for simplicity), and is at leisure for time l. We seek to derive a macroscopic utility function $U(l,\omega)$ from a microscopic utility function $U(l,\omega) = \max_{\pi|l,\omega}[U(l,\omega,\pi)]$, such that the ultimately microscopic choices of durations, and the ultimately macroscopic time allocations are all consistent with the micro-SMDP that we have derived. Here, the notation $\pi|l,\omega$ indicates that microscopic choices of leisure duration per cycle have to be consistent with the macroscopic time devoted to leisure on average, i.e., that

$$\frac{\omega}{P}\mathbb{E}_\pi\left[\tau_L\right] = l \qquad (4.8)$$

Consider the microscopic utility

$$U(l,\omega,\pi) = \frac{\omega}{P}\left(RI + \ \mathbb{E}_\pi[C_L(\tau_L)] + \frac{1}{\beta}\ H(\pi)\right) + \frac{1}{\beta}g(l,\omega). \qquad (4.9)$$

which includes the utilities of the ω/P rewards, the benefits of leisure and the entropy, and a function $\frac{1}{\beta}g(l,\omega)$, which we will choose to enforce the average reward forgone. Enforcing Eq. (4.8) via a Lagrange multiplier $\frac{\omega}{P}\xi$, we get

$$U(l,\omega,\pi,\xi) = \frac{\omega}{P}\left(RI + \ \mathbb{E}_\pi[C_L(\tau_L)] + \frac{1}{\beta}\ H(\pi) + \xi\left(l\frac{P}{\omega} - \mathbb{E}_\pi[\tau_L]\right)\right) + \frac{1}{\beta}g(l,\omega). \tag{4.10}$$

If we optimise this utility with respect to the policy π, we get

$$\begin{aligned} 0 &= \frac{\delta}{\delta\pi}\int_{\tau_L} d\tau_L\pi(\tau_L)\left[C_L(\tau_L) - \xi\tau_L - \frac{1}{\beta}\log(\pi(\tau_L))\right] \\ \Rightarrow \pi^*(\tau_L) &\propto \exp\left[\beta(C_L(\tau_L) - \xi\tau_L)\right] \end{aligned} \tag{4.11}$$

where the Lagrange multiplier ξ is chosen to satisfy Eq. (4.8). At this optimum, $\xi = \rho^* = \frac{RI+\mathbb{E}_{\pi^*}[C_L(\tau_L)]}{P+\mathbb{E}_{\pi^*}[\tau_L]}$. That is the Lagrange multiplier or, in economic terms, the "shadow price" (marginal utility of relaxing the constraint in Eq. (4.8)) is the average reward rate ρ^*. The constructed utility function in Eq. (4.10) is evaluated at this optimum, and can now be written in terms of macroscopic quantities l and ω only as

$$U(l,\omega) = \frac{\omega}{P}\left(RI + \ \mathbb{E}_{\pi^*}[C_L(\tau_L)] + \frac{1}{\beta}\ H(\pi^*)\right) + \frac{1}{\beta}g(l,\omega) \tag{4.12}$$

4.4.3 Stochastic microscopic choices

Linear $C_L(\tau_L) = K_L\tau_L$ is equivalent to the perfect substitutability case of Eq. (4.1) with $s = 1$, for which deterministic choices exclude partial allocation. However, the derived macroscopic utility in Eq. (4.12) becomes

$$U(l,\omega) = \frac{\omega}{P}RI + \ K_L\ l + \frac{\omega}{\beta\ P}\ [\log(lP/\omega) + 1] + \frac{1}{\beta}\ g(l,\omega) \tag{4.13}$$

Its ICs have negative slopes, which, for stochastic choices ($\beta \not\to \infty$), are not constant. These changes in slope generate partial time allocations (Fig.4.4A,B), when a budget constraint (BC; solid black lines) is tangent to an IC. Including an appropriate $g(\cdot,\cdot)$ (Eq.(4.27)), at the optimum, $\mathbb{E}_\pi[\tau_L] = l^*P/\omega^* = \frac{P}{\beta(RI-K_LP)-1}$ as long as $RI-K_LP \geq \frac{1}{\beta}$, and ∞ otherwise (Eqs.(4.22),(4.23)). Thus stochasticity replaces substitutability in generating partial allocation.

For $\beta \to \infty$, optimal microscopic choices are purely deterministic. The derived

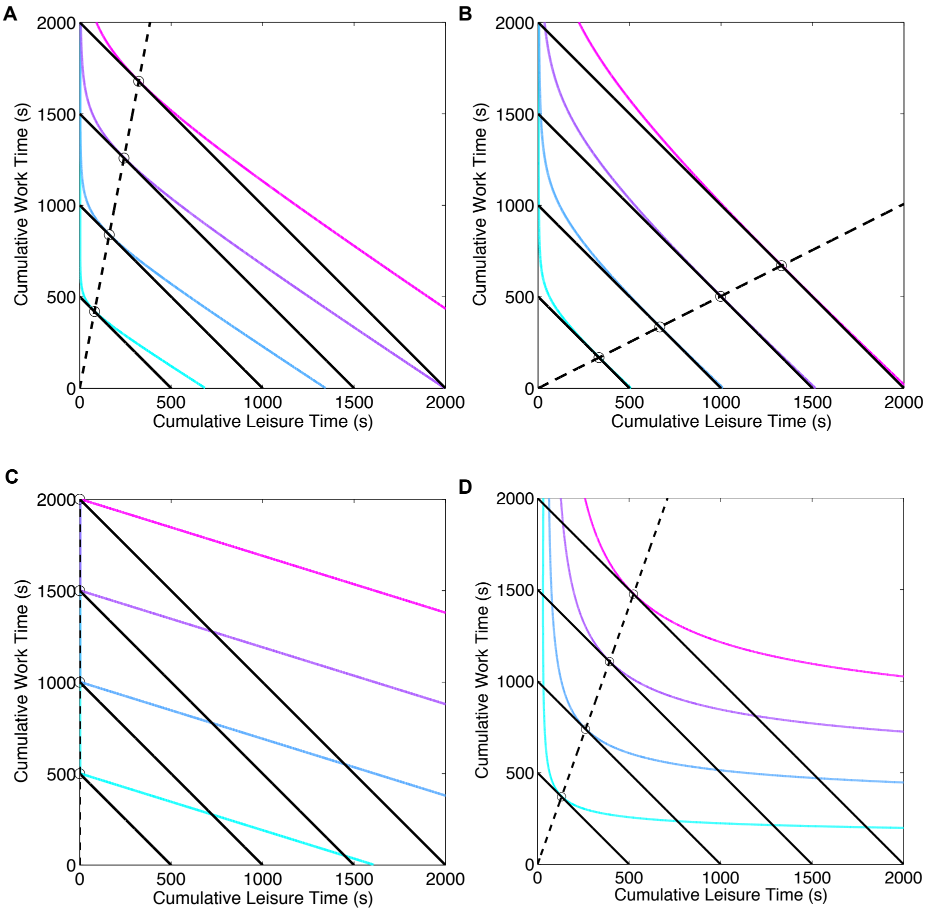


Figure 4.4: Microscopic choices yield macroscopic partial allocation even with independent marginal utilities. To compare directly with labor supply theory, we *derive* macroscopic utility functions consistent with our assumed microscopic utiities. Curves show indifference curves of the derived macroscopic utility function. Cool colours show order of increasing macroscopic utility. Solid black lines show different budget constraints $T = \omega + l$ as T is changed. Dashed black line denotes the path through theoretically predicted optimal leisure and work combinations as T is increased. A), B) Stochastic, approximately optimal microscopic choices with linear $C_L(\cdot)$ yields partial allocation (A) high and B) medium payoffs are shown). Inverse temperature $\beta = 1$. C) Deterministic, optimal microscopic choices with linear $C_L(\cdot)$ yield all-or-none allocation–work all the time if $RI > K_L P$. Inverse temperature $\beta \to \infty$. $C_L(\tau_L) = 0.7\tau_L$, Reward intensity, $RI = 9$ in A), $RI = 4.3$ in B) and C), price $P = 4$s in A-C. D) Deterministic, optimal choices with non-linear $C_L(\cdot)$ also yields partial allocation. $C_L(\tau_L) = 0.7\log(\tau_L)$, $\beta \to \infty$, $RI = 2.46$ and price $P = 4$s.

utility function in Eq.(4.13) becomes

$$U(l, \omega) = \frac{\omega}{P} RI + \ K_L \ l \tag{4.14}$$

which directly corresponds to the utility function of labor supply theory in Eq.(4.1) with $s = 1$ and would lead to total allocation to work or leisure depending on whether work or leisure is more beneficial, i.e. the sign of $RI - K_L P$ (Fig. 4.4C; compare with Fig. 4.1, upper- panels).

4.4.4 Deterministic, optimal microscopic choices

As is the case for standard labor supply theory, the assumption of stochasticity is not necessary to achieve partial allocation if the microscopic utility of leisure is a suitably non-linear function of its duration, e.g., the concave $C_L(\tau_L) = (k - 1)\log(\tau_L)$, for $k > 1$ (Fig. 4.3B, red) [2]. Importantly, for both standard labor supply and our microscopic framework , the *microscopic marginal utility* of leisure need not depend on the amount of work done. For a deterministic policy ($\beta \to \infty$), the derived macroscopic utility function (see Eq.(4.29)) is

$$U(l, \omega) = \frac{\omega}{P} \left[RI + \ (k-1) \ \log\left(l \ P/\omega\right) \ \right] \tag{4.15}$$

for which the slopes of the (*macroscopic*) ICs depend on the amount of work and leisure consumed (Fig. 4.4D) and generate partial allocation as optimal solutions. Thus, neither stochasticity nor an interaction between work and the marginal utility of leisure is necessary for partial allocation.

4.4.5 Generalized Matching Law: Mountain Model

As reviewed in Chapter 2, an alternate macroscopic characterisation of behaviour that yields smooth time allocation curves, hypothesises that subjects match (according to the generalised matching law, Eq. (2.10) Baum (1974); Herrnstein (1961)) their time allocation between work and leisure to the ratio of their payoffs Herrnstein (1961), R_W and $R_L = \frac{RI_{max}}{P_L}$, respectively Baum and Rachlin

[2] Choosing concave $C_L(\cdot)$ as a logarithmic function is for convenience; it would further be straightforward to take $C_L(\tau_L) = (k-1)\log(\tau_L + 1)$ so that the microscopic utility is defined over all $\tau_L \geq 0$.

(1969); Killeen (1972)

$$\begin{aligned}\frac{\omega}{l} &= \left(\frac{R_W}{R_L}\right)^a \\ \Rightarrow \frac{\omega}{\omega+l} &= TA = \frac{{R_W}^a}{{R_W}^a + {R_L}^a} = \frac{RI^a}{RI^a + \ (\frac{P}{P_L})^a}. \end{aligned} \tag{4.16}$$

Here, P_L is defined as the price at which, for a maximum subjective reward intensity RI_{max}, the subject allocates half the time to work, and half to leisure (see red lines in Figs.4.5 and 4.6A).

This establishes a 3-dimensional relationship between *TA*, *subjective* reward intensity and price (Fig.4.5, left panel) that is analogous to the *mountain model* Hernandez et al. (2010); Arvanitogiannis and Shizgal (2008)) (see Eq. (2.13) in Chapter 2), which plots this relationship in terms of the *objective* reward strength. *TA* is smooth, and increases and decreases monotonically with reward intensity and price, respectively, as evident in the contours in Fig. 4.5 (right panel). Stochastic *macroscopic* allocation, by virtue of generalised matching, therefore account for partial time allocation. The matching coefficient a determines how *TA* increases as a function of the payoff from work – rapidly for over-matching ($a > 1$), and slowly for under-matching (($a < 1$), Fig. 4.6B, respectively).

4.5 The microscopic mountain

By integrating the microscopic choices from our model, we can compare it with macroscopic descriptions such as the mountain model. In Chapter 3, we noted how censoring could lead to time allocation at very long prices increasing rather than decreasing with price. Here we study a more generic case, ignoring artefacts owing to issues such as censoring, and attempt to build a superset of the mountain model. We saw that linear $C_L(\cdot)$ generates partial allocation with stochasticity. It therefore generates smooth (non-step function) macroscopic time allocation curves as a function of both reward intensity and price. Consequently, 3-dimensional relationships can be derived that are qualitatively similar to those specified by the mountain model (when expressed in terms of subjective reward intensity, compare Fig. 4.7A with Fig. 4.5).

However, when $C_L(\cdot)$ is non-linear, more complicated structures arise. If the price is increased while holding the reward intensity fixed, the reward rate ρ^π (Eq. (4.2)) decreases hyperbolically and eventually asymptotes (Fig.4.8A). Consequently, unlike the mean, the mode of the gamma-like distribution does not substantially increase with the price (see Figs.4.3C and 4.8B). Since the mode

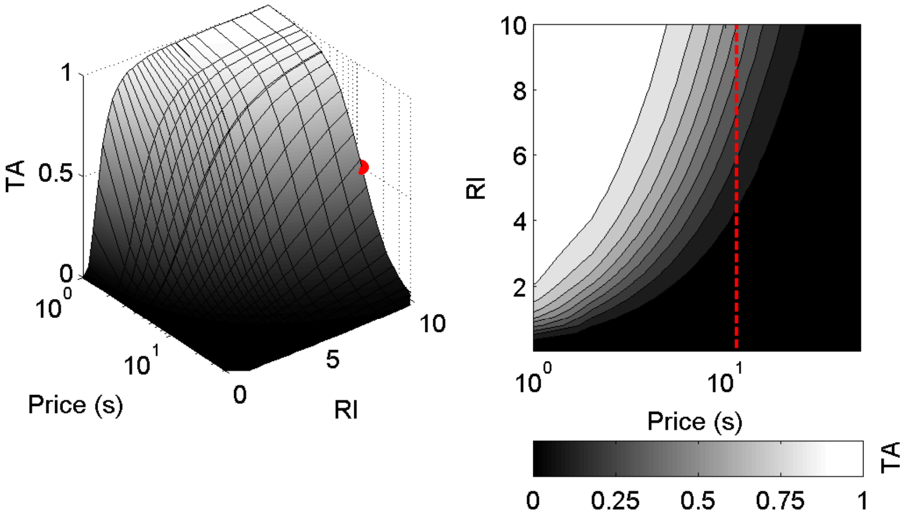


Figure 4.5: Mountain model. 3-dimensional relationship; right panel: contours of equal time allocation, as a function of reward intensity and price predicted by the mountain model using the generalised matching law. Red lines in right panel show P_L: the price at which $TA = 0.5$ for a maximal reward intensity (red dot in left panel). $a = 2.65, P_L = 11.4s$. The TA contours smoothly increase with reward intensity and smoothly decrease with price.

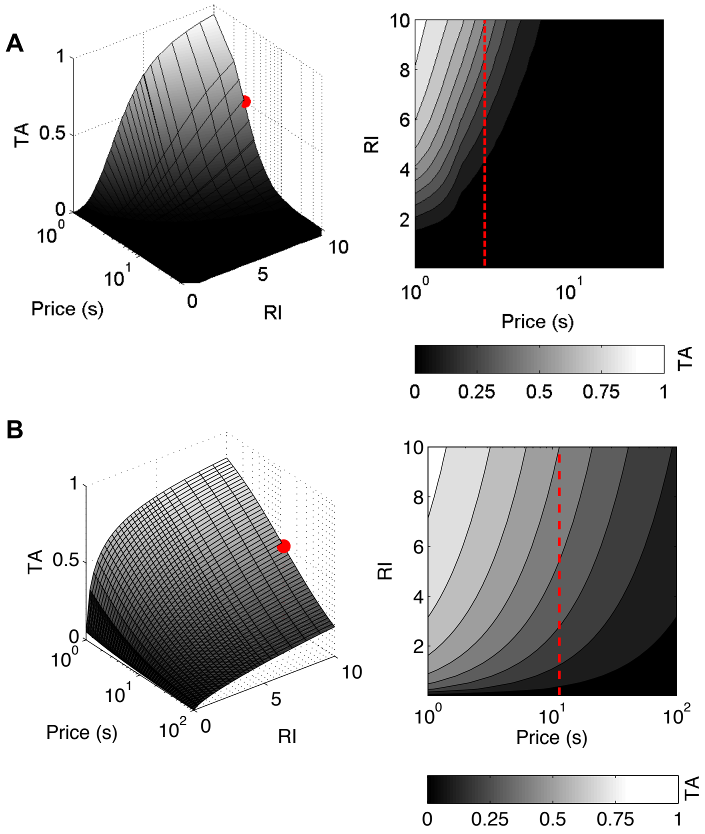


Figure 4.6: Mountain model parameters. Left 3-dimensional relationship; right panel: contours of equal time allocation, as a function of reward intensity and price predicted by the mountain model using the generalised matching law. Red lines in right panels show P_L: the price at which $TA = 0.5$ for a maximal reward intensity (red dot in left panels). A) For a small $P_L = 2.85s$, while overmatching $a = 2.65 > 1$ as in the main text and B) undermatching $a = 0.66 < 1$ while $P_L = 11.4s$ as in the main text.

determines the duration of the majority of leisure bouts, these do not increase substantially. If the subject continues to work for the entire price duration (Fig.4.8C), then, surprisingly from the macroscopic perspective of the generalized matching model, the total work time and thus the *TA* will *increase*, rather than decrease with the price (Figs.4.7B and 4.8A, lower panel). This prediction is readily amenable to experimental test.

Since for linear $C_L(\cdot)$, leisure durations are governed by substantially changing means and not modes, *TAs* are in general smaller than for strictly concave $C_L(\cdot)$, implying that higher payoffs are necessary to capture the entire *TA* range.

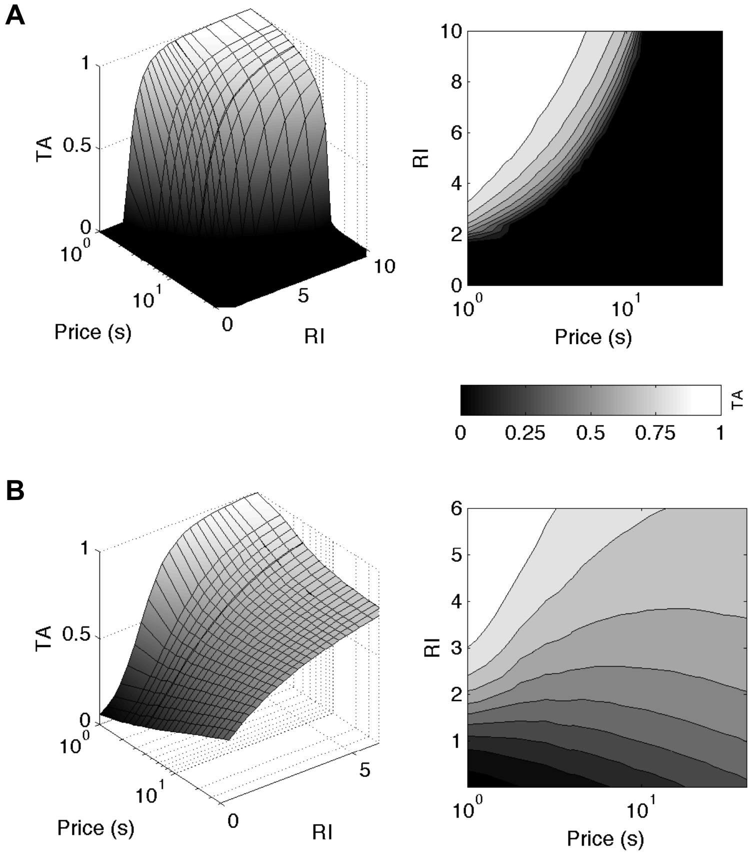


Figure 4.7: Macroscopic time allocation derived from normative, microscopic choices yields a superset of the mountain model. Left panels: 3-dimensional relationships between *TA*, reward intensity and price, right panel: contours of equal *TA*, predicted by the micro SMDP model for A) linear, B) concave $C_L(\cdot)$. The 3-dimensional relationship and smooth contours for a linear $C_L(\cdot)$ derive the mountain model in Fig.4.3. Note that an extra, higher set of reward intensities was necessary to achieve the full range of time allocation for linear $C_L(\cdot)$. The fact that contours change direction at longer prices for a non-linear $C_L(\cdot)$ rather than decrease monotonically reflects that *TA* may no longer decrease and even increase as the price is increased further.

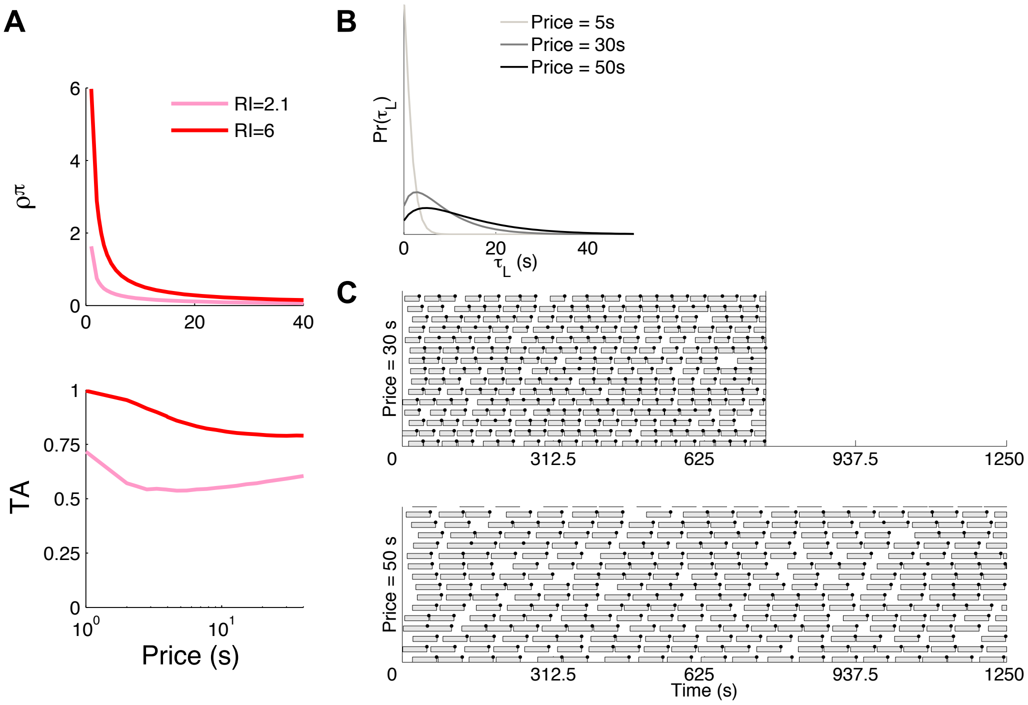


Figure 4.8: Time allocation may not decrease with price for a non-linear microscopic utility of leisure. A) Upper panel: Reward rate (ρ^π) and lower panel: time allocation (TA) for a concave microscopic utility of leisure as a function of price. A small and a high reward intensity are shown. Reward rate decreases hyperbolically with price, eventually asymptoting. B) Leisure duration distribution as a function of price for a fixed high reward intensity ($RI = 6$). At very long prices, as the price is increased further (eg. from 30s to 50s), the mode of the leisure duration distribution does not change by much although the mean does. C) Ethograms for two long prices. As price is increased, the work bouts (proportional to the price) do increase. Leisure bouts, drawn from the mode, do not change by much. Consequently, TA no longer decreases but may even increase with price (A, lower panel). This is despite the trial duration being normalised to a multiple (here 25) of the price. It is the lack of significant change in the majority of leisure durations that is critical. We normalised by the trial duration of 25 × price, instead of simply normalizing by the price, to emphasise that TA is a macroscopic quantitity and to be consistent with the procedure in the example data Figure 3.1.

4.6 Discussion

We studied the problem of partial time allocation – when reward intensities and prices are not extreme, both animals and humans divide their time between work and leisure. Traditional theories such as the microeconomic theory of labor supply, or accounts from behavioral psychology based on the generalised matching law, have characterised behavior at a macroscopic level, studying average times spent in work or leisure. While labor supply approaches have studied choices within periods of time, these have been limited to maximising utility within these time windows Blundell and Macurdy (1999); Kool and Botvinick (2012)–and thus, still average times within these windows. We proposed a normative, microscopic approach using the reinforcement learning framework of Semi-Markov Decision Processes. By integrating the microscopic choices of our model over time, we were able to account for the nature of macroscopic partial allocation.

We showed how assumptions about microscopic and macroscopic quantities relate. In labor supply theory, the marginal utility of leisure may (although not necessarily) depend on the amount of work (or rewards) consumed, and (unlike in the behavioral data) choices are classically deterministic. We considered a stochastic policy of the same form as emerges for standard random utility models, but directed at microscopic, rather than macroscopic, choices. In our case, stochasticity is reflected in the macroscopic utility function via an entropy term, and generates partial allocation even when the marginal microscopic utility of leisure is independent of work.

If the microscropic utility of leisure is linear, then the optimal allocation of time is consistent with the mountain-like products of generalized matching. However, we showed that for certain non-linearities, the time allocated to working can increase rather than decrease as the price increases, yielding complicated 3-dimensional relationships and non-monotonic contours that elude the mountain model. Whereas animals have been previously shown to consistently work more when work-requirements are greater (e.g. ostensibly owing to sunk costs Kacelnik and Marsh (2002)), the apparent anomaly discussed here only occurs at longer prices and is due to the form of the microscopic utility of leisure. This is an obvious candidate for empirical investigation, which we describe in Chapter 6.

Non-linear benefit of leisure functions can also lead to partial allocation for deterministic choices. This applies even for functions that differ from those common in labor supply theory in virtue of satisfying independence between the microscopic utilities of working and being at leisure. Of course, the marginal microscopic utility of leisure might depend on work or rewards – for instance due to fatigue or satiation. However, carefully eliminating such dependencies (by, e.g., allowing subjects sufficient rest inbetween trials, and using non-satiating rewards like BSR) may provide an avenue to quantify aspects of the microscopic utility of leisure empirically. This should help reveal why and how subjects partially allocate their time. It would then be natural to extend the study to considerations of effort, fatigue and cognitive computational costs Salamone and Correa (2002); Meyniel et al. (2013); Kool and Botvinick (2012); Botvinick et al. (2009); Kurniawan et al. (2013) (eg. from holding down weighted levers or performing cognitively demanding tasks) and the effects of manipulating motivational state Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012).

4.7 Appendix

4.7.1 Macroscopic time allocation from labor supply theory

If we consider the rewards to be continuous (instead of quantised: delivered exactly when the price is attained) or if we consider expected times spent in work or leisure only, we can construct a budget constraint (BC): the total amount of work ω and leisure l is the trial duration T

$$\omega + l = N\ P + l = T \tag{4.17}$$

where P and N are the price and number of rewards earned, respectively. Note that this budget constraint is linear in N and l.

In general, given that we maximise macroscopic utility (according to the function in Eq.(4.1)) subject to a BC, we can derive the time allocation

$$TA = \frac{(\frac{RI}{K_L})^{-\frac{1}{s-1}}}{(\frac{RI}{K_L})^{-\frac{1}{s-1}} +\ P^{\frac{s}{1-s}}} \tag{4.18}$$

which increases with RI and (for $s \geq 0$) decreases with price (Fig.4.2).

4.7.2 Linear microscopic utility of leisure yields exponentially distributed leisure durations

We repeat the derivation in Chapter 3 (but in a simplified setting, without τ_{Pav}). Suppose the microscopic utility of leisure is linear, $C_L(\tau_L) = K_L \tau_L$. Then the Q-value of engaging in leisure in the post-reward state is also linear, $Q^\pi(\tau_L) = (K_L - \rho^\pi)\tau_L + V^\pi(\text{pre})$. According to the softmax policy, the probability of choosing to engage in leisure for time τ_L in the post-reward state is proportional to the exponential of the Q-value. This probability is $\pi(\tau_L) \propto \exp\left[-\beta(\rho^\pi - K_L)\tau_L\right]$, which is an exponential distribution with mean $\mathbb{E}[\tau_L] = \frac{1}{\beta(\rho^\pi - K_L)}$. Thus, for linear $C_L(\cdot)$, leisure bout durations are always exponentially distributed with a mean which depends on the reward rate. The greater the reward rate, the shorter is the mean leisure bout.

4.7.3 Logarithmic microscopic utility of leisure yields gamma distributed leisure durations

For a logarithmic microscopic utility a leisure $C_L(\tau_L) = (k-1)\log(\tau_L)$; the Q-value of engaging in leisure in the post-reward state is a unimodal bump. The leisure duration distribution is the exponential of this bump: $\pi(\tau_L) = \frac{1}{\Gamma(1/\beta\rho^\pi)}\tau_L^{\beta(k-1)}\exp(-\beta\rho^\pi\tau_L)$. This is a gamma distribution with shape parameter $\bar{k} = \beta(k-1)+1$ and scale parameter $\frac{1}{\beta\rho^\pi}$. The mode of this gamma distribution is $\frac{k-1}{\rho^\pi}$. Thus, if the reward rate does not change substantially, neither does this mode. For the special case, of $k = 1$, the gamma distribution becomes an exponential distribution.

4.7.4 Reward rate and mean leisure duration for a linear microscopic utility of leisure

We repeat the derivation in Chapter 3 (but in a simplified setting, without τ_{Pav}). For a linear $C_L(\cdot)$, the reward rate and mean leisure duration can be analytically, self-consistently derived. The reward rate in Eq.(4.2) is simply,

$$\rho^\pi = \frac{RI + K_L\mathbb{E}[\tau_L]}{P + \mathbb{E}[\tau_L]} \tag{4.19}$$

As discussed above, leisure durations in the post-reward state are exponentially distributed with mean

$$\mathbb{E}[\tau_L] = \frac{1}{\beta(\rho^\pi - K_L)} \tag{4.20}$$

Re-arranging terms of this equation,

$$\rho^\pi = \frac{1}{\beta\ \mathbb{E}[\tau_L]} + K_L \tag{4.21}$$

Equating Eqs. (4.19) and (4.21) and solving for the mean leisure duration $\mathbb{E}[\tau_L|\text{post}]$, we derive

$$\mathbb{E}[\tau_L] = \frac{P}{\beta(RI - K_LP) - 1} \tag{4.22}$$

This is the mean leisure duration as long as $RI - K_LP > 1/\beta$, and $\mathbb{E}[\tau_L] \to \infty$ otherwise. When the former condition holds, we may substitute Eq. (4.22) into

Eq. (4.19) and solve for ρ^π

$$\begin{aligned}\rho^\pi &= \frac{(RI - K_L P)\ (\beta RI - 1)}{(RI - K_L P)\ \beta P} \\ &= \frac{\beta RI - 1}{\beta P} \qquad (4.23)\end{aligned}$$

4.7.5 Macroscopic utility derived from linear and non-linear microscopic utilities

The point of the utility function in Eq. (4.12) is to lead to choices whose macroscopic characterization is the same as those of the micro-SMDP. In particular, this means that if we maximize $U(l, \omega)$ subject to a budget constraint $l + \omega = T$ for some total duration T, then we will recover what we know to be true of the optimum $l^* P/\omega^* = \mathbb{E}_\pi[\tau_L|\text{post}] = \frac{P}{\beta(RI - K_L P) - 1}$ (Eq.(4.22)). Given the form of the optimal microscopic policy associated with Eqs. (4.3) and (4.5), we also require that the Lagrange multiplier ξ associated with Eq. (4.11) should take on the value $\rho^* = \frac{RI + \mathbb{E}_{\pi^*}[C_L(\tau_L)]}{P + \mathbb{E}_{\pi^*}[\tau_L]}$.

As required for macroscopic utility functions considered in economics, macroscopic utilities with respect to both work (or rewards) ($\frac{\partial U}{\partial \omega}$) and leisure ($\frac{\partial U}{\partial l}$) are positive. Then, since macroscopic utility is constant on an indifference curve, the total derivative with respect to a good (say leisure) is zero:

$$\begin{aligned}\frac{dU}{dl} &= \frac{\partial U}{\partial l} + \frac{\partial U}{\partial \omega}\frac{d\omega}{dl} = 0 \\ \Rightarrow \frac{d\omega}{dl} &= -\frac{\partial U}{\partial l} / \frac{\partial U}{\partial \omega} < 0 \qquad (4.24)\end{aligned}$$

This shows that indifference curves have negative slopes ($\frac{d\omega}{dl} < 0$).

The optimum (l^*, ω^*) associated with the budget constraint occurs when

$$\begin{aligned}\frac{d\omega}{dl}|_{(l^*,\omega^*)} &= \frac{d(T - (\omega + l))}{dl}|_{(l^*,\omega^*)} = -1 \\ \Rightarrow \frac{\partial U}{\partial l}|_{(l^*,\omega^*)} &= \frac{\partial U}{\partial \omega}|_{(l^*,\omega^*)} \qquad (4.25)\end{aligned}$$

Consider the case of a linear microscopic utility of leisure, $C_L(\tau_L) = K_L \tau_L$. In this case, the optimum π^* of Eq.(4.11) is exponential, implying that $\pi^*(\tau_L)|l, \omega = \frac{\omega}{lP} \exp\left[-\frac{\omega}{lP}\ \tau_L\right]$ whose entropy $H(\pi^*) = \log(lP/\omega) + 1$. Consequently, the derived macroscopic utility function in Eq.(4.12) becomes

$$U(l,\omega) = \frac{\omega}{P}RI + \ K_L\ l + \frac{\omega}{\beta\ P}\ [\log(lP/\omega) + 1] + \frac{1}{\beta}\ g(l,\omega) \tag{4.26}$$

If we define

$$g(l,\omega) = \frac{1}{P}[\ l(\log(l) - 1) + \omega(\log(\omega) - 1) - \omega(\log(P) + 1)\ ] = \frac{\omega + l}{P}[\log(l) - 1] - \frac{\omega}{P}H(\pi^*) \tag{4.27}$$

Then it turns out that the optimum has just the correct properties in terms of choice. We merely claim that this a possible $g(\cdot,\cdot)$ – it need not be unique.

If, instead, the microscopic utility of leisure is logarithmic: $C_L(\tau_L) = (k-1)\log(\tau_L)$, then, as in the case for linear $C_L(\cdot)$, we can derive $\mathbb{E}_\pi[C_L(\tau_L)]$ and $H(\pi)$ analytically in closed form for policies associated with Eq.(4.11).

$$\begin{aligned}\mathbb{E}_\pi[C_L(\tau_L)] = (k-1)\mathbb{E}_\pi[\log(\tau_L)] = (k-1)[\psi(\bar{k}) + \log(\mathbb{E}_\pi[\tau_L]) - \log(\bar{k})]\\ H(\pi) = \log(\mathbb{E}_\pi[\tau_L]) - \log(\bar{k}) + \bar{k} + (1-\bar{k})\psi(\bar{k}) + \log(\Gamma(\bar{k}))\end{aligned} \tag{4.28}$$

Here $\Gamma(\cdot)$ and $\psi(\cdot)$ represent the *gamma* and *digamma* functions, respectively and $\bar{k} = \beta(k-1)+1$ as above. It is easy to see that for the special case of $k=1$, i.e. when the gamma distribution becomes an exponential distribution, $\bar{k}$ becomes simply 1. In that case, $H(\pi) = \log(\mathbb{E}_\pi[\tau_L]) + 1$ as above, since all other quantities in Eq.(4.28) vanish. Further, if we considered a general microscopic utility which was a sum of logarithmic and linear components, $\hat{C_L}(\tau_L) = (k-1)\log(\tau_L) + K_L\tau_L$, then we could treat the linear version as a special case by simply setting $k=1$. Using the quantities in Eq.(4.28), we may derive a macroscopic utility from a microscopic logarithmic utility

$$\begin{aligned}U(l,\omega) &= \frac{\omega}{P}\left(RI + \ (k-1)\ \mathbb{E}_{\pi^*}[\log(\tau_L)] + \frac{1}{\beta}\ H(\pi^*)\right) + \frac{1}{\beta}\ g(l,\omega)\\ &= \frac{\omega}{P}\left(RI + \ (k-1)\left[\psi(\bar{k}) + \log\left(\mathbb{E}_{\pi^*}[\tau_L]\right) - \log(\bar{k})\right]\right) + \frac{1}{\beta}\left[\frac{\omega}{P}H(\pi^*) + \ g(l,\omega)\right]\\ &= \frac{\omega}{P}\left(RI + \ (k-1)\left[\psi(\bar{k}) + \log\left(l\ P/\omega\right) - \log(\bar{k})\right]\right) + \frac{1}{\beta}\left[\frac{\omega}{P}H(\pi^*) + \ g(l,\omega)\right]\end{aligned} \tag{4.29}$$

then, if

$$g(l,\omega) = \frac{1}{P}\left[\ l(\log(l)-1)+\omega(\log(l)-1))\ \right] - \frac{\omega}{P}H(\tau_L) = \frac{\omega+l}{P}[\log(l)-1] - \frac{\omega}{P}H(\pi^*) \tag{4.30}$$

then one can show that not only does maximising the derived macroscopic utility in Eq.(4.29) subject to a BC yield the appropriate mean leisure duration: $\mathbb{E}_\pi[\tau_L|\text{post}] = l^* P/\omega^*$ when $C_L(\cdot)$ is logarithmic, but also reduces to the derived macroscopic utility for a linear $C_L(\cdot)$ (when $k = 1$, see Eqs.(4.26) and (4.27)). Furthermore, as required for self-consistency, the Lagrange multiplier ξ that leads to the policy $\pi^*(\tau_L) \propto \exp(\beta(C_L(\tau_L) - \xi\tau_L))$ is the "shadow price" $\xi = \rho^* = \frac{RI+\mathbb{E}_{\pi^*}[C_L(\tau_L)]}{P+\mathbb{E}_{\pi^*}[\tau_L]}$, which is the average reward rate. If we had enforced the full budget constraint via a Lagrange multiplier, the same average reward rate would have been the shadow price for this too, i.e., the extra (macroscopic) utility arising from relaxing the total budget T, (i.e. taking an extra second total time for work and/or leisure).

Chapter 5

Fatigue and Satiation: implications for the labour-leisure tradeoff

5.1 Introduction

What to do and how long to do it for at any given moment may depend on what has been done recently. In the case of deciding whether to work or engage in leisure, the choice may depend on the recent history of work and rest, or rewards earned. For example, a labourer may rest because she is fatigued from working; an animal may stop lever pressing for food pellets and rest because it is satiated and no longer hungry. In this chapter, we seek to extend the normative, microscopic approach to the labour-leisure tradeoff by taking into account how such recent history of choices and outcomes can affect current decisions, and their implications for the characterisation of behaviour. In Chapter 4 we showed a novel prediction, that an apparent interaction between the *macroscopic* marginal utilities of work and leisure could arise even when there was no interaction between their *microscopic* marginal utilities. That is, work and leisure may appear to be imperfect substitutes macroscopically, when in reality leisure is beneficial of its own accord, and not *because* of the recent history of working. We now explore the case when in fact there is a functional interaction between the microscopic marginal utilities of work and leisure.

We specifically focus on two predominant phenomena that could underlie such an interaction between the microscopic marginal utilities: fatigue and satiation. Physical fatigue arises from performing a physical task repeatedly or for too long, with little rest in between. The physical task may not be effortful or cognitively demanding. For example, consider someone lifting a light weight at the gym, repeating this exercise many times in quick succession. She shall be physically fatigued towards the last few repetitions, and desire rest so that she can recuperate. The lightness of the weight makes lifting it once not physically effortful, it is the many repetitions that make it fatiguing. Furthermore, this barely involves much cognitive effort, the computational complexity of keeping track of the number of repetitions so far pales in comparison to that of, e.g. a chess player planning her next move. Most work on the effect of fatigue on decision-making studies cognitive fatigue. Decision-making is computationally expensive, and making too many decisions, especially computationally complex ones, leads to mental fatigue. This led to an influential account called ego-depletion in which voluntary cognitive effort declines after performing several bouts of forced cognitively demanding tasks Baumeister (2002); Baumeister and Bratslavsky (1998); Vohs and Baumeister (2008).

However, these studies do not explore the effect of fatigue on the temporal topography of choice. Whereas algorithmic explanations of the effect of physical effort on the microstructure of choices have been proposed Meyniel et al. (2013), here we put forward a normative theory of how physical fatigue affects the microstructure of work and leisure choices. This could be due to fatigue making working more costly or leisure more beneficial. In the latter case, leisure is beneficial *because* of the recent history of working–there is an interaction between the marginal microscopic utilities of work and leisure.

We consider satiation–or the diminishing marginal utility of a reward, as it is known in economics, as a dual of fatigue. In a counterintuitive set of findings, humans and animal subjects, tested in the laboratory and studied in real-life situations were observed to work less when wage-rates are increased, rather than more Hanoch (1965); Bigelow and Liebson (1972); Meisch and Thompson (1973, 1974a,b); Barofsky and Hurwitz (1968); Collier and Jennings (1969); Battalio et al. (1981); Green et al. (1987); Camerer et al. (1997). A macroscopic explanation of this posits that subjects work till they attain a particular target income, and quit working when it is attained. Since attaining such target income is easier when the wage-rate is greater, subjects work less. Similarly, in behavioural psychology experiments, response rates have been observed to have a bitonic shape, increasing at the beginning of an experimental session and decreasing thereafter owing to satiation Fischer and Fantino (1968); McSweeney et al. (1991); McSweeney and Roll (1993); McSweeney and Johnson (1994); Killeen (1995). The decrease in response rates was faster for the higher reinforcer rates McSweeney

(1992), durations and sizes Bizo et al. (1998), suggesting subjects satiate faster for greater reward values.

We seek to build a normative, microscopic framework for decision-making as a subject satiates from consuming rewards. This could be either due to rewards becoming less rewarding or leisure becoming more beneficial with satiation. Similar to fatigue, the latter case implies an interaction between the marginal utilities of leisure and rewards consumed. We shall derive the above counterintuitive macroscopic phenomenon from our microscopic theory.

We approach these issues from two perspectives according to which decisions may be made: prospectively or retrospectively. For example, an individual deciding whether and how long to rest on a Sunday before a long week decides prospectively, e.g. taking into account how fatigued she shall be during the week. On the other hand, an individual deciding whether and how long to rest on a Saturday after a long week is deciding retrospectively, e.g. taking into account the extra benefit of leisure when already fatigued. We consider each of these underlying factors in turn.

5.2 Fatigue

We consider a simple mechanism in which fatigue accumulates as the subject works and dissipates as it engages in leisure. We represent this as a fatigue variable $\nu(t)$, which low-pass filters the recent history of work and leisure (Fig. 5.1). This yields an exponential kernel of past work and leisure durations and the fatigue variable ν' at the end of a bout of duration τ_a is linearly related to that at the start of the bout ν

$$\nu' = \nu(t+\tau_a) = z_a \; [1-(1-\alpha)^{\tau_a}] + (1-\alpha)^{\tau_a} \; \nu \tag{5.1}$$

where z_a=1 if the subject works ($a = W$) and 0 if it engages in leisure ($a = L$). The rate of filtering $\alpha \in [0, 1]$ represents the inverse of the time-constant with which fatigue dissipates or builds up: a small α signifies a slow time-constant. $\nu = 0$ represents no fatigue while $\nu = 1$ signifies maximum fatigue. We shall consider the same CHT task and action space as before. We once again assume that the subject works the entire price and then engages in leisure, and characterise steady-state behaviour using the recurrent micro-SMDP discussed in Chapter 4. This maintains the pre and post-reward state space of that model, except that it is augmented with the fatigue variable ν. We do not separately model a possible set of transient states wherein fatigue accumulates and, and focus only on behaviour as the subject transitions between recurrent states. We explore two cases of how fatigue can affect microscopic work and leisure choices, either by making working more costly when fatigued, or leisure more beneficial. We can consider a cost of working whilst fatigued $C_{F_W}(\nu)$ or a utility of leisure that depends on fatigue $C_L(\nu, \tau_L)$.

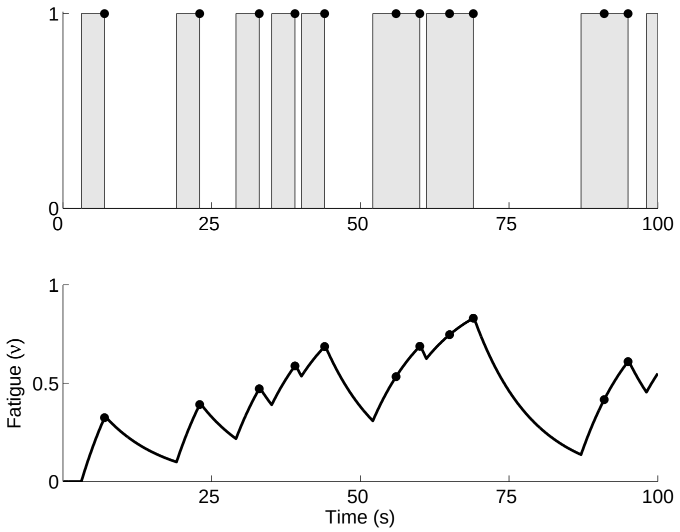


Figure 5.1: Fatigue as a low pass-filtered variable ν of the recent history of work and leisure. Fatigue accumulates as the subject works and dissipates as it engages in leisure, exponentially in both cases.

The Q-values of working and engaging in leisure then become

$$\begin{aligned} Q^\pi([\text{pre}, \nu], [W, P]) &= RI - \rho^\pi P - C_{F_W}(\nu) + V^\pi(\text{post}, \nu') \\ Q^\pi([\text{post}, \nu], [L, \tau_L]) &= C_L(\nu, \tau_L) - \rho^\pi \tau_L + V^\pi(\text{pre}, \nu') \end{aligned} \tag{5.2}$$

which depend on the value $V^\pi(\cdot, \nu')$ of the fatigue state that the subject transitions to: more fatigued after working and less fatigued after engaging in leisure (see Eq.(5.1)).

As we shall show, the two cases of work being costly or leisure being more beneficial when fatigued make similar predictions about the microscopic (and thus also macroscopic) structure of work and leisure choices. They may be potentially distinguished by considering experiments in which choices are made prospectively: taking into account how fatigued one will be in the future, or retrospectively: because one is already fatigued. We call these prospective and retrospective fatigue, respectively.

5.2.1 Prospective fatigue: cost of working whilst fatigued

Suppose one is considering whether and how long to rest on a Sunday before a long week. If working is costly when fatigued, then one should take plenty of rest on a Sunday, prospectively taking into account how fatigued one will be otherwise. To demonstrate this using the CHT paradigm, suppose further that the subject is *forced* to work for the entire price each time it works. We had previously assumed that subjects choose to work for the entire price duration, but in this case, the employer enforces this. For simplicity, we assume a cost of working that is independent of the duration of a work bout and increases linearly with fatigue $C_{F_W}(\nu) = K_F\ \nu$. While considering a cost of working that depends on the duration of the work bout is perhaps more natural, this could be due to effort costs. Our simplifying assumption avoids such a confound and attempts to isolate the cost of working whilst fatigued. Further, we assume a linear microscopic utility of leisure $C_L(\tau_L) = K_L \tau_L$, *independent* of how fatigued the subject is.

We explain the algorithmic mechanics underlying our model of prospective fatigue: the Q-value of leisure in the post-reward state depends on the value of the pre-reward (forced work) state $V^\pi([\text{pre}, \nu'])$ to which the subject transitions to as a consequence of engaging leisure for duration τ_L: $V^\pi([\text{pre}, \nu']) = RI - \rho^\pi P - K_{F_W}\nu' + V^\pi([\text{post}, \nu''])$, where ν'' is the post-reward fatigue state that the subject further transitions to. Which pre-reward fatigue state ν' a leisure bout transitions the subject into depends exponentially on the current state of fatigue ν through $(1-\alpha)^{\tau_L}\nu$ (Fig. 5.2A(i) and see Eq. (5.1)). If the initial state is non-fatigued ($\nu = 0$), then further leisure cannot further reduce it ($\nu' = 0$). Otherwise, it decays exponentially from the initial fatigue state.

For a sufficiently large K_{F_W}, the value $V^\pi([\text{pre}, \nu'])$ is dominated by the cost of working whilst fatigued: $-K_{F_W}\nu'$. This is just the negative of the above exponential decay (multiplied by a constant K_{F_W}) and shifted by a fatigue independent term $RI - \rho^\pi P$ (Fig. 5.2A(ii))

The difference between the linear utility of leisure $K_L\tau_L$ (solid line in Fig. 5.2A(iii)) and the, also linear, opportunity cost of time $-\rho^\pi\tau_L$ (dashed line), yields the bold line, as in Chapter 3. This is then added to the value of the pre-reward fatigue state that the leisure bout takes the subject into $V^\pi([\text{pre}, \nu'])$ to yield the Q-value of leisure as a function of initial fatigue (Fig. 5.2 A(iv)). As

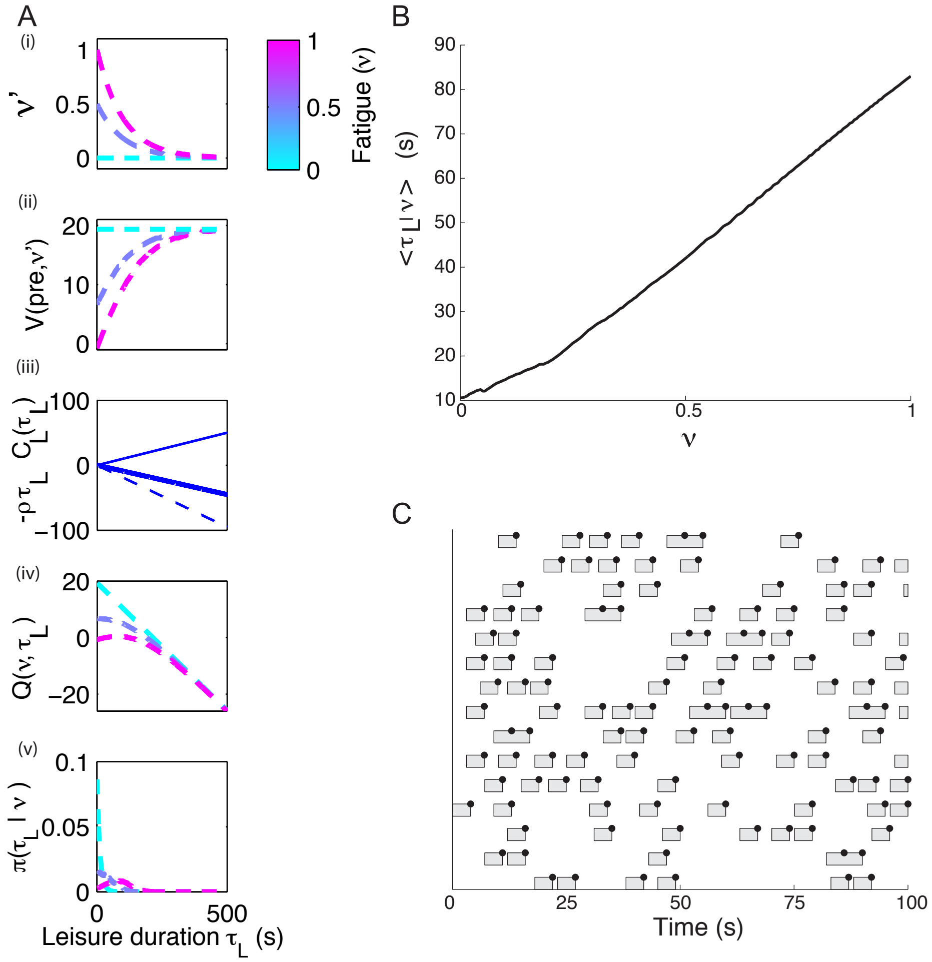


Figure 5.2: A) Prospective fatigue model mechanics. From top to bottom: (i) $\nu' = \nu_{t+\tau_a} = z_a\ [1-(1-\alpha)^{\tau_a}] + (1-\alpha)^{\tau}_{a}\ \nu$, where z_a=1 if $a = W$ and 0 if $a = L$; long leisure resets fatigue. Blue to red curves show increasing initial fatigue levels (ν). (ii) the value of pre-reward (forced work) state $V^{\pi}([\text{pre}, \nu'])$ to which the subject transitions to as a consequence of taking leisure for duration τ_L: $V^{\pi}([\text{pre}, \nu']) = RI - \rho^{\pi} P - K_{F_W}\nu' + V^{\pi}(\text{post}, \nu'')$. This is dominated by the $-K_{F_W}\nu'$ term. (iii) the linear utility of leisure $K_L\tau_L$ (solid line) and the opportunity cost of time $-\rho^{\pi}\tau_L$ (dashed line) are added (bold line) and then (iv) added with the value of the pre-reward state to yield the Q-value of leisure for duration τ_L starting from fatigue state ν. (v) Finally, the Q-value is sent through the softmax to yield the policy $\pi(\tau_L|[\text{post}, \nu])$. Note that in the absence of fatigue ($\nu = 0$) this is an exponential distribution; whereas it is a gamma distribution with a longer mode for greater initial fatigue. B) The mean leisure duration increases with fatigue. C) Ethograms show runs as a fatigue builds up with each work bout and is alleviated by longer leisure bouts. $RI = 3, P = 4s, K_{F_W} = 3$.

discussed previously, the Q-value of leisure in the absence of fatigue ($\nu = 0$) is linearly related to its duration. For initially fatigued states ($\nu > 0$), the Q value is a unimodal bump, with the peak of the bump increasing with fatigue.

When the Q-value is sent through the soft-max to yield the policy $\pi(\tau_L|[\text{post}, \nu])$, we achieve the usual exponential distribution in the absence of fatigue ($\nu = 0$). However, when the subject is already fatigued ($\nu > 0$), the leisure duration distribution is gamma-like, with its mode and tail increasing with the degree of initial fatigue (Fig. 5.2 A(v)). Consequently, the mean leisure duration increases as a function of initial fatigue (Fig. 5.2 B). Ethograms generated according to this policy yield a temporal pattern of *runs* of work bouts interspersed by short leisure bouts as fatigue accumulates, followed by a longer leisure bout to alleviate fatigue (Fig. 5.2C).

As the price is increased, the opportunity cost of time decreases, and consequently so does the mean leisure duration (Fig. 5.3A). Long leisure is particularly desirable for high initial fatigue levels, since being forced to work longer would otherwise exacerbate fatigue even more. This is particularly deleterious if the cost of working whilst fatigued is high (compare left and right panels in Fig. 5.3A). The microstructure of work and leisure clearly shows how leisure breaks in between runs of working become longer as the cost of working whilst fatigued is amplified (compare left and right panels in Fig. 5.3B).

5.2.2 Retrospective fatigue: benefit of resting whilst tired

The more intuitive case of fatigue is retrospective. Suppose one is deciding whether and how long to rest on a Saturday *after* a long, tiring week. In this case the utility of leisure is greater if one is more fatigued. The microscopic utility of leisure $C_L(\nu, \tau_L)$ depends on the fatigue level and duration of leisure τ_L, but fatigue dynamically decreases from ν to ν' as the subject engages in leisure. $C_L(\nu, \tau_L)$ thus represents the microscopic utility of leisure, integrating over its microscopic marginal utility $\frac{\partial C_L}{\partial d\tau_L}$.

$$C_L(\nu, \tau_L) = \int_0^{\tau_L} d\tau_L \frac{\partial C_L\left(\nu'(\tau_L|\nu), \cdot\right)}{\partial d\tau_L} \tag{5.3}$$

For simplicity, let us assume a constant momentary marginal utility of leisure $\frac{\partial C_L(\nu'(\tau_L|\nu), \cdot)}{\partial d\tau_L} = K_{L_F}(\nu'(\tau_L|\nu))$, but which increases with the level of fatigue. This corresponds to a constant slope of the momentary microscopic utility of leisure as a function of its duration, but with the slope that is larger if the subject

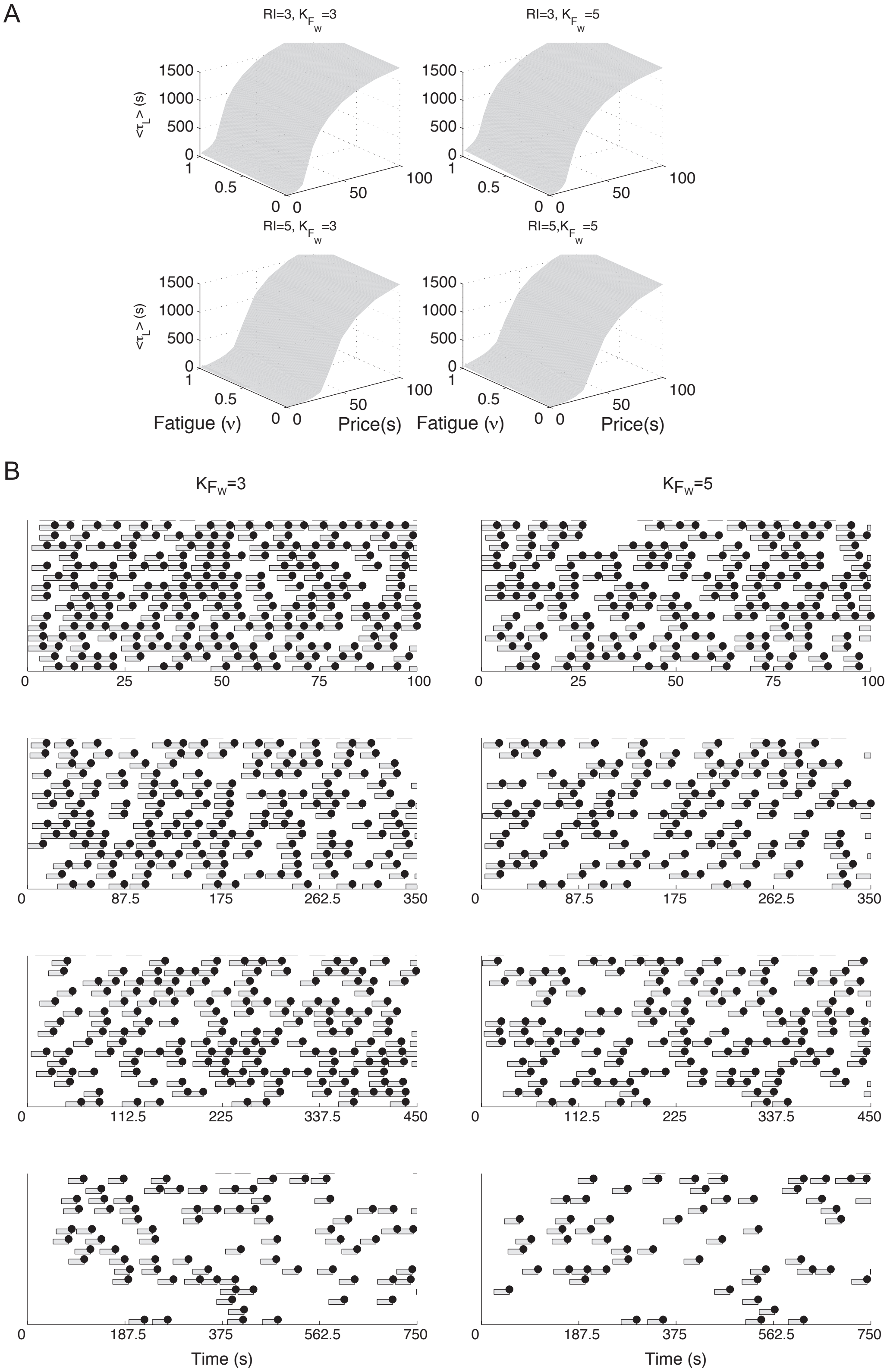


Figure 5.3: A) Effect of Price and cost of working on leisure durations. Mean leisure duration increases with price owing to the reduced opportunity cost of time. But longer leisure is is desired the more fatigued the subject is. Upper right panel: Note how the surface is shifted up at high fatigue levels. Short leisure would lead to the subject having to work in a highly fatigued state; which is exacerbated if the cost of working is amplified. B) Ethograms showing runs for lesser $K_{F_W} = 3$ (left) and greater $K_{F_W} = 5$ (right) costs of working whilst fatigued. Leisure breaks between runs of working are longer if the cost of working is greater. Price increases from top to bottom: 4s, 14s, 18s and 30s; note the different scales on the x-axis as price is increased.

is more fatigued (Fig. 5.4A (i)). To maintain simplicity, let this microscopic marginal utility be a linearly increasing function of fatigue $K_{L_f}(\nu) = K_L \cdot [\nu + 1]$. We use the $+1$ so that there is some marginal utility of leisure K_L even when there is no fatigue ($\nu = 0$). Then $K_{L_f}(\nu'(\tau_L|\nu) + 1) = K_L \cdot [(\nu'(\tau_L|\nu)) + 1] = K_L \cdot [(1-\alpha)^{\tau_L}\nu) + 1]$. In effect, this slope is decreasing as the duration of leisure increases. We can therefore derive the integrated microscopic utility of leisure as function of initial fatigue and duration of leisure

$$\begin{aligned} C_L(\nu, \tau_L) &= \int_0^{\tau_L} d\tau_L \; \frac{\partial C_L\left(\nu'(\tau_L|\nu), \cdot\right)}{\partial d\tau_L} \\ &= \int_0^{\tau_L} d\tau_L \; K_L[(1-\alpha)^{\tau}\nu) + 1] \\ &= K_L\nu \left[\frac{(1-\alpha)^{\tau_L} - 1}{\log(1-\alpha)}\right] + K_L\tau_L \end{aligned} \tag{5.4}$$

which is the sum of a Gamma and a linear function (Fig. 5.4A (ii)). Thus, we have derived a concave microscopic utility of leisure, which increases with the level of fatigue, eventually becoming a linear function for very long leisure bouts. This expresses in microscopic terms a combination of two commonly championed properties of a utility of leisure function in economics (i) being concave i.e., with decreasing marginal utility, and (ii) being interactive, i.e, leisure is beneficial *because* of the recent history of work i.e.. the marginal utility of leisure depends on work (here due to fatigue).

The Q-value of leisure given an initial level of fatigue ν, is simply the difference between this microscopic utility and the opportunity cost of time (Fig. 5.4A (iii)). As before, the Q-value is linear in the absence of fatigue ($\nu = 0$). For non-zero levels of fatigue, it is a unimodal bump, whose peak increases with fatigue. The policy derived from sending the Q-value through a soft-max is exponential in the absence of fatigue, and gamma-like in its presence, with the mode of the gamma distribution increasing with the level of fatigue (Fig. 5.4A (iv)). Leisure durations drawn according to this policy lead to a pattern of runs of work bouts followed by a long leisure bout that reduces fatigue (Fig. 5.4B). These runs are particularly evident at longer prices (compare top to bottom panels of Fig. 5.4B depicting the effect of the price increasing). We thus note that the microscopic behaviour owing to retrospective fatigue is similar to that in the prospective case. Teasing apart which type of decision-making underlies such behaviour could require careful experimental designs which control for one while testing the other.

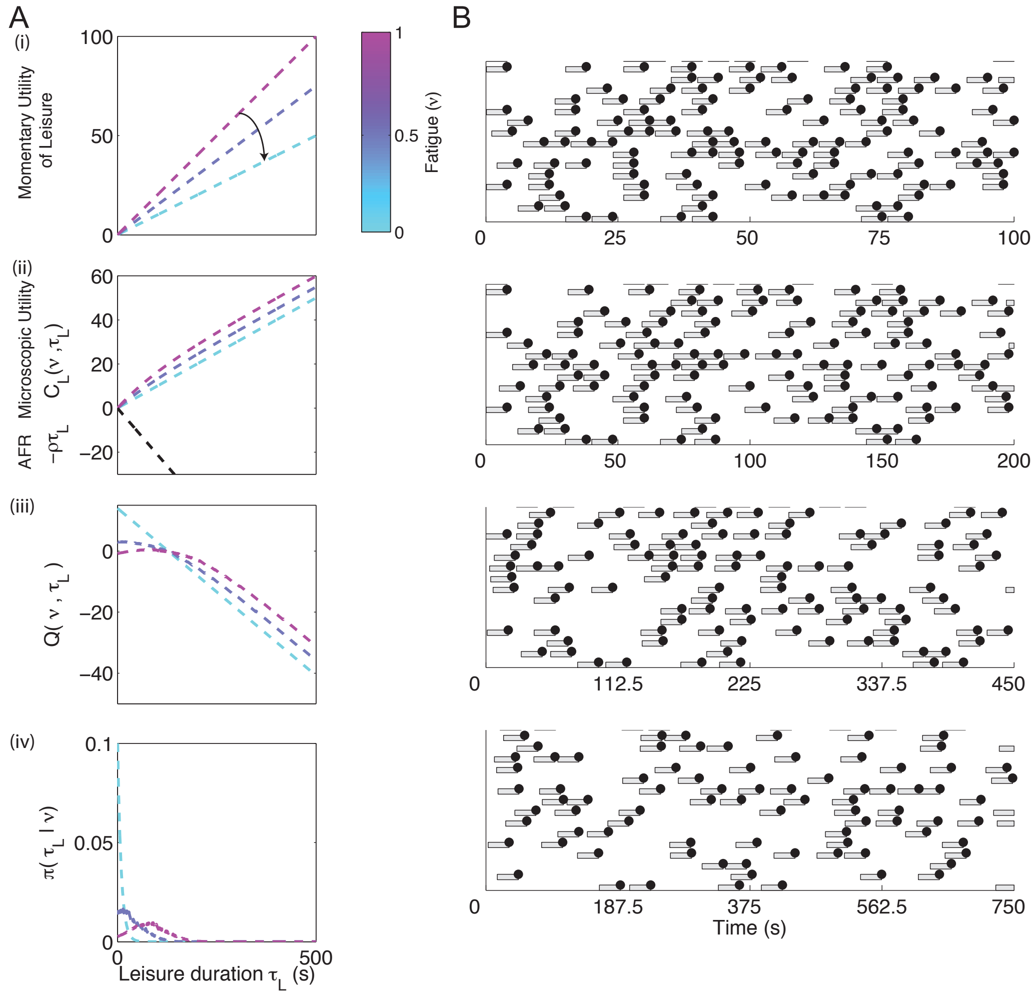


Figure 5.4: Retrospective fatigue A) mechanics, from top to bottom: (i) linear momentary microscopic utility of leisure, but increases with fatigue. This represents a microscopic utility function with a constant microscopic marginal utility $\frac{\partial C_L(\nu'(\tau_L|\nu),\cdot)}{\partial d\tau_L} = K_{L_F}(\nu'(\tau_L|\nu))$, but which increases with the level of fatigue. (ii) the integrated microscopic utility of leisure $C_L(\nu, \tau_L)$ for a bout of duration τ_L, starting from a fatigue level ν is a sum of a Gamma and linear function. This integrates over the momentary microscopic utility of leisure taking into account that fatigue decreases while the subject is engaging in leisure. It is a concave function which increases with the initial level of fatigue. Dashed black line shows the opportunity cost of time. (iii) Q-value of leisure for duration τ_L starting from fatigue state ν linear in the absence of fatigue ($\nu = 0$), but is a bump whose peak increases with initial level of fatigue. (iv) Finally, the Q-value is sent through the softmax to yield the policy $\pi(\tau_L|[post, \nu])$. Note that in the absence of fatigue ($\nu = 0$) this is an exponential distribution; whereas it is a gamma distribution with a longer mode for greater initial fatigue. B) Ethograms show runs as fatigue builds up with each work bout and is alleviated by longer leisure bouts. Price increases from top to bottom: 4s, 8s, 18s and 30s; note the different scales on the x-axis as price is increased. The leisure bouts in between runs are longer for longer prices, since the subject is more fatigued. $RI = 3, K_{F_W} = 3$.

5.3 The backward bending labour supply curve

We shall develop a dynamic theory of satiation, as a dual to that of fatigue delineated above, and use it to build a normative, microscopic account for a counterintuitive phenomenon from economics and behavioural economics.

The macroscopic and microscopic theories considered in Chapters 2, 3 and 4, as well as the normative, microscopic theory of fatigue developed above, all predict that subjects shall work more if the payoff (wage-rate) is higher. We now address a counterintuitive observation in both humans and animals: subjects sometimes work less and engage in more leisure when wage-rates are increased. We shall explain this counterintuitive phenomenon by developing a normative, microscopic theory of satiation. We start, however, by reviewing the relevant literature on this phenomenon.

5.3.1 Predictions from labour supply theory

For the macroscopic utility functions considered so far, labour supply predicts that subjects work more as the wage rate increases (as reward intensity increases and/or price decreases). An interesting, normative prediction from macroscopic labour supply theory is that wage rate increases that are not income-compensated (as in the cases we have considered) may result in the subject working less and not more (Fig. 5.5A, upper panel). As noted in Chapter 2, and shown in Fig. 5.5B an income uncompensated wage rate change can be decomposed into (i) an income compensated wage increase that would allow the subject to maintain the same level of income consumption and leisure (with the budget constraint line passing through the same income-leisure combination), and (ii) income effect owing to the increased wage (with the budget constraint being shifted upwards). Despite having the opportunity to maintain the same income-leisure combination, the effect of an income compensated wage increase is that the budget constraint is now tangent to a different indifference curve with greater utility. At this optimum, the subject substitutes more work for leisure (substitution effect). It is thus clear that income compensated wage changes will only result in the substitution effect. However, as the wage-rate is increased in the absence of any compensation, the substitution effect may be dominated by the income effect. The budget constraint is tangent to an indifference curve at which the subject engages in more leisure, and works less. In other words, the subject can purchase more leisure due its increased income and therefore works less. When the amount of work performed / labour supplied (abscissa) is plotted against the wage rate (ordinate) we obtain a *backward bending labour supply curve* (Fig. 5.5A, lower panel). The subject works less as the wage rate is increased. We plot the dependent variable: labour supplied, on the abscissa as is conventional in economics when plotting the supply of a good. The forward bending part of the labour supply curve arises from the usual case where substitution effects dominate income effects: subjects substitute more work for leisure as the wage rate is increased.

It is important to note that not all macroscopic utility functions can produce this backward bending labour supply curve. For example, the constant elasticity of substitution utility functions that we have discussed in Chapter 4 only allow forward, but not backward bending labour supply curves. The conditions necessary to produce the backward bending labour supply curve from using labour supply theory in are described in Hanoch (1965).

5.3.2 Behavioural laboratory experiments

Evidence for the backward bending labour supply curve comes from laboratory studies in both humans and animals working on ratio schedules to gain a variety of different rewards. This includes humans and rats pressing levers to gain access to alcohol Bigelow and Liebson (1972); Meisch and Thompson (1973, 1974a,b), rats depressing levers to receive food pellets Barofsky and Hurwitz (1968) and sucrose solutions Collier and Jennings (1969), and pigeons pecking a response key to gain access to food Battalio et al. (1981); Green et al. (1987). These all display the forward and then backward bends of the labour supply curve.

5.3.3 Behavioural economics

Behavioural economists have investigated the backward bending labour supply curve in field studies. While less controlled than laboratory experiments, they provide a real-life measure of labour supply in humans from which causes underlying the curve maybe gleaned.

5.3.3.1 Taxi drivers

One of the most prominent studies is that of New York City taxi drivers, conducted by Camerer et al. (1997). They used the trip sheets of drivers in the late 1980s and early 1990s, which logged details of each fare. They computed daily wage rate (R_W) as the total daily income (m) divided by the total daily hours ω. They obtained a significantly negative correlation when they regressed the

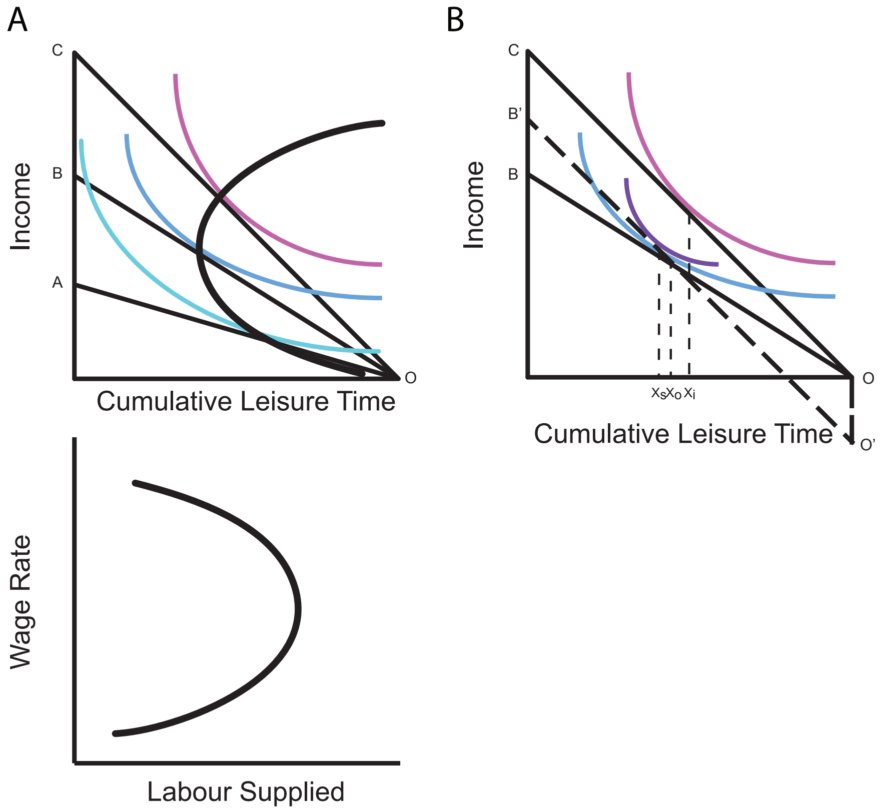


Figure 5.5: Backward bending labour supply curve A) For some macroscopic utility functions (here displayed as a function of income and cumulative leisure time), income uncompensated wage rate increases lead to the amount of work decreasing rather than increasing with wage rate. Upper panel: As the wage rate increases, the budget constraint rotates from OA to OB to OC reflecting the opportunity to earn greater income for the same amount of labour supplied. The point of tangency between an indifference curve (coloured curves) and the budget constraint yields the optimal combination of income and leisure. As the budget constraint rotates from OA to OB, the optimal cumulative leisure duration decreases. However, as the wage rate is increased further, the budget constraint rotates from OB to OC. The new optimal cumulative leisure time increases. Lower panel: when labour supplied (cumulative work time, i.e. the opposite of the axis in panel A, abscissa) is plotted against the wage rate (ordinate), we obtain a forward and then backward bending labour supply curve. B) The backward bending segment of the curve (e.g. due to the wage rate increasing and the budget constraint rotating from OB to OC) in panel A can be decomposed into two effects. First, the substitution effect, which would be due to an imaginary income compensated wage increase. The budget constraint would then shift from OB to O'B', leaving the subject the opportunity to consume the same income-leisure combination (X_o). However, this imaginary budget constraint is tangent to an indifference curve with a greater utility. The optimal allocation is to allocate less time to leisure, and work more (X_s). The substitution effect thus always leads to an increase in labour supply as wage rate increases. Second, the income effect. The increased wage rate enables the subject to gain more income. The budget constraint is shifted upward from O'B' in parallel, to OC. The new budget constraint OC is tangent to an indifference curve for which the optimal combination of income and leisure (X_i) involves the cumulative leisure time increasing compared to the original level (X_o). Thus, when the income effect (shift from X_o to X_i) dominates the substitution effect (magnitude of the shift fromX_o to X_s) the subject should work less more rather than more. Note that the subject still consumes more income as result of the wage rate increase. This greater income enables the subject to purchase more leisure, and work less.

logarithm of total daily hours worked against the logarithm of the daily wage rate.

$$\log(\omega) = \eta \log(R_W) + Z + \epsilon = -\eta \log(\omega) + \eta \log(m) + Z + \epsilon \tag{5.5}$$

where Z represents other, fixed effects terms, ϵ is Gaussian distributed noise.

This indicates that hours worked decreases for a percentage increase in wage rate. They concluded that New York City taxi drivers show a backward bending labour supply curve. They also found that wage rates were positively autocorrelated between hours within a day–and hence stable within a day, but not correlated across days. Taking these together, they propose that New York City taxi drivers have a daily income target in mind and quit once this is achieved. This was also suggested by several drivers and just more than half of the fleet managers surveyed by them. Camerer et al. (1997) proposed that taxi drivers are loss-averse to daily incomes below the target. This finding of negative correlations between log daily hours worked and log daily wage rate was replicated by Chou (2002) in taxi drivers in Singapore.

When Camerer et al. (1997) sorted drivers according to experience, by classifying those with license plate numbers less than the median as experienced, they noted that, except in one dataset, experienced drivers had a positive correlation between log daily hours worked and log daily wage rate. Inexperienced drivers consistently showed a negative correlation. By this they suggested that inexperienced drivers may be using income targeting and quitting once that target is attained, while experienced drivers work longer on the more profitable, high wage-rate days.

The negative correlations were obtained by regressing macroscopic quantities: the daily total hours worked against the daily wage rate. As noted by Camerer et al. (1997) themselves, and subsequently by Farber (2005), calculating daily wage rates by using the total daily hours worked can potentially exaggerate the negativity of this regression coefficient. log hours worked is both the dependent variable and the main independent variable (with a negative coefficient) in Eq. (5.5). Any measurement error in the hours worked would inflate the negativity of this coefficient.

Farber (2005) analysed his own dataset of New York City taxi drivers, as well as that of Camerer et al. (1997), using a miniscopic approach. He computed the probability that drivers would quit after a trip as a function of the cumulative number of hours worked so far and the income earned so far during the shift, and found that this was more significantly related to the former than the latter.

When the data were further partitioned according to experience of the drivers, a significant positive correlation was obtained for cumulative hours worked so far, while there was no significant positive, and sometimes actually negative, correlations with the income earned so far for both experienced and inexperienced drivers. While the dataset of Farber (2005) showed macroscopic negative wage elasticities, just as in Camerer et al. (1997); Chou (2002) , the strong relationship between the probability of quitting for the day with cumulative hours worked led to the conclusion that drivers prefer to work either a fixed set of hours or until they are tired. It must be noted however, that unlike Camerer et al. (1997), Farber (2005)'s data showed no autocorrelation in wage rates within a day–indicating that the fluctuating wage rates could be a reason why strong effects of income earned are not observed.

5.3.3.2 Bicycle messengers

The analyses of taxi drivers were purely correlational, and the conclusions drawn were susceptible to the type of analysis performed. Specifically, they did not partition their data between high and low wage rate days. Fehr and Goette (2007) studied the labour supply of bicycle messengers, performing the causal manipulation of increasing their commission rates were increased for a month, and then resetting them to normal. They found that messengers worked more in the high commission rate month, but did so by working less per day and more days in that month. They concluded from this that messengers had a daily income target in mind, and quit earlier when that target was achieved.

The cleanest evidence of the effects of wage rate increase on the time course of labour supply comes from the miniscopic analyses of Goette and Huffman (2006). They investigated the revenues earned per hour of bicycle messengers whose commission rate was increased compared to those employed by another firm for whom commission rates remained the same. Firstly, they showed the temporal profile of revenues earned. Messengers increased work from the beginning of the day, followed by a reduction around lunch time and a subsequent increase and decay. Secondly, they found that there was no increase in net daily labour supply due to the increased commission rate. Thirdly, messengers in the firm with the increased commission rate worked more and quit earlier in the day. The increased labour earlier in the day was compensated by the reduced labour later on, leaving net daily labour supply unaffected. Goette and Huffman (2006) propose that messengers have a daily income target in mind, and the utility of a reward increases until the target is attained, after which it drops discontinuously, and sharply. Finally, messengers showed evidence of increasing their income per hour as they gained experience over many months on the job. However, they increased their work early in the day and reduced that later on. This is contrary to the claim of Camerer et al. (1997), who concluded that experienced taxi drivers work more throughout the day when the wage rate is higher.

While these studies of bicycle messengers lend further support to the income targeting hypotheses, they remain miniscopic– reporting revenues or labour supplied per hour. They do not investigate the duration of leisure breaks, and temporal topography of work and leisure. The idea that utility of rewards drops sharply as soon as a target is attained is not verified by the data which shows a more gradual reduction in labour supply. Further the idea of having an income target in mind seems somewhat ad-hoc. We now develop a normative, microscopic approach to understanding the backward bending labour supply curve proposing satiation as an underlying factor.

5.4 Satiation

Whereas fatigue affects either the cost of working or the utility of leisure owing to the recent history of work, satiation reduces the utility of a reward or increases the utility of leisure (consider the greater utility of rest after a sumptuous lunch) owing to the recent history of rewards. We derive a macroscopic backward bending labour supply curve by formulating a normative model of microscopic satiation. Similar to fatigue the two cases of satiation, namely, reducing the utility of a reward or increasing the utility of leisure, can be distinguished by considering decisions being made prospectively or retrospectively. We consider satiation to be the dual of fatigue, and define a dynamically changing satiation variable $\psi(t) \in [0,1]$, which decays when there is no reward, but jumps each time a reward is received (Fig.5.6A). Let rewards be received at times $\{t_i\}$. The satiation variable at the next, (infinitesimally small) time-step $t + dt$ is given by

$$\psi(t+dt) = \begin{cases} \psi(t) + (1-\psi(t))\alpha_+ RI & t = t_i \\ \alpha_- \psi(t) & \text{otherwise} \end{cases}$$

We express $\psi(t + dt)$ at a discrete time-step, to show the jump after a reward is received clearly; in the limit of infinitesimally small time-steps, $\psi(t)$ is a continuous variable. The satiation variable increases with the utility of reward RI. The $(1 - \psi(t))$ term ensures that the satiation variable saturates at 1. For the purposes of flexibility, we consider different time-constants for the decay of satiation ($1/\alpha_-$) in the absence of rewards, and the jump at the time of a reward ($1/\alpha_+$). The satiation variable decays exponentially from an initial level ψ after a bout of duration τ_a to $\psi' = \alpha_-^{\tau_a}\psi$ in the absence of a reward. For a work bout of duration $\tau_W = P$, this is $\alpha_-^P\psi$. If we assume that rewards are received only when the price is attained, and that if a subject works, it works continuously for the entire price, then starting from ψ, the satiation variable after receiving a reward is: $\psi' = \alpha_-^P\psi + (1 - \alpha_-^P\psi)\alpha_+ RI$. We use the same recurrent SMDP as for the case of fatigue, except that the fatigue variable ν is replaced by the satiation variable ψ. While our assumptions below are arbitrary, we use them for simplicity to show the effects of dynamic satiation on decision-making.

5.4.1 Prospective satiation / satiation reduces the net utility of a reward

Consider the case of someone assuming whether to have a light tea in order to enjoy a sumptuous dinner. A filling tea would make one less hungry and more satiated, reducing the utility of the food at dinner. We term this prospective satiation. Let us assume that the only effect of satiation is to reduce the utility of a reward, leaving the microscopic utility of leisure unchanged. For extreme simplicity, we suggest that the utility of the reward is reduced to $(1 - \psi)RI$. Then the Q values of working and engaging in leisure are

$$\begin{aligned} Q^\pi([\text{pre}, \psi], [W, P]) &= (1-\psi)RI - \rho^\pi P + V^\pi([\text{post}, \psi']) \\ Q^\pi([\text{post}, \psi], [L, \tau_L]) &= C_L(\tau_L) - \rho^\pi \tau_L + V^\pi([\text{pre}, \psi']) \end{aligned} \tag{5.6}$$

As before, for the prospective case, we assume a linear microscopic utility $C_L(\tau_L) = K_L\tau_L$.

The mechanics of how the Q-values and policies are rendered for prospective satiation are very similar to those for prospective fatigue. When choices are generated according to the softmax policy, we once again obtain runs of working and gaining rewards followed by a long leisure bout to reduce satiation. Since satiation increases with reward intensity, the Q value of working and gaining a reward decreases. Consequently, a subject will engage in longer leisure so that the prospects of working and gaining a reward are greater. The greater the reward intensity, the faster and more frequently the subject becomes satiated. The leisure bouts required to reduce satiation therefore become longer as reward intensity increases. As a result, when we average over these temporal patterns, time allocation, and hence the labour supplied, decreases as reward intensity increases. We thus derive the backward bending labour supply curve from a normative microscopic approach using prospective satiation (Fig. 5.6). For much lower reward intensities, the subject does not satiate, as satiation builds up very slowly and is continually alleviated by leisure bouts. For reward intensities in this range, the subject therefore works more to receive more rewards without satiating. The macroscopic labour supply curve obtained from averaging across these cases is forward bending. Taken together, we provide a novel, normative, microscopic explanation to the entire forward and then backward bending labour supply curve.

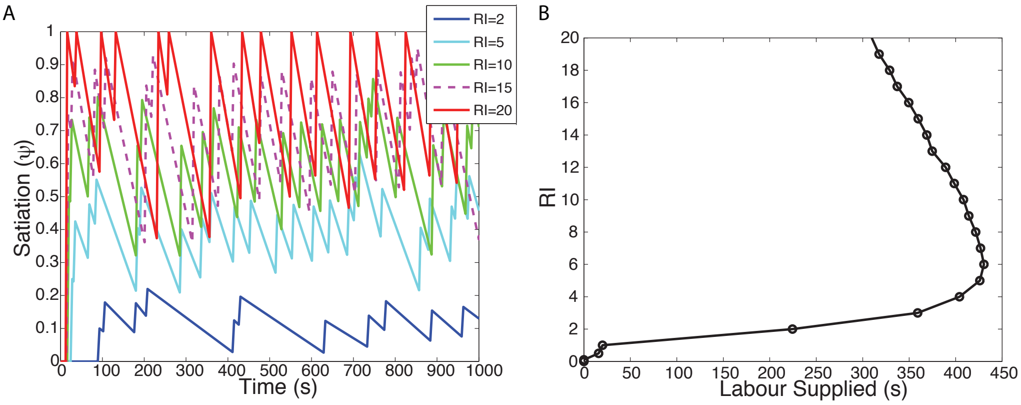


Figure 5.6: Prospective satiation A) The dynamic satiation variable $\psi(t)$ jumps, proportional to the reward intensity (RI), each time a reward is received, and decays exponentially with the duration of a bout in its absence.The greater the reward intensity, the faster the subject becomes satiated. Consequently, the subject will prospectively engage in longer leisure so that it is less satiated when it receives a reward and can enjoy it more. The leisure bouts required to reduce satiation therefore become longer as reward intensity increases. B) Backward bending labour supply curve derived from microscopic prospective satiation. The dependent variable: labour supplied is plotted on the abscissa while the independent variable: reward intensity, is on the ordinate. When the temporal pattern in A) is averaged over trials, the macroscopic time allocation, and hence the labour supplied decreases as the reward intensity increases, yielding the backward bending labour supply curve. For much lower reward intensities (eg. $RI = 2$ shown in the blue trace) the subject does not satiate, as satiation builds up very slowly and is continually reduced by long leisure bouts. For reward intensities in this range, the subject therefore works more to receive more rewards without fully satiating. The macroscopic labour supply curve obtained from averaging across these cases is forward bending. Taken together, this yields a forward and then backward bending labour supply curve. Microscopic marginal utility of leisure $K_L = 0.1$, Price= $4s$, inverse time-constants for satiation decaying ($\alpha_- = 0.99$) in the absence of rewards and its jumping at the time of rewards ($\alpha_+ = 0.05$).

5.4.2 Retrospective satiation / satiation increases the utility of leisure

As noted, a standard microeconomic explanation for the backward bending labour supply curve is that leisure is more beneficial when one is richer. We follow the same principle, but cast in a microscopic version.

We define a microscopic utility of leisure function that increases with the current level of satiation and the duration of leisure: $C_L(\psi, \tau_L)$. We seek to show that at least one such utility function exists that leads to a backward bending labour supply curve. Note that this function may not be unique. Since the degree of satiation increases with the reward intensity, for ease of exposition only, let us break with the previous conventions and set $\psi = RI$. In Chapter 3 we had introduced a sigmoidal microscopic utility of leisure function, whose microscopic marginal utility increases and then decreases. Now consider the *satiation dependent* sigmoidal utility function

$$C_L(RI, \tau_L) = \begin{cases} RI_{min}\ \sigma(\tau_L - C_{Lshift}) & RI \leq RI_{min} \\ g(RI)\ \sigma(\tau_L - h(RI)) & RI > RI_{min} \end{cases}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic function. RI_{min} is a threshold level before which satiation is assumed to be negligible, C_{Lshift} is the shift in the logistic function, and $g(\cdot)$ and $h(\cdot)$ are increasing functions of satiation, which we shall specify such that labour supply curve bends backwards (Fig. 5.7A). $g(\cdot)$ and $h(\cdot)$ are chosen such that the microscopic utility of leisure consistently increases with satiation for all durations of leisure τ_L. That is, the utility of a duration of leisure τ_L will be greater when the subject is more satiated than when it is less. Here, we choose $g(RI) = RI^2$ and $h(RI) = \log(RI)$, although other functional forms may also suffice. For the deterministic, greedy policy we can show (see Section 5.6.1) that the optimal leisure duration τ_L^* increases with $h(RI)$ as long as $RI > RI_{min}$ (Fig. 5.7B, bottom panel)

$$\tau_L^* = \zeta(RI) + h(RI) = \zeta(RI) + h(RI) \tag{5.7}$$

where $\zeta(RI)$ is the logit function, i.e, the inverse of the logistic function. It is negligible compared to $h(RI)$. Consequently, as long as $h(RI)$ is an increasing function, the optimal leisure duration increases with reward intensity.

When microscopic choices are generated according to this optimal leisure duration, and averaged across time, macroscopic labour supplied decreases with

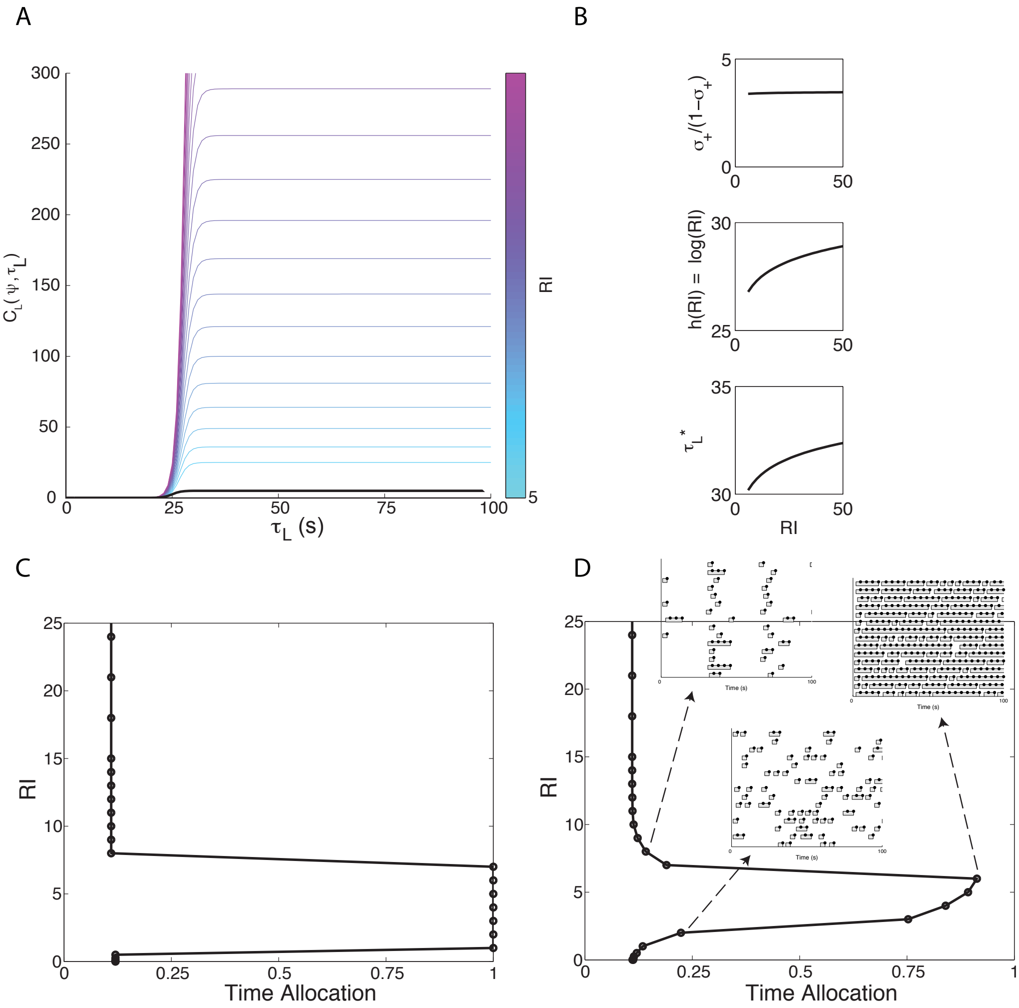


Figure 5.7: Retrospective satiation A) The microscopic utility of leisure $C_L(\psi, \tau_L)$ is sigmoidal, whose maximum and shift increase with the level of satiation (see Eq.(5.7)). Here we assume that, above a threshold level RI_{min}, satiation is simply proportional to reward intensity RI; the maximum of the utility function increases quadratically and the shift logarithmically with RI. Below the threshold RI_{min} the utility of leisure has a fixed maximum at RI_{min} (black curve). B) The optimal leisure duration for the greedy policy τ_L^* (bottom panel) is the sum of $\log[\sigma(\cdot)/(1-\sigma(\cdot))]$ (top panel) and $\log(RI)$ (centre panel, see Eq.(5.7)), where $\sigma(\cdot)$ is the logistic function. Thus, the optimal leisure duration increases with RI. $RI_{min} = 5$,$C_{Lshift} = 25$. C) When leisure bouts are generated accordingly, and the temporal structure is averaged across, we obtain a backward bending labour supply curve (the labour supplied is normalised and here shown in terms of time allocation). The forward bending part of this curve occurs when satiation is below a threshold level and the utility of leisure does not vary with satiation. Then the subject works more as reward intensity increases. D) Using a stochastic, softmax policy rather than a deterministic, greedy one makes the curve smoother. Insets show microscopic ethograms, which are averaged across to yield points on the macroscopic curve.

the reward intensity, producing the backward bending labour supply curve (Fig. 5.7C). If $h(RI) = \log(RI)$, then the optimal leisure duration τ_L^* increases slowly (logarithmically) as reward intensity is increased. If we had used polynomial functions for $h(RI)$, which would make the the optimal leisure duration increase faster with reward intensity, although they would have made the microscopic utility of leisure functions inconsistent, as mentioned above.

When satiation is below its threshold level $RI < RI_{min}$ and hence can be neglected, then we should see labour supply increasing with reward intensity. As as the reward intensity increases, so does the reward rate ρ. The Q-value of leisure is increasingly dominated by the opportunity cost of time $-\rho\tau_L$ and leisure durations consequently become shorter. As discussed in Chapters 3 and 4, consequently, macroscopic labour supply increases with reward intensity, producing the forward bending part of the labour supply curve.

We do not claim that the sigmoidal utility in Eq.(5.7) this is the only microscopic utility of leisure function that can achieve this, but an example of one that does. The forward and backward bending parts of the curve are smoother if we use a stochastic, e.g. softmax, policy (Fig. 5.7D) rather than the deterministic greedy one to generate microscopic choices and average over them.

5.5 Discussion

Although choices about which actions to take and how long to persist with them for may depend on the benefits and costs associated with those dimensions of choice, they may also depend on the recent history of actions and rewards. We explored the case where the choice of whether to work or engage in leisure depended on the recent history of work and leisure and/or reward received, based on the phenomena of physical fatigue and satiation as underling causes. We represented fatigue as a dynamic variable that increases whilst working and decreases during leisure. While we used a low-pass filtered fatigue variable for simplicity, other representations would have been possible. Most research on how fatigue affects decision-making has focussed on cognitive fatigue. An influential account called ego-depletion in which voluntary cognitive effort declines after performing several bouts of forced cognitively demanding tasks Baumeister (2002); Baumeister and Bratslavsky (1998). This decline has been found to be correlated with a depletion of blood glucose Baumeister et al. (2007); Gailliot and Baumeister (2007), although this has been called into question Kurzban (2010). The effect of cognitive fatigue on the labour-leisure tradeoff has been recently studied by

Kool and Botvinick (2012), who showed that human subjects substitute from the less demanding to the more demanding task in accordance with predictions from labour supply theory. A recent study, focussing on how cognitive effort affects the labour leisure tradeoff showed that human subjects would switch from a less cognitively demanding to a more demanding task when they received an income-compensated wage-rate increase for performing the more demanding task, in the form of more M&M candies, in line with the predictions from labour supply theory Kool and Botvinick (2012). However, it remains to seen whether performing a less cognitively demanding task is actually leisure or just less demanding work. Here we developed a novel, normative microscopic theory of physical fatigue.

We attempted to tease apart two ways in which fatigue may affect cost-benefit decisions: by making working more costly when fatigued or leisure more beneficial. Both led to a microscopic behavioural prediction of runs of work bouts interspersed by short leisure bouts as fatigue accumulated, and was alleviated by a long leisure bout. This would lead to autocorrelation between consecutive leisure bouts, making them non independent and identically distributed. This is an important behavioural prediction which can be tested for when microscopic work and leisure data are analysed. The two cases of fatigue that we assumed can be further distinguished by appropriate experiments designed to tease them apart. The cost of working whilst fatigued can be explored by forcing subjects to work for an employer determined period, and asking them to choose how long to rest prospectively, taking into account how fatigued they will be when the start working otherwise. If the period of work is long, and the costs of working whilst fatigue are high, then we predict that subjects should prospectively engage in long leisure. Such prospective fatigue is akin to the common problem of choosing how long to rest on a Sunday *before* a long week.

The more intuitive case of fatigue is when it makes leisure more beneficial. That is, leisure is more beneficial *because* of fatigue. This is akin to the case of choosing how long to rest on a Saturday *after* a long week. This case implies that the marginal utility of leisure is dependent on that of the recent history of working (manifested as fatigue), contrary to the assumptions in Chapter 3. By starting from a very simple, linear momentary microscopic utility of leisure, which increased with fatigue, we derived an integrated microscopic utility of leisure which was concave in its duration, but whose utility was greater if the subject was more fatigued. We thereby derived, rather than assumed two properties of utility of leisure functions popular among economists: a decreasing marginal utility, favouring many short leisure bouts over one long one, and an interaction between the marginal utilities of work and leisure. The former implies that a preference for many short leisure bouts arise from actual indifference, but owing to fatigue. The second implies that work and leisure are microscopically imperfect substitutes as well.

We then explored the curious case predicted by labour supply economics that time allocated to working can decrease rather than increase as wage-rates are increased without adjusting for the increased income possible. This can happen when income effects dominate substitution effects. This is akin to the common notion that the subject can purchase more leisure with the increased income, and thus spends more time engaging in leisure. Consequently, the labour supply curve is backward bending. Both laboratory experiments and behavioural economic studies have showed evidence for this backward bending curve. Camerer et al. (1997) studied this in New York City taxi drivers, proposing an algorithmic mechanism that drivers have an income target in mind and quit when it is reached. Since the target is more quickly attained on high wage-rate days (eg. when the subway is on strike), drivers work less on those days. They posited that having a daily target in mind assists the driver in two self-control problems. Firstly, it provides a simple heuristic of when to quit for the day, reducing the computational complexity of calculating when sufficient work has been done based on the wage rate. Secondly, drivers could intertemporally substitute labour: working more on a high wage-rate day (eg. a day where the underground/subway is on strike) and saving that earned income for low wage-rate days. However, as Camerer et al. (1997) claim, driving around New York City with $250-300 would tempt the driver into spending that income immediately. Income targeting thus acts as a precommitment device preventing drivers from spending their daily income on temptations.

Camerer et al. (1997)'s findings involved purely correlating macroscopic quantities: daily hours worked and daily wage rate. Their claims of income targeting were contradicted by Farber (2005)'s who studied the probability of quitting in a given hour, based on how long the driver had driven and how much income s/he had earned so far. Farber (2005) found that the probability of quitting depended more on the cumulative hours worked so far. While these studies were purely correlational, Fehr and Goette (2007) and Goette and Huffman (2006) conducted more causal experiments in bicycle messengers, whose commission rates were increased compared to a control group. Goette and Huffman (2006) provided the cleanest evidence of income targeting. Their miniscopic analysis showed that messengers who received a higher commission rate worked more and quit early in the day; and the subsequently reduced labour compensated for this—keeping daily labour supply unaffected.

Animal experiments from behavioural psychology have shown that miniscopic response rates show a bitonic profile, increasing at the beginning of an experimental session and decreasing thereafter. The decrease in response rates were faster for the higher reinforcer rates McSweeney (1992), durations and sizes Bizo et al. (1998). This is similar to the finding of Goette and Huffman (2006) in bicycle messengers. The reduction in response rates in animal experiments were attributed to satiation Fischer and Fantino (1968); McSweeney et al. (1991); McSweeney and Roll (1993); McSweeney and Johnson (1994); Killeen (1995)), and were teased apart from habituation (see McSweeney and Murphy (2000) for a review). The findings suggested that subjects satiate faster for greater reward utilities (owing to size, frequency or duration of the reward).

Whether macroscopic or miniscopic, these studies still investigate average times and response rates, and ignore the crucial durations of leisure in between work bouts. We developed a normative microscopic theory of dynamic satiation to derive the macroscopic backward bending labour supply curve. We once again studied two cases: satiation reducing the net utility of a reward (e.g. a hungry animal will work more than a satiated one, Dinsmoor (1952)) or making leisure more beneficial. As for the case of fatigue, these can be distinguished by considering prospective and retrospective decisions. By considering satiation to increase when rewards are received, we noted how subjects would be more quickly satiated when reward intensities were larger. This would need to be alleviated by long leisure bouts. Consequently, when averaged macroscopically, we obtained a backward bending labour supply curve from normative, microscopic principles.

The case of leisure being more beneficial due to satiation implies an interaction between the marginal utilities of leisure and rewards, as assumed in labour supply theory (leisure is more beneficial because one is richer). This is the macroscopic assumption underlying the backward bending labour supply curve. We showed that a microscopic utility function, whose utility increases with the duration of leisure and degree of satiation, that can generate the backward bend of the curve. We proved that as satiation increases with reward intensity, so does the optimal leisure duration—consequently reducing the amount of labour supplied. Although we used a logistic utility function, it is quite possible that other functions exist.

We thus provided a novel, normative and microscopic reason underlying the macroscopic backward bending labour supply curve–owing to microscopic satiation. The income targeting assumptions of Camerer et al. (1997) are somewhat ad-hoc. A question becomes, what constitutes the income target. This could be used when the income earned from work can be saved and then spent on essential commodities and leisure activities Dupas and Robinson (2013). Once sufficient quantities of the latter can be guaranteed, there is no need to earn further income from work. Income-targeting is an extreme case of our normative, microscopic formulation of satiation, one for which the net utility of a reward or marginal utility of leisure drops to zero as soon as the income target is reached, rather than merely decaying with leisure.

An alternative is to consider income-targeting as the solution to an optimal stopping problem. Suppose subjects do not save their earned daily income for future days. Then since there is little point working so long that there is no time left in the day to spend the earned income on leisure, subjects will quit working when the utility of leisure for the remaining time exceeds the marginal utility of income. Although this is a viable alternative, its dynamics are not as detailed as the case of satiation we have considered. Finally, much like Niv et al. (2007), we could consider a subject considering choosing its labour according to the average reward rate in the environment. In an environment where high reward rates (e.g. due to subway strikes) are rare, subjects should normatively work more and give up early on days with a high reward rate, compared to usual days in which the reward rate is low. Such a model would be highly rich, but complicated; we leave it for future consideration.

Finally, we can use our normative, microscopic theories to explore the neural underpinnings of cost-benefit decision-making under fatigue and satiation. Previous laboratory and field experiments with animals and humans have studied cost-benefit decisions involving physical effort costs, employing higher ratio schedules of reinforcement (higher number of lever presses), heavier weights of levers, higher metabolic requirements, longer travelling distance during foraging, higher barrier climbing and greater hand force investments Collier and Levitsky (1968); Collier et al. (1975); Kanarek and Collier (1973); Floresco et al. (2008a); Stevens et al. (2005); Prévost et al. (2010); Salamone et al. (1994); Salamone and Correa (2002); Salamone et al. (2007); Rudebeck et al. (2006); Croxson et al. (2009); Walton et al. (2002, 2003, 2005, 2006, 2007, 2009); Phillips et al. (2007); Kennerley et al. (2006, 2009); Hosokawa et al. (2013); Gan et al. (2010); Wanat et al. (2010); Floresco and Ghods-Sharifi (2007); Floresco et al. (2008b); Ghods-Sharifi and Floresco (2010); Kurniawan et al. (2010, 2013). The neuromodulator dopamine, along with the striatum and anterior cingulate cortex (ACC) have been implicated in such decisions.

Depletions of dopamine shift preference away from higher rewards when tasks require climbing a barrier to attain them, without affecting preference when such energetic costs are minimal Salamone et al. (1994); Salamone and Correa (2002); Salamone et al. (2007). There are distinctions within the ventral striatum, with a stronger effect of dopamine depletions on effort-based decision-making at the core than shell subregion of the nucleus accumbens Ghods-Sharifi and Floresco (2010). However, measuring sub-second phasic dopamine release onto the core of the accumbens showed correlations with rewards but not anticipated effort costs Gan et al. (2010); Wanat et al. (2010).

Human fMRI studies have showed that activity in the dorsal striatum is correlated with the anticipated effort of an action Croxson et al. (2009). Further, higher dorsolateral striatal activity (especially in the putamen) is observed when choosing low compared to high effort options in a physical effort task Kurniawan et al. (2010), and higher ventral striatal activity during a low cognitive demand block compared to a high cognitive demand block in a mental effort task Botvinick et al. (2009).

Lesion studies in rodents performing T-maze tasks consistently show impairments in effort-based decision-making following ACC lesions. Similar to dopamine depletions, ACC lesions lead to in a shift in preference away from the high reward arm when a barrier is required to be climbed to attain it Walton et al. (2002, 2003, 2009); Rudebeck et al. (2006). This shift in preference was not due to immobility as normal preference is restored when both arms require equal effort Floresco and Ghods-Sharifi (2007); Walton et al. (2002, 2003, 2009). Further, monkeys with ACC lesions were impaired in choosing appropriate responses that required the integration across past contingencies between actions and rewards Kennerley et al. (2006). Monkey single-cell recordings Kennerley et al. (2009); Hosokawa et al. (2013) and human imaging experiments with passive action valuation Croxson et al. (2009); Prévost et al. (2010); Kurniawan et al. (2013) or mental loads Botvinick et al. (2009) have also revealed increased ACC activity with increasing anticipated effort.

While the neural basis of physical and cognitive effort costs have been examined, they have been confounded with possible effects of fatigue Meyniel et al. (2013). Whilst we considered a simple cost of working whilst fatigued, it is possible that this could interact with effort (see Meyniel et al. (2014)), and a neural signature of these has been found in the insula Meyniel et al. (2013). The two make different predictions behaviourally: effort costs would lead to shorter work bouts whilst working whereas fatigue would lead to runs of long work bouts interspersed by short leisure breaks. Studying their interaction and neural basis could be useful both for basic neuroscience and for sports medicine. Additionally, our predictions for prospective fatigue can be readily extended to investigations of anticipated effort costs. Similarly, the neural basis of satiation and its effect on decision-making can be studied by comparing the effects of satiating (eg. food, water) rewards against non-satiating (eg. BSR) ones.

5.6 Appendix

5.6.1 Optimal duration of leisure increases with reward intensity

Consider the microscopic utility function in Eq. (5.7), which depends both on the duration of leisure τ_L and initial level of satiation ψ. The Q-value of engaging in leisure is

$$Q^\pi([\text{post}, \psi], [L, \tau_L]) \;=\; C_L(\psi, \tau_L) - \rho^\pi \tau_L + V^\pi([\text{pre}, \psi']) \tag{5.8}$$

Since we are considering retrospective satiation, and assuming that satiation only affects the utility of leisure but not that of rewards, we may neglect the value of the (less-satiated) pre-reward state the subject transitions to $V^\pi([\text{pre}, \psi'])$. Further, since the degree of satiation increases with the reward intensity, for ease of exposition only, let us assume $\psi = RI$. We wish to find the optimal duration of leisure τ_L^*.

As we assumed in Eq. (5.7), as long as satiation is above a threshold level ($RI > RI_{min}$), the microscopic utility of leisure $C_L(RI, \tau_L) = g(RI)\ \sigma(\tau_L - h(RI))$, where $\sigma(\cdot)$ is the logistic function. We need only solve for $\sigma(\cdot)$; with $\tau_L^* = \sigma^{-1}(\cdot) + h(RI) = \log\left[(\sigma(\cdot)/(1-\sigma(\cdot))\right] + h(RI)$ as in Eq. (5.7).

For the optimal duration of leisure (under the greedy policy π^*), the marginal utility of leisure is approximately equal to the reward rate

$$\begin{aligned} \frac{\partial Q\left([\text{post}, RI], [L, \tau_L]\right)}{\partial \tau_L} &= 0 \\ \frac{\partial C_L(RI, \tau_L)}{\partial \tau_L} - \rho &\approxeq 0 \end{aligned} \tag{5.9}$$

which, for the utility function we considered implies

$$\begin{aligned} g(RI)\ \sigma(\cdot)\ (1-\sigma(\cdot)) &= \rho \\ &= \frac{RI + C_L(RI, \tau_L)}{P + \tau_L} \\ &= \frac{RI + g(RI)\ \sigma(\cdot)}{P + \zeta(RI) + h(RI)} \end{aligned} \tag{5.10}$$

where $\zeta(RI) = \sigma^{-1}(\cdot)$ is a negligible term at higher reward intensities. Simplifying Eq. (5.10), $\sigma(\cdot)$ is the solution of a quadratic equation, with two roots (which we may denote $\sigma_+(\cdot)$ and $\sigma_-(\cdot)$. As long as the roots exist, and the solution $\tau_L^* \geq 0$, we can conclude that $\tau_L^* = \log\left[(\sigma(\cdot)/(1-\sigma(\cdot))\right] + h(RI)$, which is dominated by $h(RI)$. Since $h(RI)$ is an increasing function of reward intensity, and satiation increases with reward intensity, the optimal duration of leisure τ_L^* increases with reward intensity.

Chapter 6

The microscopic utility of leisure

6.1 Introduction

In previous chapters, we had developed a normative, microscopic theory–of how a subject chooses to work or engage in leisure. Central to this theory was the microscopic utility of leisure function, which reflected the subject's innate preference for durations of leisure, irrespective of all other rewards and costs. While much research has been devoted to quantifying the subjective value of external rewards, the intrinsic utility of leisure is yet to be quantitatively and empirically studied. This is the intent of this Chapter.

Since we are interested in isolating a subject's preference for leisure, independent of satiation, we study data collected from rat subjects working for brain stimulation reward, which is widely believed not to satiate. Effort costs were controlled by requiring that the rats depress a very light lever. Finally, fatigue was controlled for by allowing the subjects to rest in between trials. We quantitatively fit a utility function for leisure to an entire dataset of microscopic behaviour across different experimenter determined conditions, and thereby infer which functional form best accounts for the observed experimental data.

6.2 Experiment

We analyse the data collected by Rebecca Solomon from 6 rat subjects. We introduced some of the experimental details in Chapter 3; for completeness, we reiterate them here and additionally mention those specific to this dataset. Subjects worked for BSR rewards by depressing a very light lever on a CHT task. BSR rewards of a certain reward intensity (RI, which we somewhat arbitrarily assume to be between 0 and 5; restricting it between 0 and 1 could lead to numerical problems, while if the maximum is too large then reward rates would be too high to allow our distributions to be fit) were provided when the experimenter determined price was attained. The lever was then retracted and reintroduced following a 2s delay during which the trial time-clock was frozen. These 2s delays were not counted within the trial duration. Trial duration was 25 × price for prices longer than or equal to 1s, and 25s for prices shorter than 1s. The latter was used to ensure there was sufficient data on such short prices. Leisure bouts that did not immediately follow after a reward and which were shorter than 1 second are considered to be part of the previous work bout (since subjects remain at the lever during this period). *Graphically*, this makes some work bouts *appear* longer than others. Data were recorded at a precision of 0.1s.

In the random world environment, subjects faced triads of trials: 'leading', 'test', then 'trailing'. Leading and trailing trials involved maximal and minimal stimulation frequencies (reward intensities) respectively, and a 1s price. Each trial was separated by a 10s cue during which house-lights are switched on. We analyse the sandwiched test trials, which span a range of prices and reward intensities. Leading and trailing trials allow calibration, so subjects can stably assess *RI* and price on test trials. Subjects tend to be at leisure on trailing trials, limiting physical fatigue. Subjects repeatedly experience each test reward intensity and price over many months, and so can readily appreciate them after minimal experience on a given trial without uncertainty. However, on a given test trial, subjects are not initially aware of the RI and price, and must actively infer them by working early on in the trial. We call this period a sampling period, and preprocess the data to exclude data in this period.

On a given test trial, the RI and price were generated from one of 9 pseudo sweeps through parameter space (we shall henceforth refer to them as 'sweeps'). These are pesudosweeps since the RI and price are actually not swept from trial to trial, but randomly sampled from the relevant parts of the parameter space. There are 7 RI 'sweeps' in which the price is fixed while the RI is increased. These prices were 0.125s, 0.25s,0.5s,1s,2s,4s and 8s (red ray in Fig.6.1, upper panel for a price of 4s and rays with cool colours corresponding to increasing prices for the others). The price 'sweep' (blue ray in Fig.6.1, upper panel) involves the RI remaining fixed at its highest value while the price is decreased. Finally, the radial 'sweep' (green ray in Fig.6.1, upper panel) involves both RI increasing and price decreasing.

6.2.1 Preprocessing data

We preprocessed the data, ignoring work and leisure bouts on a trial when the subject is sampling: trying actively to infer the reward intensity and price; we only consider the subjects' choices when they know the reward intensity and price with certainty. Work bouts during the sampling phase are coloured yellow and data up to the end of the last of such work bouts are excluded from subsequent analyses. We shall see that this only discards very little data.

6.2.2 Macroscopic Time Allocation

As shown in Fig. 6.1, macroscopic time allocation increases with reward intensity and largely decreases with price for all 'sweeps'. A new prediction we had made in Chapter 3 is that time allocation would be observed not to decrease, and even increase with the price, a prediction not made by any existing macroscopic model. Whereas animals have been previously shown to consistently work more when work-requirements are greater (eg. ostensibly owing to sunk costs Kacelnik and Marsh (2002)), the apparent anomaly we discussed only occurred at very long prices, and was unexpected from a macroscopic perspective, but revealing from a microscopic perspective. We tested whether this is experimentally true. For a fixed, high reward intensity, time allocation is indeed observed to increase rather than decrease with price at the highest prices (see the price 'sweep': blue curve in Fig. 6.1, lower panel).

6.2.3 Ethograms

Figs. 6.2, 6.3 and 6.4 depict ethograms of subject F9, revealing the microstructure of choices. As mentioned in Chapter 3, the microscopic characteristics of the data include: (i) at high payoffs, subjects work almost continuously, engaging in little leisure inbetween work bouts; (ii) at low payoffs, they engage in leisure all at once, in long bouts after working, rather than distributing the same amount of leisure time into multiple short leisure bouts; (iii) subjects work continuously for the entire price duration, as long as the price is not very long (compare the short e.g. Price= 2.8s with the longest prices: Price=57.5s,34.7s on the price 'sweep' shown in blue in Fig.6.2); (iv) the duration of leisure bouts is variable. At the very long prices (e.g. Price=57.5s, 34.7s–arranged in order of payoff–on

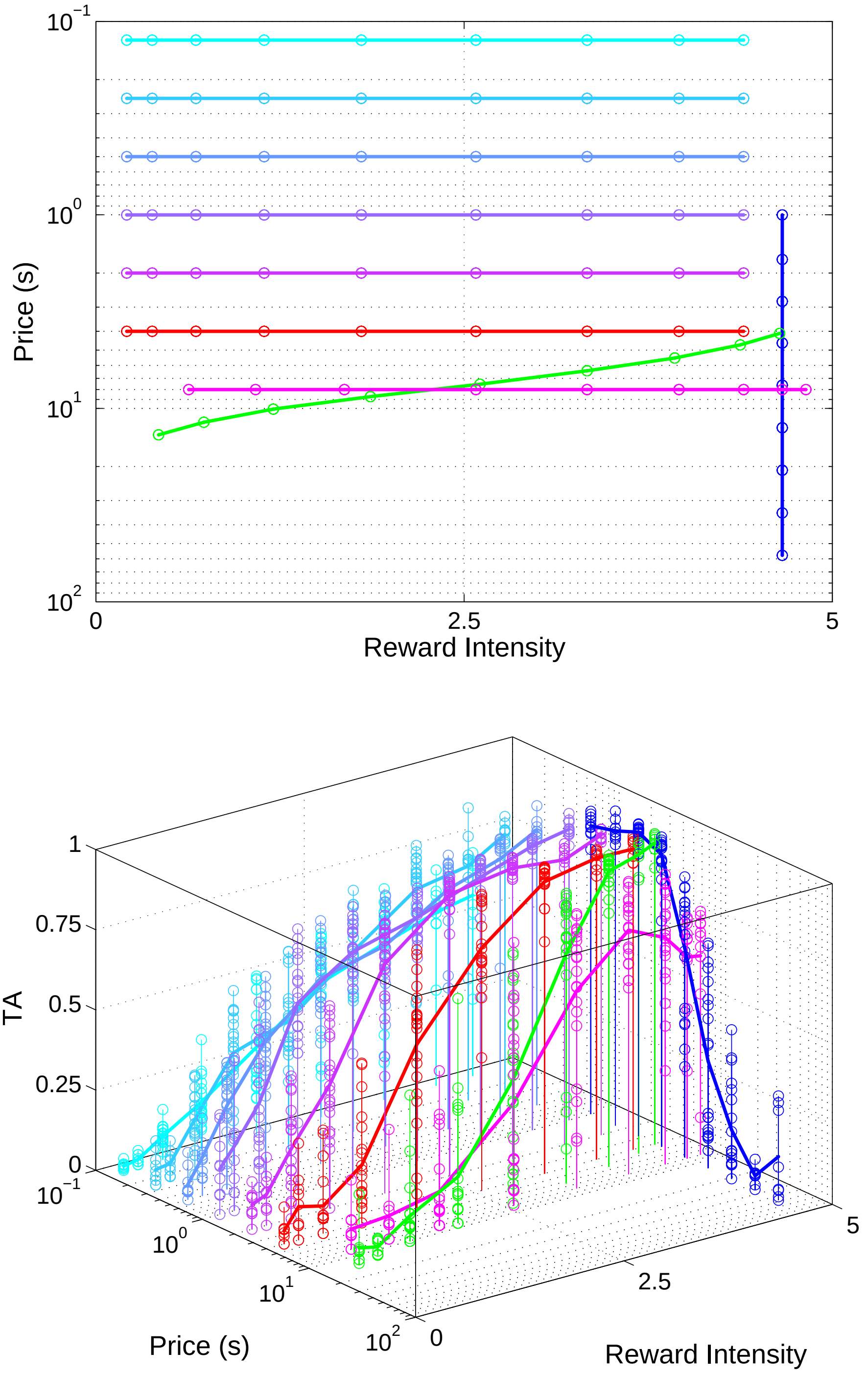


Figure 6.1: Reward Intensity and Price pseudo sweeps and Time Allocation for subject F9. Upper panel: Reward Intensity and prices were generated from one of 9 pseudo sweeps through parameter space. For the 7 RI 'sweeps', the price is fixed while the RI is increased. These prices were 0.125s, 0.25s,0.5s,1s,2s,4s and 8s (red ray shows a price of 4s and rays with cool colours showing increasing prices for the others). The price 'sweep' (blue ray) involves the RI remaining fixed at its highest value while the price is decreased. Finally, the radial 'sweep' (green ray) involves both RI increasing and price decreasing. Lower panel: Time allocation: proportion of the trial duration allocated to working. Note that time allocation increases rather than decreasing with price at the highest price of the price 'sweep'.

the price 'sweep' shown in blue, or Price= 13.7s for the lowest payoff in the radial 'sweep' shown in green), subjects engage in a long leisure bout before resuming working. Integrating and averaging these across time leads to the macroscopic time allocation not decreasing, but increasing as price is increased (Fig.6.1, lower panel). Subjects work more on trials with very short prices even when the reward intensity is low (compare Prices of 0.125s and 0.25s or 1s in Figs. 6.3 and 6.4 with that of 4s in Fig.6.2). Finally, on medium payoffs, subjects display a pattern of working with short leisure bouts interspersed in between work bouts, followed by a long leisure bout and a resumption of this pattern, unless this leisure bout is censored by the end of the trial (this pattern is clearer for medium e.g. 8s prices, Fig.6.4).

6.3 Procedures for fitting microscopic data

We wish to recover the microscopic utility of leisure function $C_L(\tau_L)$, which is innate to an individual subject, and is assumed to be dependent only on the duration of leisure and independent of all other rewards and costs. In Chapter 3 we showed how, given a $C_L(\cdot)$, the leisure duration distribution changes as the reward intensity and price are manipulated. We therefore fit one $C_L(\cdot)$ for each subject across all reward intensity and price conditions.

6.3.1 Microscopic utility of leisure

We are interested in characterising an animal's preference for leisure, and quantifying it relative to its utility of reward (RI). While there is an infinite number of functional forms for the utility of leisure that we could have considered, the canonical forms we considered in Chapters 3 and 4: linear, concave (we use a logarithmic function for simplicity, since as we showed in Chapter 4, it leads to gamma distributed leisure durations) and sigmoid make starkly different predictions about a subject's preference for leisure durations. We therefore assumed that these would suffice for inferring subject's preferences. For each subject we tested which of these fit the data best. The linear and concave utilities each have only one parameter: the (maximal) slope K_L, whereas the sigmoid has two more: $C_{L_{max}}$ and $C_{L_{shift}}$ are the maximal utility and shift, i.e. a total of three parameters.

Since a sigmoid implies a constrained maximum at which the utility of leisure saturates, we also used a weighted combination of a linear and sigmoid function

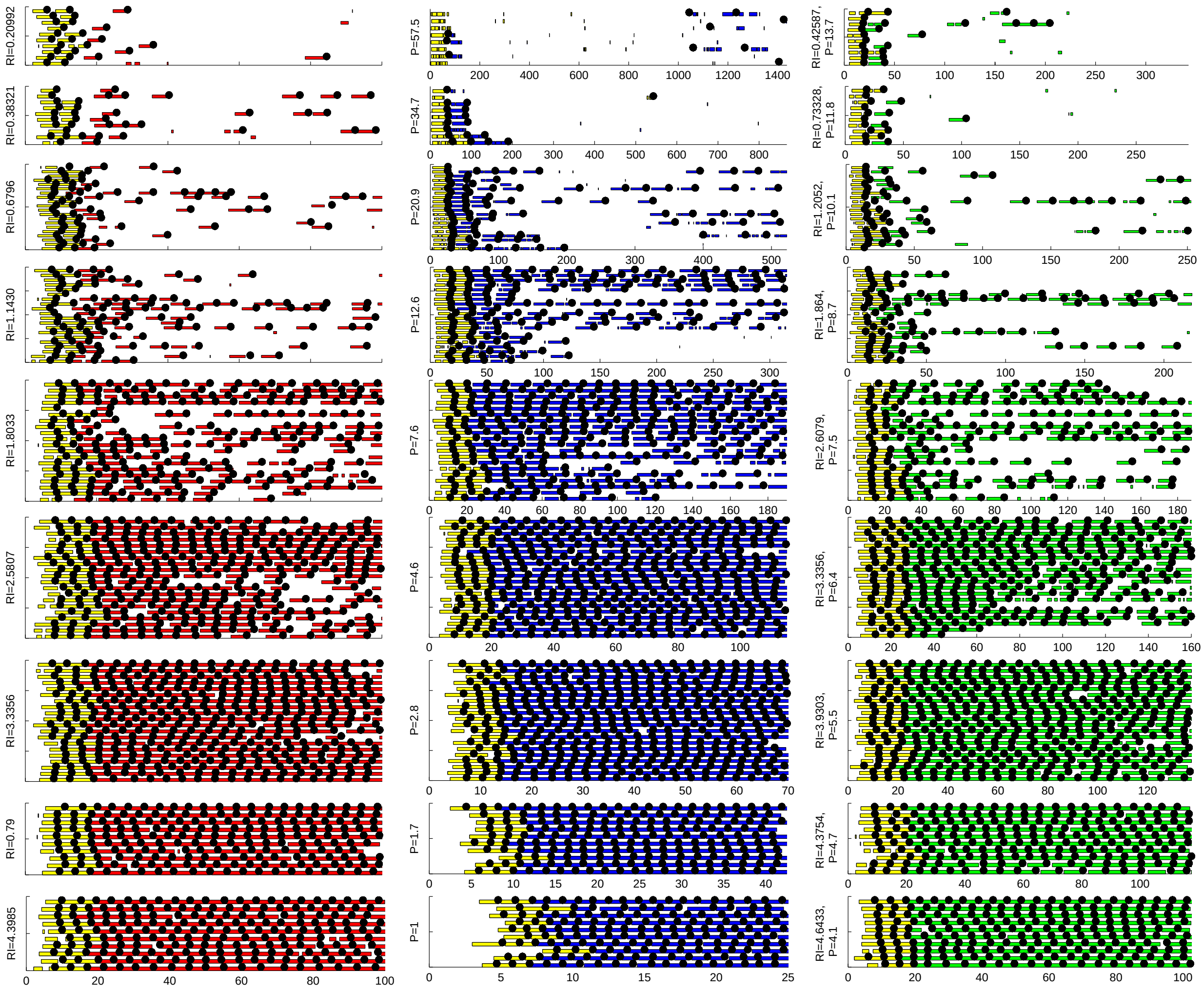


Figure 6.2: Ethograms for subject F9. Price= 4s RI sweep, Price sweep and Radial sweep. Payoff increases from top to bottom. Coloured bars show work, white spaces show leisure and black dots show reward delivery. Red (left column), blue (middle column) and green (right column) show Price=4s RI 'sweep', price 'sweep' and radial 'sweep', respectively (see Fig. 6.1, upper panel). Trial duration is 25 × price. Leisure bouts that do not immediately follow after a reward and which are shorter than 1 second are considered part of the previous work bout (since subjects remain at the lever during this period). *Graphically*, this makes some work bouts *appear* longer than others. Work bouts during the sampling phase when the subject does not know the reward intensity and price with certainty are coloured yellow and data up to the end of the last of such work bouts are excluded from subsequent analyses.

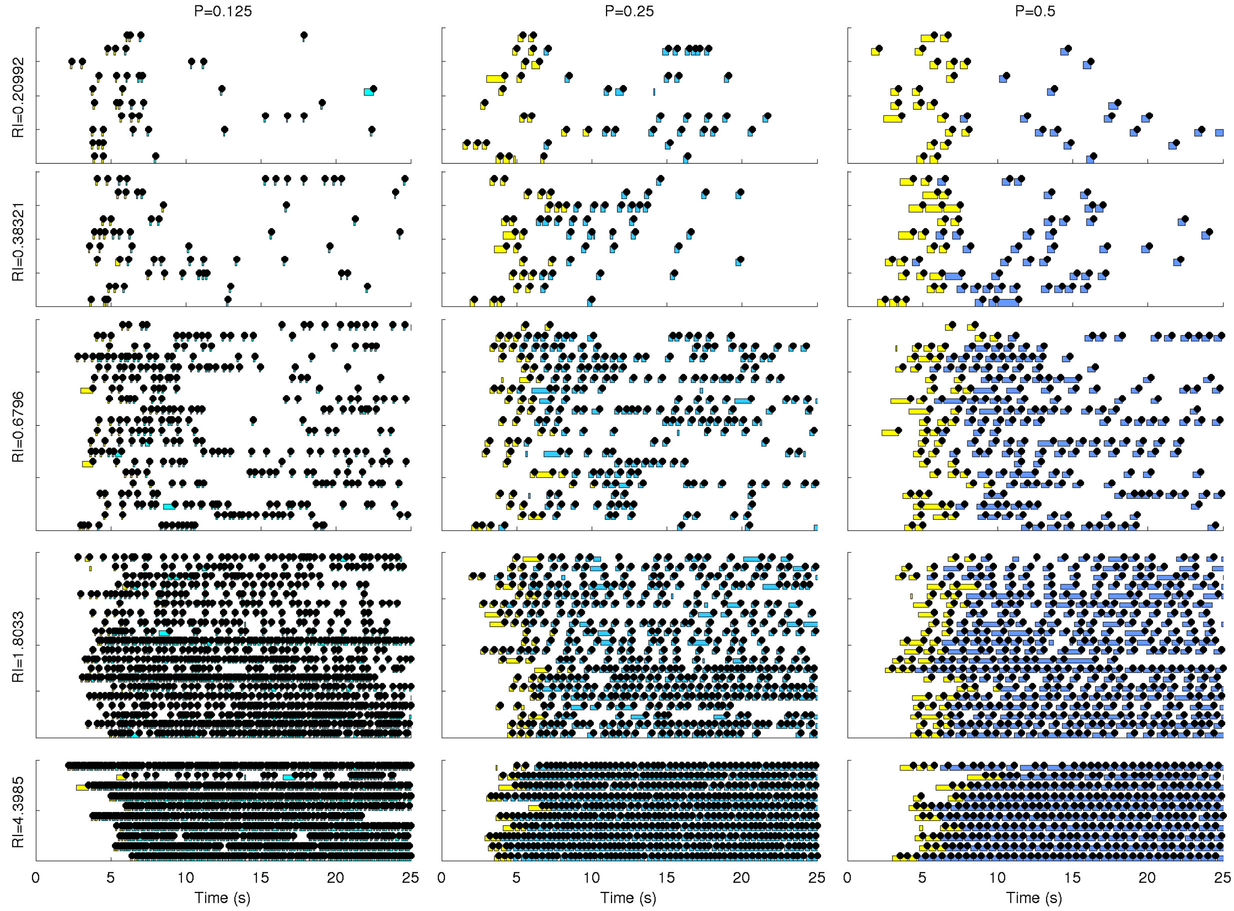


Figure 6.3: Ethograms for subject F9. RI sweeps for sub-second prices. Payoff increases from top to bottom. Coloured bars show work, white spaces show leisure and black dots show reward delivery. Light cyan to blue colours show RI sweeps with Price=0.125 (left column) , 0.25s (middle column) and 0.5s (right column) , respectively (see Fig. 6.1, upper panel). Trial duration is 25s to enable sufficient data to be collected. The rare leisure bouts that do not immediately follow after a reward and which are shorter than 1 second are considered part of the previous work bout (since subjects remain at the lever during this period). Only the three lowest, one medium and the highest payoffs are shown. Other conventions same as in Fig. 6.2.

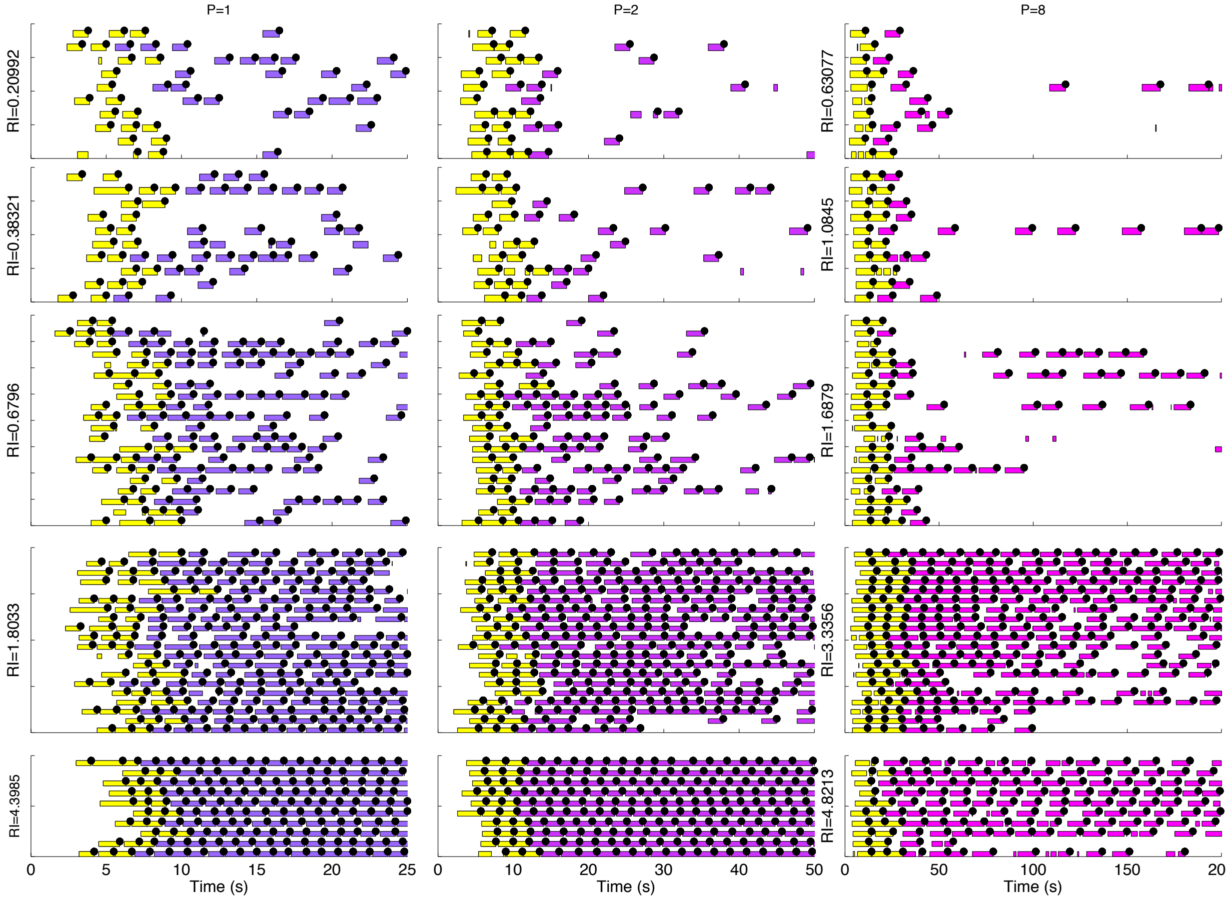


Figure 6.4: Ethograms for subject F9. RI sweeps for Price=1s, 2s and 8s. Payoff increases from top to bottom. Coloured bars show work, white spaces show leisure and black dots show reward delivery. Purple to magenta colours show RI sweeps with Price=1s (left column), 2s (middle column) and 8s (right column), respectively (see Fig. 6.1, upper panel). Only the three lowest, one medium and the highest payoffs are shown. Other conventions same as in Fig. 6.2.

to represent the class of initially supra-linear but eventually sub-linear utility functions.

$$C_L(\tau) = \alpha\ K_L\ \tau + (1-\alpha)\ \frac{C_{L_{max}}}{1+\exp\Big[\ -4\frac{K_L}{C_{L_{max}}}(\tau - C_{L_{shift}})\ \Big]} \tag{6.1}$$

$\alpha \in [0,1]$ is the weight on the linear component (see Fig. 3.3B). This extra parameter (the weight α) enables greater flexibility of fitting the data. We penalise this extra flexibility afforded by more complex models by reporting BIC scores.

6.3.2 Pavlovian component of leisure τ_{Pav}

In Chapter 3 we had assumed a deterministic Pavlovian component of leisure $\tau_{Pav} = f_{\text{Pav}}(RI, P)$, which decreases with payoff – i.e., increases with price and decreases with reward intensity (Figure 3.3C). However, when fitting (uncensored) data this implies a constraint on τ_{Pav} such that the PRP is at least as long as the Pavlovian component. This is specifically detrimental for high reward intensity and medium length prices, since τ_{Pav} has to be curtailed to values as low as 0.1. A more reasonable assumption is a probabilistic $\tau_{Pav} \sim Pr\,(\tau_{Pav}|f_{\text{Pav}}(RI, P))$. We could then convolve $Pr(PRP) = Pr(\tau_L + \tau_{Pav}) = \pi(L, \tau_L|\vec{s} = \text{post}) * Pr\,(\tau_{Pav}|f_{\text{Pav}}(RI, P))$, allowing us to easily fit arbitrarily small PRPs any given model of τ_{Pav} . We wish to minimise the number of additional free-parameters for τ_{Pav}. If we assume τ_{Pav} to be exponentially distributed (a one parameter distribution) with mean $f_{\text{Pav}}(RI, P)$, and if instrumental leisure is exponentially distributed as well (as is true for high payoffs, or in general, the exponential part of the bimodal distribution) then the convolved PRP distribution is gamma distributed (or its left mode is gamma distributed). Under the assumption that Pavlovian leisure is dominated by instrumental leisure at low payoffs, we bound τ_{Pav} at a maximum of 10s. For simplicity we assume f_{Pav} to be a sigmoidal function of the inverse payoff $\frac{P}{RI}$. We fix the slope of this function at a gentle 0.25 allowing τ_{Pav} to cover the entire range from 0s to 10s for the payoffs used experimentally. This leaves us with only one free parameter, the shift: Pav_{shift} to fit. Larger values of Pav_{shift} imply shorter mean Pavlovian components of leisure. While our assumptions are slightly arbitrary in the absence of independent data, they can be justified as above, and leave one free parameter to be fit.

6.3.3 Policies

As in Chapter we assume a soft-max choice rule over Q-values to generate the data. Since quantities constituting the Q-values can be scaled, we set the inverse-temperature parameter β to 1 the soft-max policy without loss of generality.

6.3.4 Likelihoods

For each subject, suppose we have D combinations of (RI, P) conditions tested by the experimenter, each comprising K_d iid trials each. For brevity of notation, let us denote $y_d = \{RI_d, P_d\}$ as a condition. Also suppose that on each trial k_d, we observe a sequence of N_{k_d} PRPs ($\{PRP\}_{i_{k_d}}$) and pre-reward work and leisure bouts ($\{\tau_W \ , \ \tau_L\}_{\vec{s}=[\text{pre},w]_{i_{k_d}}}$), where $i_{k_d} = 1, \ldots N_{k_d}$ is each individual bout and $\vec{s} = [\text{pre}, w]$ is a pre-reward state with the subject having worked for a cumulative time $w \in [0, P)$ of the price.

For an individual subject, the likelihood of observing the microscopic sequence of work and leisure bouts, across the entire dataset is, given a set of model parameters $\vec{\theta}$

$$l(\vec{\theta}) = \prod_{d=1}^{D} \prod_{k_d}^{K_d} Pr\left(\left[\{PRP\}_{i_{k_d}}, \{\tau_W \ , \ \tau_L\}_{\vec{s}=[\text{pre},w]_{i_{k_d}}}\right]_{i_{k_d}=1}^{N_{k_d}} \mid y_d, \vec{\theta}\right) \tag{6.2}$$

The normative, microscopic model in Chapter 3 assumes, in the absence of fatigue or satiation, all PRPs and pre-reward work and leisure bouts are independent. We shall test whether this assumption is valid in Section 6.6. We can therefore re-write this likelihood in terms of the distributions of PRPs: $Pr(\tau_L + \tau_{Pav})$, and pre-reward work $\pi(\tau_W \mid w)$ and leisure durations $\pi(\tau_L \mid w)$.

$$l(\vec{\theta}) = \prod_{d=1}^{D} \prod_{k_d}^{K_d} \prod_{i_{k_d}}^{N_{k_d}} Pr\left(\{PRP\}_{i_{k_d}} | y_d, \vec{\theta}\right) \cdot \pi\left(\{\tau_W \mid w\}_{i_{k_d}}; y_d, \vec{\theta}\right) \cdot \pi\left(\{\tau_L \mid w\}_{i_{k_d}}; y_d, \vec{\theta}\right) \tag{6.3}$$

6.3.5 Censoring

Since trials end at 25 × *price*, some bouts are censored, i.e. curtailed by the end of the trial. Furthermore, work bouts intended to be longer can be curtailed by the price being attained and the lever retracted. Since the duration of such a censored bout can be anything that is longer than the observed duration when it was curtailed, we use the cumulative distribution function rather than the probability distribution function for these bouts. Specifically, a likelihood comprising both uncensored and censored durations is

$$l(\vec{\theta}) = \prod_{d=1}^{D} \prod_{k_d}^{K_d} \prod_{i_{k_d}}^{N_{k_d}} Pr(X_{i_{k_d}} = x_{unc,i_{k_d}}) \cdot \Big(1 - Pr(X_{i_{k_d}} \leq x_{cens,i_{k_d}})\Big) \tag{6.4}$$

where $X_{i_{k_d}}$ is a random variable, $x_{unc,i_{k_d}}$ and $x_{cens,i_{k_d}}$ are the observed durations for uncensored and censored bouts.

6.4 PRP fits using 2-state microSMDP

As shown in Fig. 6.2, and Chapter 3 subjects work continuously for the entire price duration, as long as the price is not very long. Leisure bouts are therefore mostly post-reward. Similar to Chapters 4 and 5 we thus start from the initial assumption that subjects work continuously for the entire price duration. We then use the simplified, 2-state micro-SMDP to model PRPs only. The likelihood in Eq. (6.3) reduces to

$$l(\vec{\theta}) = \prod_{d=1}^{D} \prod_{k_d}^{K_d} \prod_{i_{k_d}}^{N_{k_d}} Pr\left(PRP_{i_{k_d}} | y_d, \vec{\theta}\right) \tag{6.5}$$

τ_L was discretized into 1s time-steps up to a total of 2000s. We did not use a prior probability density for durations as for the full micro-SMDP model laid out in Chapter 3. In the full model, since the policy is over all action-durations ($[a, \tau_a]$), irrespective of whether they are of work and leisure, arbitrarily long leisure durations would have a greater effect on the reward rate than work durations. Including a prior that makes longer leisure durations less likely to be chosen normalises the contributions of durations of work and leisure to the reward rate, affording both an equal role. However, in the case of our simplified 2-state SMDP, the subject chooses to work pre-reward and engage in leisure post-reward, keeping these choices separate. The reward rate in our 2-state SMDP without any priors for durations was similar to that for the full micro-SMDP model with an exponential prior for leisure durations, indicating that our simplification would not adversely affect our results.

It must be noted that when we fit the data, we are in fact fitting a normative, average-reward micro SMDP model, including recursively solving for for policies and reward-rates, as discussed in Chapter 2 and 3.

6.4.1 Best fit parameters

For the subject F9 discussed above the best fit supra-linear to sub-linear utility function was a sigmoid ($\alpha = 0$, see Fig. 6.5 upper panel, green and red curves are superimposed). The extra weight parameter α did not improve the fit to the data. The sigmoid had a maximum $C_{L_{max}} = 48.2$, a shift $C_{L_{shift}} = 144.7$s and a slope $K_L = 0.24\text{s}^{-1}$. Note that the scaling of the sigmoid is relative to that of the reward intensities (which are assumed to be between 0 and 5). The shift at 144.7s indicates the duration at which the microscopic marginal utility was maximum (Fig.6.5 lower panel). This indicates that, according to the best fit sigmoidal utility to the given data, the subject preferred to engage in a long (optimally 144.7s) leisure bout, all at one go.

The best fit linear microscopic utility has a shallow slope ($K_L = 0.01\text{s}^{-1}$, see blue line in Fig. 6.5 lower panel), while a concave utility has a moderate maximal slope ($K_L = 0.41\text{s}^{-1}$, see cyan curve, lower panel). The shift of the sigmoidal function representing the mean Pavlovian component of leisure was larger when $C_L(\cdot)$ was linear ($Pav_{shift} = 3.70$s) or concave ($Pav_{shift} = 3.55$s), than when it was sigmoidal ($Pav_{shift} = 2.83$s). This implies the mean Pavlovian component of post-reward leisure is shorter for the best-fit linear and concave $C_L(\cdot)$ than sigmoidal ones.

6.4.2 Predicted distributions and ethograms

We illustrate how our model quantitatively fits the data using the price 'sweep' (Fig. 6.6). These are trials in which the reward intensity is fixed at a high level. Payoff is therefore determined by the price. We begin by comparing the sigmoid and the linear microscopic utilities.

Our model predicts the data at high payoffs well, irrespective of the utility of leisure function $C_L(\cdot)$. The data are roughly gamma distributed with very short means. Since at high payoffs, the opportunity cost of time dominates the microscopic utility of leisure, the predicted leisure duration distributions will be the same, irrespective of the choice of utility function. The revealing difference in fits occurs at long prices, when the payoff is low but trials (proportional to 25 × price) are sufficiently long to observe long, uncensored leisure bouts. Note, however that since we assume that subjects work continuously for the entire price duration, an

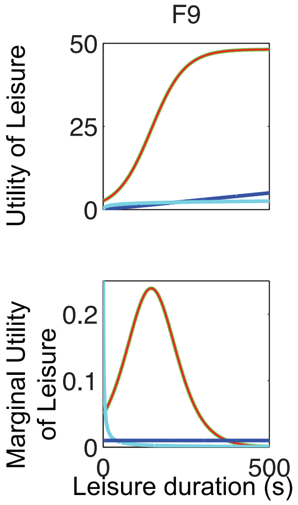


Figure 6.5: The microscopic utility of leisure for subject F9. Upper and lower panels show the best fit microscopic utility of leisure $C_L(\cdot)$ and the corresponding microscopic marginal utility of leisure, respectively. Blue, cyan, red and green curves denote linear, concave (logarithmic), sigmoid and initially supra-linear but eventually sub-linear $C_L(\cdot)$, respectively. The best fit supra-linear to sub-linear $C_L(\cdot)$ for this subject is, in fact, the sigmoid ($\alpha = 0$). Hence, the red and green curves are superimposed.

assumption that is not valid at very long prices, the work bouts predicted by our models cannot be expected to correspond to those in the experimental data. This will lead to more (and longer) work bouts being displayed on the ethograms generated from our models than observed experimentally. To put it another way, that we assume work bouts that exactly equal the price makes the leisure bouts in the ethograms generated from the best-fit sigmoid $C_L(\cdot)$ *appear* less stochastic than in the data. Thus, for the longer prices, only the leisure bouts should be compared between experiment and models. The full model in Chapter 3 predicts that pre-reward leisure bouts on low payoffs should have similar distributions to PRPs. Thus, for comparison purposes, at long prices, the PRPs predicted by our model should be considered to reflect pre-reward leisure bouts. Note, however that we in fact did not fit these pre-reward leisure bouts. We shall fit the entire dataset, including pre-reward leisure in Section 6.5.

At longer prices (see the highlighted upper panels, arranged in order of payoff: Price= 57.5s and 34.7s conditions) the sigmoidal $C_L(\cdot)$ fits the data better than the linear one. This is particularly evident in the lower negative log-likelihood for the sigmoid $C_L(\cdot)$ than the linear for the Price= 34.7s condition; the same is not reflected for the Price= 57.5s condition because, as discussed above, we include only PRPs, and exclude the majority of leisure bouts, which, for this condition, occur pre-reward.

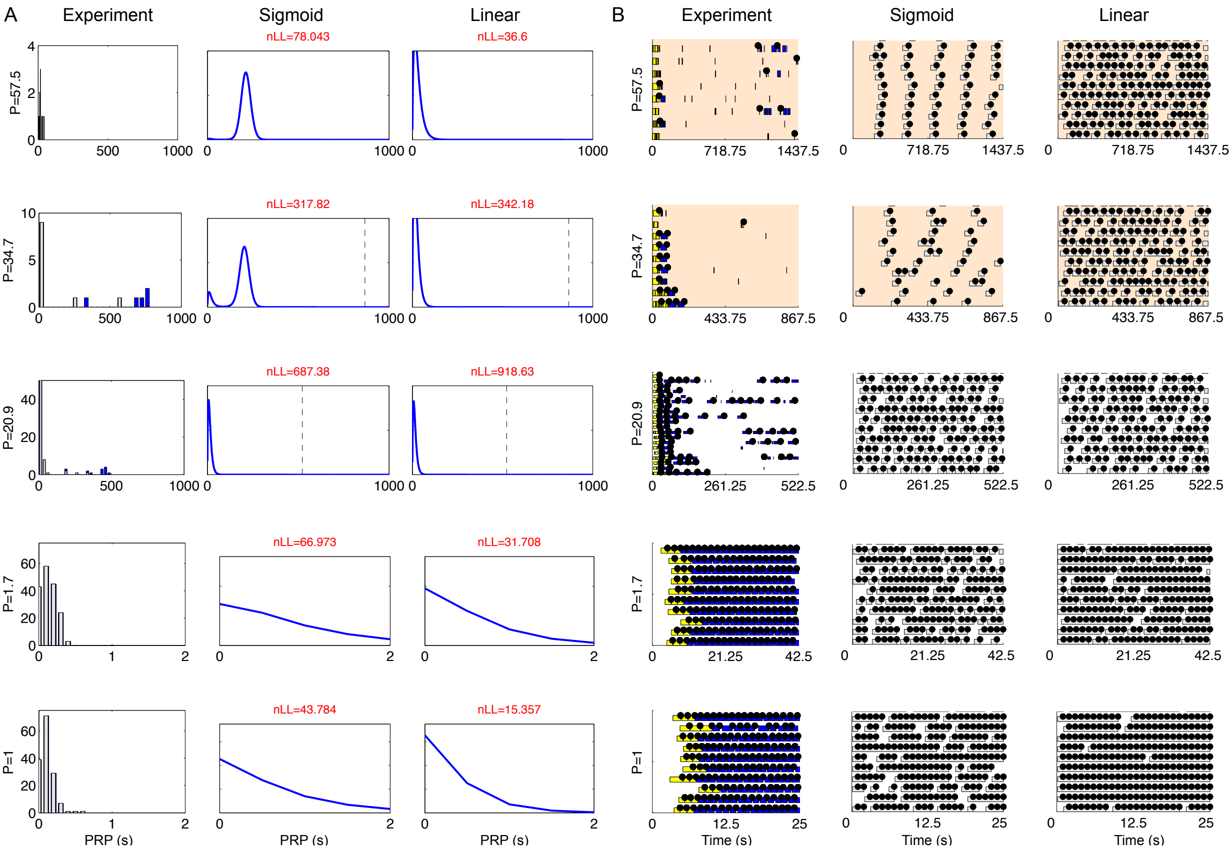


Figure 6.6: Model fits to PRPs on price 'sweep' for subject (F9). Price decreases (payoff increases) from top to bottom panels, as denoted in the labels. Only the lowest three and highest two payoffs are shown since the difference in model fits are clearest on these. A) PRP distributions. Left to right: Experiment, distribution predicted by best fit sigmoid and linear microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. The x-axis for the models is the same as that for the data. For very short prices, the 25 × Price line is not shown to allow for comparison with the data. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison B) Ethograms. Left to right: Experiment and ethograms predicted by best fit sigmoid and linear microscopic utilities of leisure. Note that since we assume subjects work continuously for the entire price duration, an assumption that is not valid at very long prices, the work bouts predicted by our models cannot be expected to correspond closely to those in the experimental data. Only the leisure bouts should be compared between model and experiment. According to our full model in Chapter 3, pre-reward leisure bouts on long prices should have similar distributions as PRPs, so the PRPs predicted by our model should be considered to reflect those.

For the sigmoidal microscopic utility of leisure, the microscopic marginal utility increases and then decreases after attaining a maximum–subjects prefer to engage in a long leisure bout, all at one go. The linear $C_L(\cdot)$ cannot accommodate this, and when data are fit using this function, the best fit linear $C_L(\cdot)$ predicts that subjects will engage in much shorter leisure bouts than observed in the experiment at these long prices.

The poorer fit of linear $C_L(\cdot)$ owes to the fact that if leisure durations across RI, P conditions are inherently generated from a bimodal distribution, comprising a mixture of gamma distributions, then attempting to fit such data with a unimodal exponential or gamma distribution will be impaired. A unimodal distribution would fit either the short or long modes, but not both. Since across conditions, the majority of the uncensored leisure bouts observed are shorter, a unimodal distribution would be biased towards shorter modes than longer ones. Thus, the predicted leisure bouts at long prices would be shorter than those observed experimentally. We shall return to the question of how much the precision our data fits could be impeded by censoring in Section 6.5.1.

It is clearer to see that PRPs are inherently generated from a mixture of two gamma distributions when we analyse the data for RI 'sweeps'. These collect the conditions for a fixed price, with the reward intensity increasing between conditions. This is clearest for a RI 'sweep' with a longer Price= 8s (Fig. 6.7; see also Fig. 6.16 for comparing with the 'RI 'sweep' with the Price= 4s). Despite the fact that trials end at 200s, causing a large proportion of PRPs at low RIs to be censored, we can see that the PRPs are bimodally distributed, with the mixture weight on the shorter mode increasing with payoff. By analysing the top three panels of Figs. 6.7, it is clear why, despite the proportion of censoring, the sigmoidal $C_L(\cdot)$ fits the bimodal data much better than the linear or concave $C_L(\cdot)$. The latter attempt to fit the long PRPs by having a long-tail. This is further verified by analysing the radial 'sweep' which compares across conditions in which both RI and price differ (Fig. 6.7). These afford us the opportunity to analyse and fit the PRP distributions at low RIs, except that prices are also long enough to somewhat reduce censoring.

The larger shift in the Pavlovian component for the linear $C_L(\cdot)$ mentioned above, implying shorter mean Pavlovian components of post-reward leisure, affords greater flexibility in fitting longer leisure distributions. The shorter Pavlovian components attempt to capture the shorter mode of the bimodally distributed leisure durations, while the shallow slope of the linear $C_L(\cdot)$ attempts to fit the longer leisure bouts with a long-tailed exponential. However, as noted above, provided there is a significant proportion of long, uncensored leisure bouts gen-

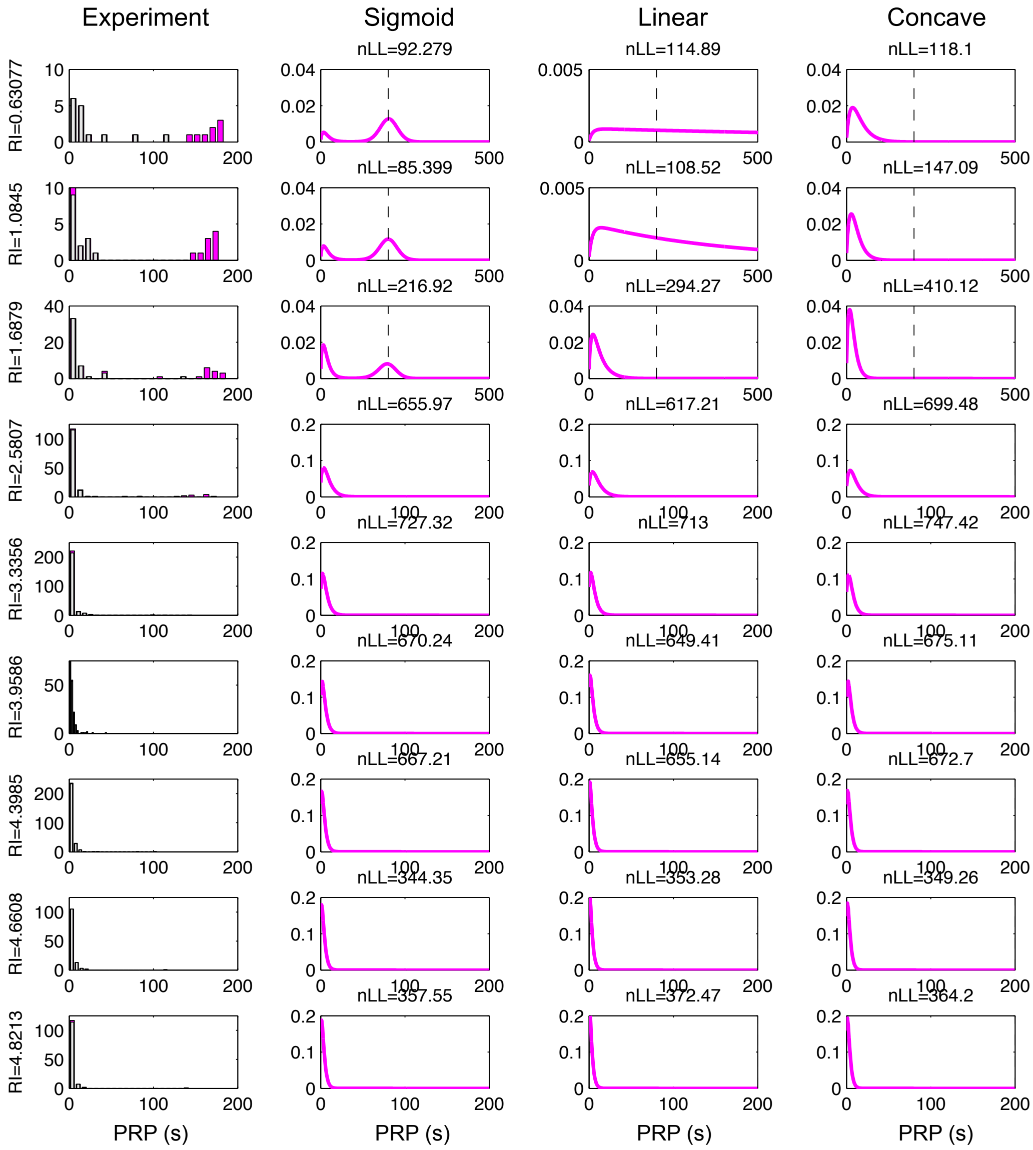


Figure 6.7: Model fits to PRPs on a RI 'sweep', Price= 8s for subject (F9). RI increases from top to bottom panels, as denoted in the labels on the left. Left to right: Experiment, distribution predicted by best fit sigmoid, linear, and concave microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison. Note the different x-axis scales for experimental data.

erated from the long mode of a bimodal distribution, a long-tailed exponential will, literally and figuratively, fall short of fitting them.

6.4.3 Model comparisons

That the sigmoid was the best fit, amongst the four models we considered, to these inherently bimodally distributed data was confirmed by analysing the BIC scores of the model with each $C_L(\cdot)$ (Fig.6.8 lower left panel). When integrated across all 81 RI, P conditions for this subject, the PRPs were best fit (lowest BIC score) by the sigmoidal $C_L(\cdot)$ ($BIC = 23,709$, see gold bar in Fig.6.8 lower left panel). The concave $C_L(\cdot)$ ($BIC = 24,212$ fits better than the linear one ($BIC = 28,161$). The best fit supra-linear but eventually sub-linear $C_L(\cdot)$ is the sigmoidal but its extra weight parameter (α) leads to a larger BIC score ($BIC = 23,714$).

6.4.4 Between subject model comparisons

When we compare across subjects (Fig 6.8), we notice that quantitatively the pure sigmoidal is the best fit for subject F9 only. For subjects F3 and F12, despite the best fit supra-linear-sub-linear functions having a maximal microscopic marginal utility at long durations, the PRP data are slightly better fit by concave utility functions with moderate to steep slopes ($K_L = 0.52$ for F3 and 0.25 for F12). For the other three subjects (F16, F17 and F18), the supra-linear-sub-linear function is the best fit, with the weight on the linear component exceeding that on the sigmoid. For subjects F17 and F18 especially, the extra parameter enables a very shallow sloped sigmoid, which implies a broad, slowly increasing and slowly decreasing marginal utility curve. The PRPs for these subjects have a broader, more unimodal distribution than for the tighter, bimodally distributed PRPs of F9. Consequently, they are better fit by the weighted combination of a linear and sigmoid. However, the linear $C_L(\cdot)$ is the worst fit for every subject, implying that subjects are not indifferent to the division of leisure durations.

6.5 Fitting all data using full micro-SMDP

Fitting PRPs only requires some amount of data to be discarded. For instance, the long, uncensored instrumental leisure bouts observed at long prices are revealing aspects of the data which need to be accounted for. Similarly, short

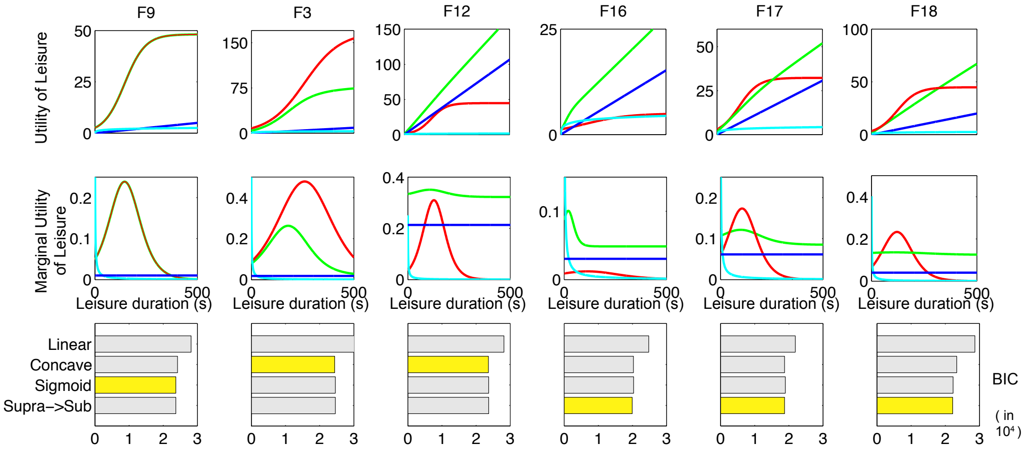


Figure 6.8: The microscopic utility of leisure between subjects. Each column shows a different subject. Upper and middle panels show the best fit microscopic utility of leisure $C_L(\cdot)$ and the corresponding microscopic marginal utility of leisure, respectively. Blue, cyan, red and green curves denote linear, concave, sigmoid and initially supra-linear but eventually sub-linear $C_L(\cdot)$, respectively. Lower panels show BIC scores (expressed in 10^4) for each best fit $C_L(\cdot)$; lower BIC scores reflect better fits. The lowest BIC score, corresponding to the model with the $C_L(\cdot)$ that fits the PRPs the best is highlighted in gold.

instrumental leisure bouts which occur pre-reward could favour a different microscopic utility of leisure function. We now fit the entire dataset, comprising all work and leisure choices and their durations for each individual subject.

We use the same full micro-SMDP that we introduced in Chapter 3, except for a minor alteration that restricts transitions to those between work and leisure, precluding a long sequence of e.g. 'work-work-work ...' choices. More crucially, in the model considered in Chapter 3, at low payoffs, if behaviour is too deterministic subjects shall repeatedly choose to engage in long leisure bouts ('leisure-leisure-leisure ...' choices) and remain in the same pre-reward state. The behavioural cycle from pre- to post-reward can then fail to complete (leading to non-ergodicity). Further, while a set of consecutive leisure bouts is sensible when the subject is indifferent to the duration of leisure, a set of consecutive long leisure bouts is less meaningful when choosing according to e.g. a sigmoidal utility of leisure. We prevent the above problems by restricting transitions to those between work and leisure.

We choose a finer discretisation of 0.5s time-steps for τ_L and a discretisation of 1s time-steps for chosen work durations τ_W. Since subjects usually pre-commit to working continuously for the entire price duration, this discretisation should not adversely affect our results. As in the model in Chapter 3, we employed an exponential prior for leisure $\mu_L(\tau_L) = \lambda \exp(-\lambda \tau_L)$ with mean $1/\lambda = 10P$; the exponential prior for work durations $\mu_W(\tau_W)$ did not matter as long its mean was not so short that it made attaining of the price much unlikely. The reward rate in our new full micro-SMDP, which restricts transitions to those between work and leisure, was similar to that for the full micro-SMDP model in Chapter 3, suggesting that the results of fitting the data with either would be similar.

We fit the data by maximising the log of the likelihood given in Eq.(6.3) of observing the entire dataset of microscopic choices.

6.5.1 Censoring impedes precision of fits

A problem that we face in fitting the experimental data we possess is that several of the long durations are censored by the end of trials (which are 25 × price). Another possibility is that we have too little data. These may impair our ability to estimate precisely the underlying parameters of our models. To investigate whether it is the amount of data or censoring which may lead to lower precision in our estimates, we generated artificial data from a sigmoidal microscopic utility (black curve in Fig. 6.9, upper and lower panels. We used 1000 trials per RI, P condition (compared to the 9 to 20 trials in the experiment), providing us with plenty of data to fit. As in the experiment, these trials were terminated at 25 × price. When we fit this dataset, irrespective of the starting point of negative log-likelihood minimisation, the recovered best fit $C_L(\cdot)$ has its shift $C_{L_{shift}}$ in the appropriate place, but its maximum $C_{L_{max}}$ is underestimated and slope K_L is overestimated (see coloured curves in Fig. 6.9, upper panel). That the shift is recovered accurately implies that the model accurately predicts the mode of the underlying distribution, although the true mixture weights of the generated bimodal distribution may be less accurately estimated.

To investigate whether removing censoring could alleviate this underestimation, we generated 1000 trials per RI, P condition with trial durations set to a long, 5000s (or 500s for prices that were less than 1s). This led to most data being observed and not censored (Fig. 6.10, left panels). By fitting all this data, we were able to recover the true generating parameters (Fig.6.9, lower panel). This is clearly reflected in the near identical distributions predicted by the true and recovered parameters (Fig.6.10 middle and right panels, respectively).

We have thus established censoring to be a key issue, which may impede our ability to fit the data precisely. In general, longer durations are more likely to be censored than short ones. If we generate data as in the experiment, even if we generate large quantities of it, we are still left with the problem that longer durations, which come from the right mode of the bimodal get censored. If those critical durations were not censored, then we would be extremely confident about the maximum ($C_{L_{max}}$) and slope (K_L) parameters. However, in the presence of censoring, we shall still be able to accurately estimate the long mode of the bimodal distribution.

Since a larger $C_{L_{max}}$ leads to a larger reward rate (hence a greater opportunity cost of time), predicted distributions have larger mixture weights on the short-tailed exponential than the long-mode gamma component, making it costlier in terms of negative log-likelihood to fit long leisure bouts. To further emphasise this, suppose we only fit data from medium and high payoffs with a sigmoid $C_L(\cdot)$. These could be better fit by a linear $C_L(\cdot)$. But to fit that range with a sigmoid $C_L(\cdot)$, the $C_{L_{max}}$ would have to be low enough such that reward rate is small and the policies are roughly exponential with medium-long means. Thus, when we have limited data or the long mode is systematically censored, the relative costliness in negative log-likelihood of having larger reward rates biases $C_{L_{max}}$to be underestimated. The slope K_L is overestimated to compensate this underestimation.

6.5.2 Predicted distributions and ethograms for experimental data

Compared to when we fit only PRPs, when we fit all the experimental data, including the long, uncensored pre-reward leisure bouts, we can see even more clearly that our model with a sigmoidal $C_L(\cdot)$ quantitatively captures the temporal topography of choice better than a linear $C_L(\cdot)$ (Fig. 6.11). When we analyse the price 'sweep', at the longer prices (see Price= 57.5s and 34.7s, 20.9s conditions) our sigmoidal $C_L(\cdot)$ predicts that the subject shall engage in a long leisure bout before resuming work. These predicted long instrumental leisure bouts occur pre-reward for the Prices= 57.5s and 34.7s, as in the data. The set of extra work bouts seen in the ethograms predicted by our model, compared to that in the experimental trials is due to the fact that we start our simulated trials from the post-reward state, with a long PRP, followed by the first work bout. For our simulated ethograms to appear exactly as that in the experiment, we should shift the start our simulated trials to coincide with the last work bout when the subject does not know the reward and the price. We have not done so because we did not model the sampling period.

That our model with a sigmoid $C_L(\cdot)$ fits the data better is even more clearly seen when we analyse the low payoffs of the RI 'sweep' for a price of 8s (Fig. 6.12).

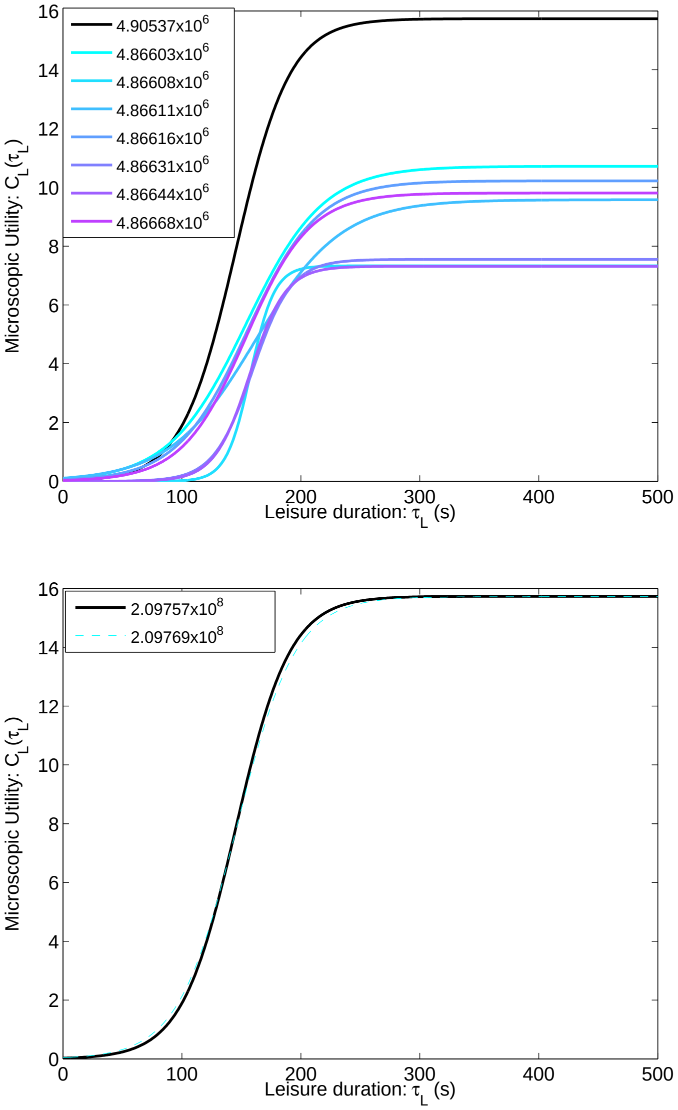


Figure 6.9: Censoring impedes precision of fits. 1000 trials per RI, P condition of data were generated the sigmoid $C_L(\cdot)$ shown by the black curve. Upper panel: For trial durations of 25 × price as in the experiment a large proportion of leisure bouts are censored by the end of the trial. These data were then fit by our model, coloured curves show best fit $C_L^*(\cdot)$ obtained by starting the negative log-likelihood minimisation procedure from different points, red to blue cool colours show decreasing negative log-likelihood (see legend for the negative log-likelihood; note how small the differences are). The recovered ${C_L}^*(\cdot)$ has the correct shift $C_{L_{shift}}$, but underestimates the maximum $C_{L_{max}}$ (and overestimates the slope K_L). Lower panel: When trial durations were extended to 5000s (or 500s for prices less than 1s), the majority of leisure bouts were observed and not censored. The best fit best fit $C_L^*(\cdot)$ (coloured dashed curve) recovered the true generating parameters, irrespective of the starting point of negative log-likelihood miminisation.

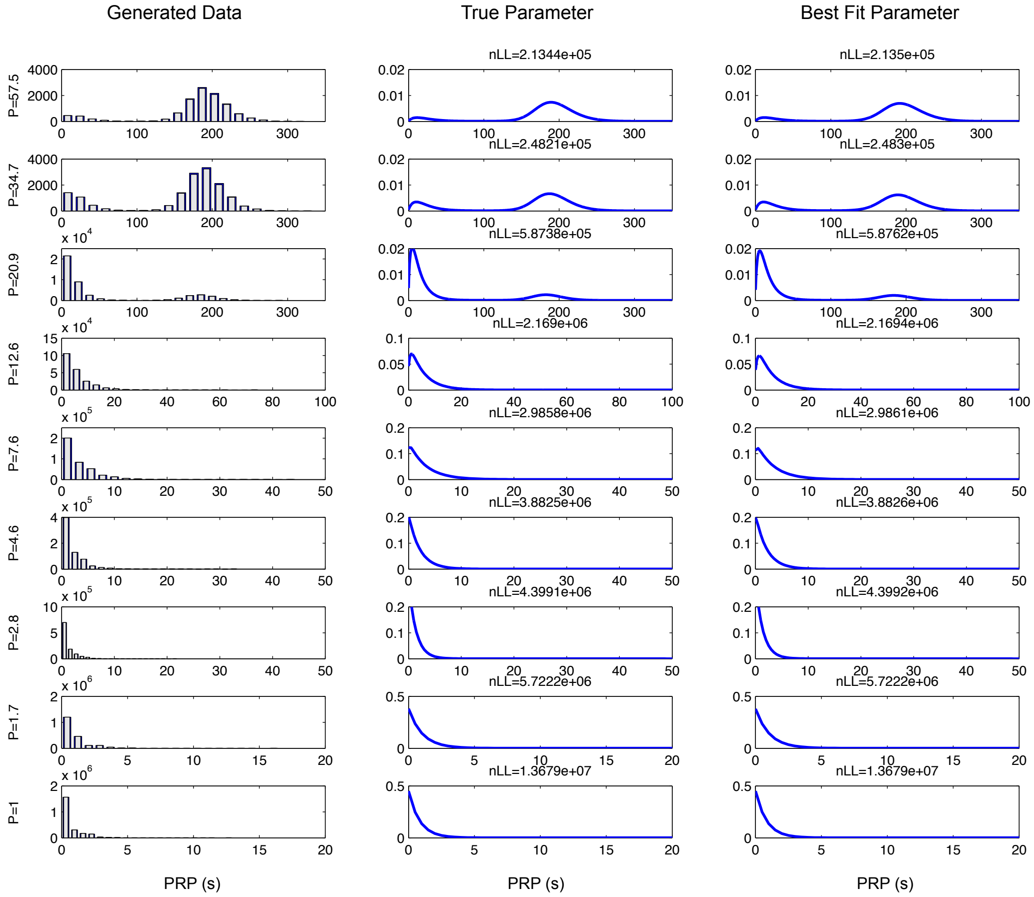


Figure 6.10: Improved fits for uncensored data. Generated price 'sweep'. Price decreases (payoff increases) from top to bottom panels, as denoted in the labels on the left. 1000 trials per RI, P condition were generated with trial duration of 5000s. Left: generated data PRP distribution, middle: true PRP distribution, right: PRP distribution from best fit sigmoid. For left panel, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. Note the small proportion of censored data. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison.

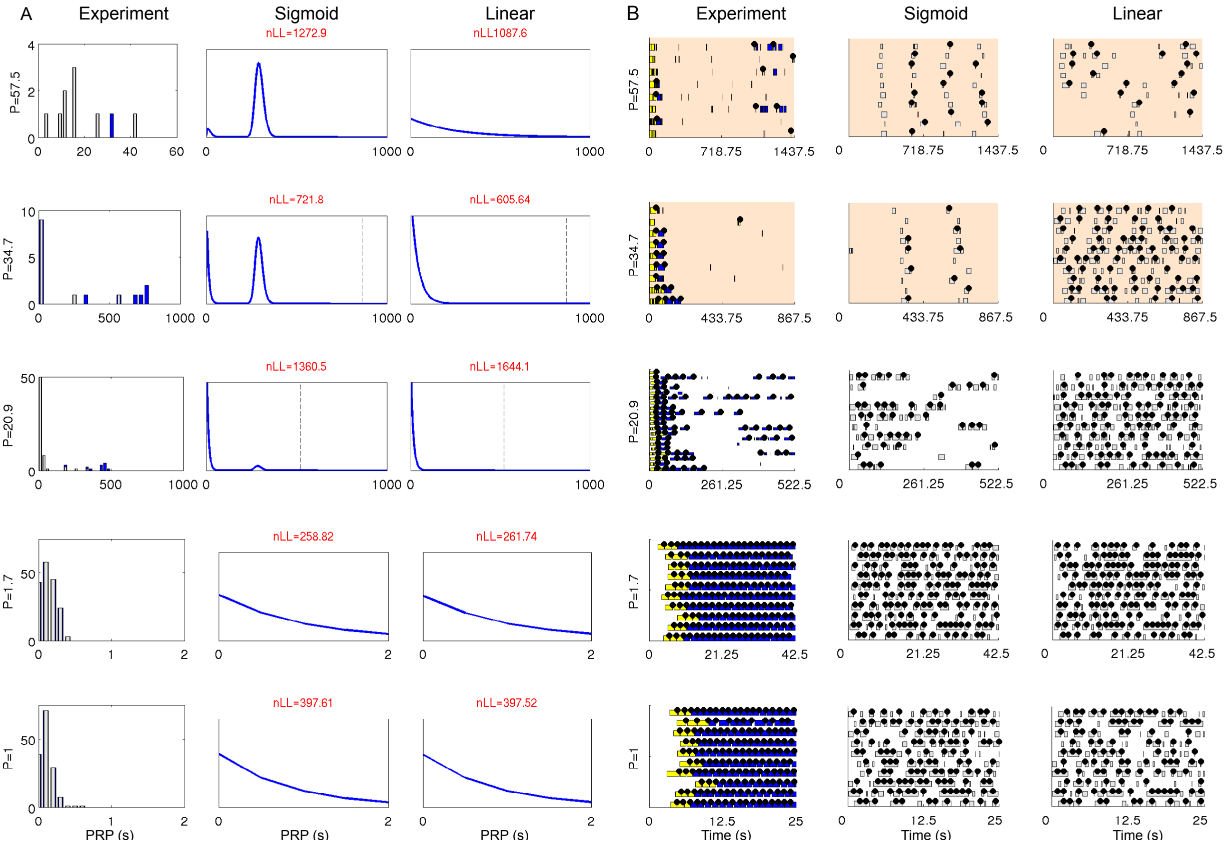


Figure 6.11: Model fits to all data, shown for price 'sweep' for subject (F9). Price decreases (payoff increases) from top to bottom panels, as denoted in the labels. Only the lowest three and highest two payoffs are shown since the difference in model fits are clearest on these. A) PRP distributions. Left to right: Experiment, distribution predicted by best fit sigmoid and linear microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. The x-axis for the models is the same as that for the data. For very short prices, the 25 × Price line is not shown to allow for comparison with the data. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison B) Ethograms. Left to right: Experiment and ethograms predicted by best fit sigmoid and linear microscopic utilities of leisure.

As discussed above, the linear $C_L(\cdot)$ is unable to fit such bimodally distributed data.

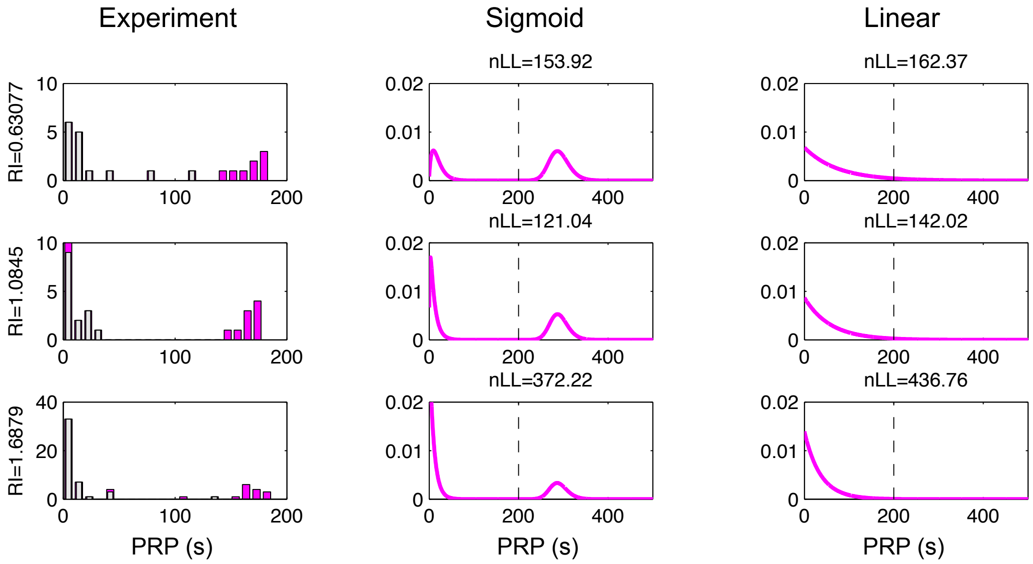


Figure 6.12: Model fits to all data, shown on a RI 'sweep', Price= 8s for subject (F9). RI increases from top to bottom panels, as denoted in the labels on the left. Only the lowest three payoffs are shown since the difference in model fits are clearest on these. Left to right: Experiment, distribution predicted by best fit sigmoid, linear, and concave microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. Note the different x-axis scales for experimental data.

6.6 PRPs on medium prices may not be independent and identically distributed

In the above sections, we attempted to determine quantitatively subjects' microscopic utilities of leisure, reflecting their innate preference for leisure independent of fatigue, satiation, and all other rewards and costs. We assumed the data were independent and identically distributed (i.i.d), as predicted by our model. However, if fatigue or satiation play a role, then these data shall no longer be i.i.d. Specifically, if working leads to physical fatigue, then we should expect runs of work bouts interspersed by short PRPs, followed by a long PRP. We therefore tested whether PRPs in our data were non-i.i.d. We performed two non-parametric frequentist statistical tests. (i) the Ljung-Box Q test for non-zero autocorrelations was performed on each trial. This is a portmanteau test of overall autocorrelation/randomness. We report Bonferronni corrected results. (ii) Wald-Wolfowitz Runs Test was used to test for runs in each PRP time-series, with binary values: greater than or less than median of uncensored PRPs. Since this is a weak test when there are few data points per trial, as in our case, we bootstrapped PRPs and then re-ran the test.

Fig. 6.13 shows sequences of PRPs, on trials belonging to a RI 'sweep' with a medium price of 8s. Each column corresponds to a different RI condition, with each row representing a different trial. The axes on these are coloured light pink if the post-bootstrap runs test came out significantly positive (there were runs in the data) at an alpha level of 0.05, light yellow when the LjungBox Autocorrelation test was successful (there was significant autocorrelation in the data) at the appropriately Bonferronni corrected alpha level, and gold when both tests were successful. On trials on which PRPs were significantly non-i.i.d. according to the tests (e.g. the fourth panel from the top for the $RI = 4.6608$ condition), subjects engaged in a set of longer PRPs after a string of short PRPs. Importantly, these long PRPs did not happen at random parts of the trial, thus making the case against them being i.i.d draws.

To summarise these, we split the sequence of all PRPs (without pre-processing them to exclude the sampling period) on each trial into 3 parts and computed the mean PRP for that part of that trial (Fig. 6.14). Further summarising this summary analysis, we collapsed the first third, second third, last third PRPs (red, green and blue histograms in Fig. 6.15A) across all trials (per RI condition). These show that the first third of a trial has short PRPs. These partly reflect the sampling period when the subject does not know the reward and price. We may henceforth exclude them from consideration. As the trial goes on, PRPs become longer, PRPs on the final third of the trial are longer than those on the second.

Taken together, these tests suggest that fatigue may play a role on these trials when the price is medium (8s). However, fatigue alone does not explain these results; any effect of fatigue is in addition to that of payoff, since PRPs are shorter for higher payoffs. Furthermore, these effects are not observed on shorter (e.g. 1s) prices (PRPs on different parts of a trial are similarly distributed Fig.6.15B), suggesting that physical fatigue may play a role only at longer prices. However, overall, these accounted for around 10% of the trials, even at long prices, suggesting that fatigue is less likely to be a major underlying explanation of these data.

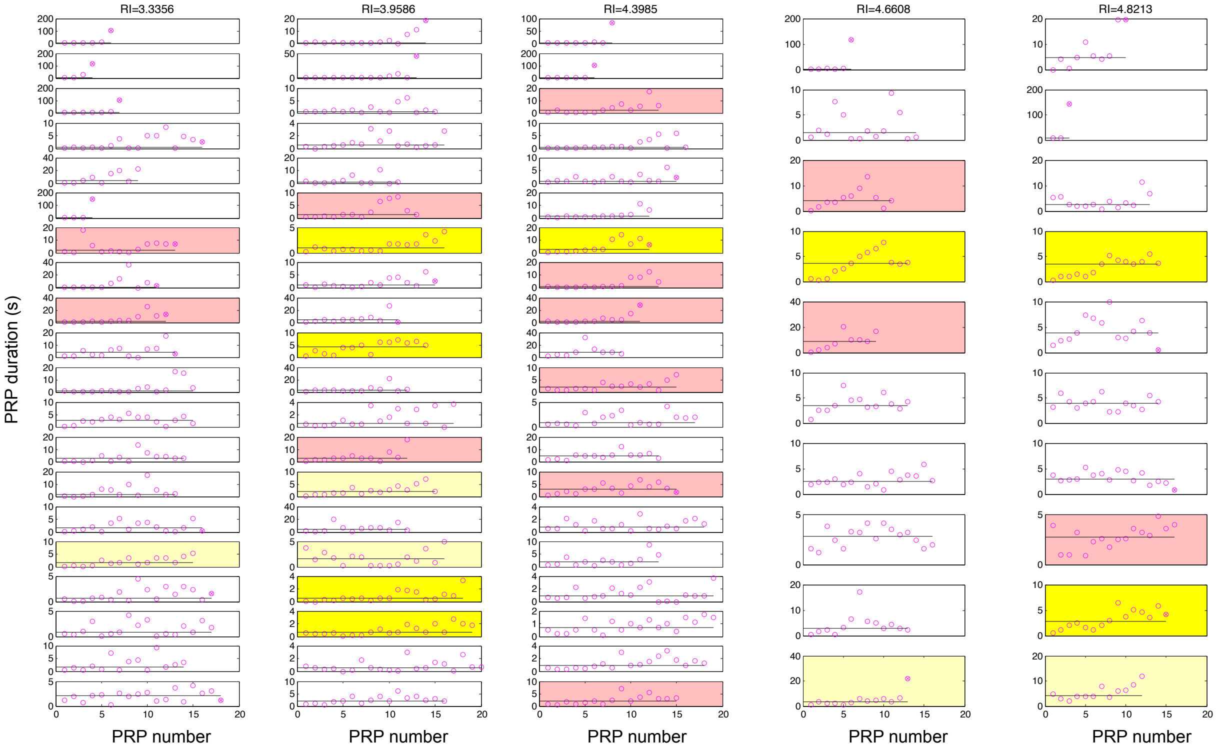


Figure 6.13: PRPs on medium prices may not be i.i.d Sequence of PRPs shown on a RI 'sweep', Price= 8s for subject (F9). y-axis: PRP duration, x-axis: PRP number in the trial; only uncensored PRPs are shown. Each column corresponds to a different *RI* condition, with each row representing a different trial. *RI* increases from left to right, with only the highest 5 *RI* shown here. Black line shows the median of uncensored PRPs on that trial. Data are preprocessed to exclude the initial sampling period when the subject does not know the reward intensity and price. Axes are coloured light pink if the post-bootstrap runs test was significantly positive at alpha level of 0.05 (there were runs in on that trial), light yellow when the LjungBox Autocorrelation test was successful (there was significant autocorrelation in the PRPs on that trial) at the appropriately Bonferronni corrected alpha level, and gold when both tests were successful.

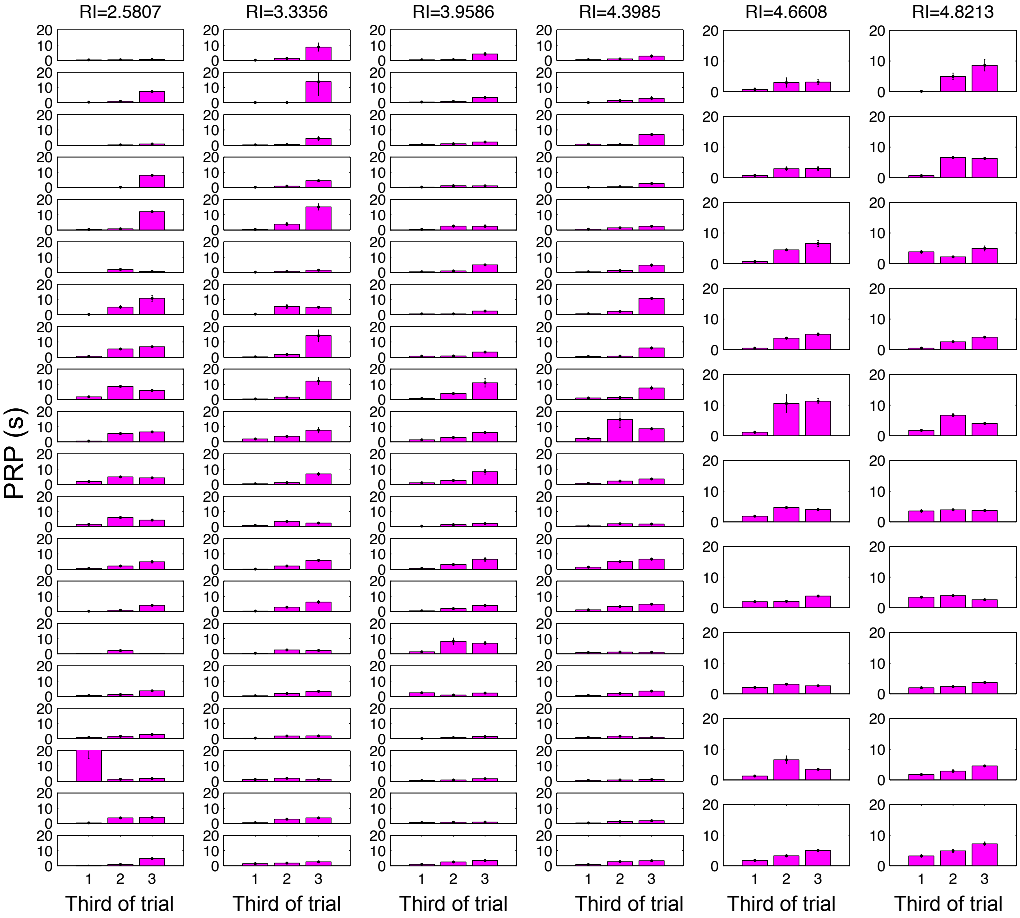


Figure 6.14: PRPs may increase through a trial for medium prices. Mean PRPs for the first, second and final thirds of a trial shown on a RI 'sweep', Price= 8s for subject F9. y-axis: PRP duration. Only uncensored PRPs are accounted for. Each column corresponds to a different *RI* condition, with each row representing a different trial. *RI* increases from left to right, with only the highest 6 *RI* shown here. Error bars show standard deviations, normalised by the number of data points for each third. Data are not preprocessed to exclude the initial sampling period when the subject does not know the reward intensity and price.

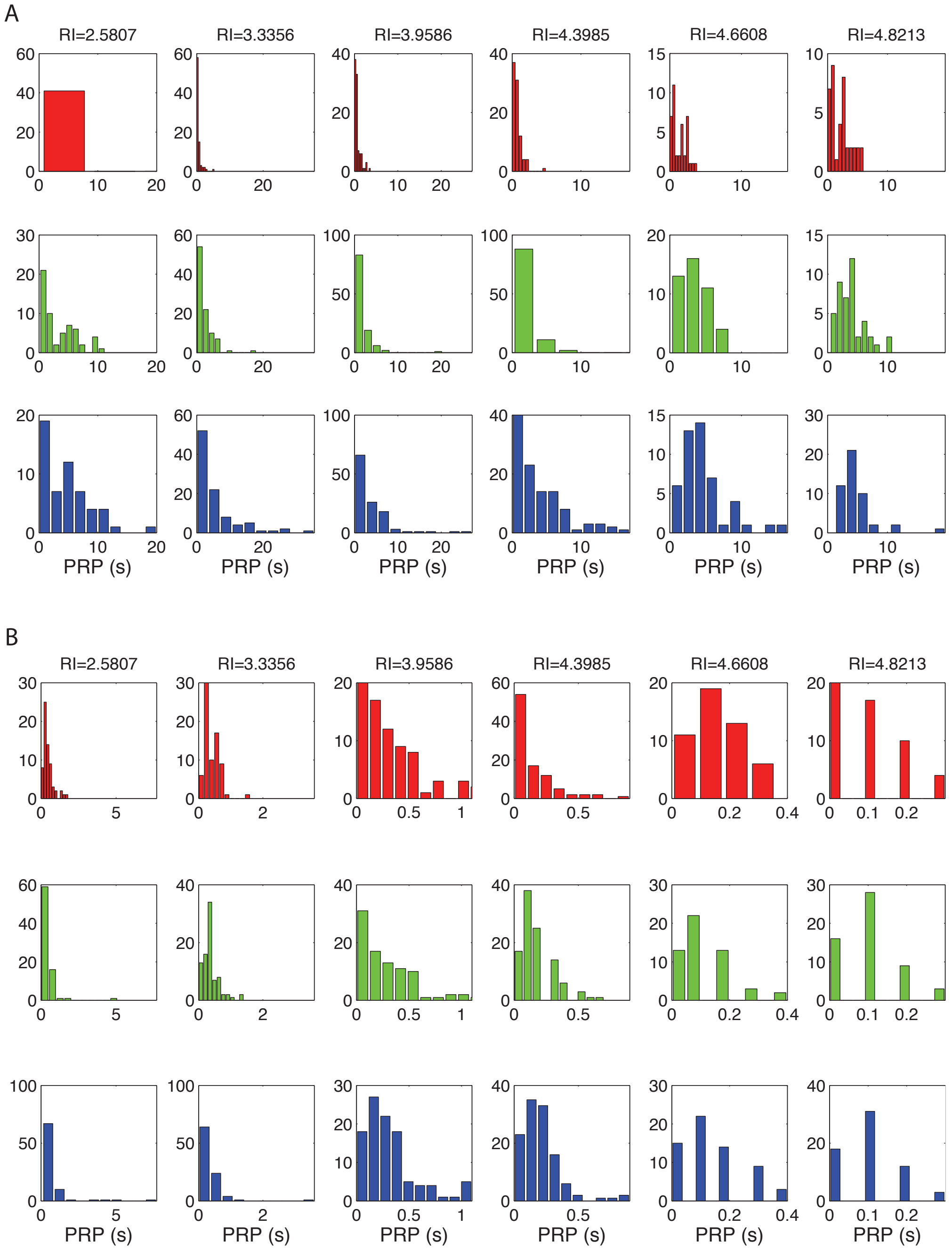


Figure 6.15: PRPs may increase through a trial for medium prices but not short prices. PRP duration histograms for the first (red), second (green) and final (blue) thirds of a trial are collapsed across trials and shown on a RI 'sweep' for subject F9. x-axis: PRP duration. Only uncensored PRPs are accounted for. Each column corresponds to a different *RI* condition. *RI* increases from left to right, with only the highest 6 *RI* shown here. A) Medium price of 8s. B) Short price of 1s. Data are not preprocessed to exclude the initial sampling period when the subject does not know the reward intensity and price.

6.7 Summary

The microscopic utility of leisure $C_L(\tau_L)$ quantifies a subject's innate preference for a duration of leisure, independent of all other rewards and costs, fatigue or satiation. Here we empirically determined the utility of leisure in rat subjects by quantitatively fitting their microstructure of choices. We first noted that at long prices, time allocation indeed increases rather than decreasing with price. Our analysis of the temporal structure of behaviour revealed that at long prices, subjects engaged in a long leisure bout before resuming work. By integrating and averaging this across time, we could explain the experimental observation that time allocation may increase rather than decrease with price at long prices.

The resumption to work after a long leisure bout is predicted by our normative, microscopic model with an initially supra-linear but eventually sub-linear $C_L(\cdot)$, e.g. a sigmoid. To test whether subject's preferences for leisure could be quantified by such a utility we fit one such utility function $C_L(\cdot)$ across the entire dataset of experimenter determined RI and price conditions. As discussed in Chapters 3 and 4, the difference between the innate microscopic utility of of leisure and the opportunity cost of time determines the leisure duration distribution. Since the opportunity cost of time is largely determined by the payoff (i.e., RI/P), for a given $C_L(\cdot)$ the subject's leisure duration distribution changes as the payoff is manipulated by the experimenter.

While there is an infinite space of possible microscopic utility of leisure functions to test, we chose those that had few parameters, and made starkly different qualitative predictions about the preference for durations of leisure. We therefore chose the canonical $C_L(\cdot)$ from Chapters 3 and 4: namely the linear (indifferent in preference), concave (which prefers many short leisure bouts to one long bout) and sigmoid (which prefers one long leisure bout to many short ones). We further considered a weighted combination of a linear and a sigmoid to represent the generic class of initially supra-linear but eventually sub-linear $C_L(\cdot)$. This functional form is also a more mathematically continuous version of the quasi-concave functions popular in economics.

An alternative approach to the one we pursued would be to descriptively fit leisure durations first with standard parametric distributions Breton (2013) and then invert the generative process to infer utility functions. We had originally performed such preliminary analyses (not shown here), which suggested that PRPs were most likely to be generated from bimodal gamma distributions. However, in general, inverting the process to infer utility functions could lead to arbitrarily complex utility functions with many parameters. Interpreting what these would predict about preferences, from a normative perspective, would be cumbersome. Instead, we approach the data from an approximately normative perspective, fitting micro-SMDP models, including self-consistent policies and reward-rates. Our full micro-SMDP model has the further advantage of also being able to provide a normative perspective on the duration of work bouts and not just leisure bouts. It should be noted, however, that according to our micro SMDP models, exponential and gamma distributed leisure durations are consistent with subjects having linear and logarithmic $C_L(\cdot)$, under a softmax policy over leisure durations. In these cases, going from descriptive analyses to normative perspectives or vice versa yields the same result.

Since, for short prices, subjects pre-commit to working continuously for the entire price duration, leisure bouts are mostly PRPs. As in Chapters 4 and 5 we first assumed that subjects work continuously for the entire price duration and use the simplified, 2-state micro-SMDP to model PRPs only. We found that for a subject of interest (F9), the sigmoidal $C_L(\cdot)$ fit the data much better than a linear $C_L(\cdot)$. Specifically, the best fit linear $C_L(\cdot)$ predicted PRPs that were much shorter than those experimentally observed. A linear or concave $C_L(\cdot)$ predicts unimodal distributions. However, if the PRPs are generated from a bimodal distribution consisting of a mixture of gamma distributions with short and long modes, then attempting to fit such bimodal distributions with a unimodal will be extremely disadvantaged. Since a larger proportion of PRPs are observed to be short, while long PRPs are censored, a unimodal fit would be biased towards the shorter mode. That is, predicted leisure durations will be much shorter. Our model with a sigmoidal $C_L(\cdot)$ predicts and accurately captures such bimodally distributed data, with the mixture weight on the shorter mode increasing as payoff increases. We can conclude that this subject innately prefers long leisure bouts, all at one go, rather than many short ones for the same total duration. For this subject we found that the best fit initially-supra-linear but eventually sub-linear $C_L(\cdot)$ was intact a sigmoid, with its extra parameter not affording a better fit.

For two other subjects, the concave was the best fit, although only slightly better than the sigmoidal. This could either reflect that these subjects innately prefer many short leisure bouts over one long bout for the same total duration, or be due to the fact that a large proportion of PRPs are censored. In future work, we shall analyse the data further to understand which is more likely. For three others the supra-linear-sub-linear $C_L(\cdot)$ function was the best fit, with the weight on the linear component exceeding that on the sigmoid. The best fit $C_L(\cdot)$ functions for these subjects had shallow slopes, implying a slowly increasing and decreasing marginal utility. This may be due to the fact the PRPs for these subjects had a broader distribution with larger variance. However, the linear $C_L(\cdot)$ was the worst fit for every subject. We therefore conclude that rat subjects are not indifferent in preference to the duration of leisure.

The simplification that subjects work continuously for the entire price duration and leisure bouts are PRPs is invalid at long prices. For very long prices, subjects engage in long leisure bouts pre-reward before resuming work. Ignoring such data could have slightly impaired the quality of our model fits. We therefore fit the full micro-SMDP model developed in Chapter 3, with a very minor alteration, to all the microscopic work and leisure choices in the dataset, for each individual rat. For subject F9, we were able to better capture the temporal topography of choice; our model with a sigmoidal $C_L(\cdot)$ predicted these long, pre-reward instrumental leisure bouts after which a subject resumed working. As before, the linear $C_L(\cdot)$ fell short of accounting for bimodally distributed leisure durations. In future work, we shall fit all the data with this full micro-SMDP model, for concave and other $C_L(\cdot)$ as well, and for other subjects. It is possible that accounting for the crucial pre-reward data could lead to different best-fit utility functions for the other subjects, than those reported here while fitting only PRPs.

The large proportion of long, censored, leisure bouts is a key issue that may impede our ability to determine precisely a subject's utility of leisure. Despite having a wealth of data, since longer durations are more likely to be censored than short ones, we are able to more precisely fit the shorter mode than the longer mode, biasing our inferred utility of leisure functions. Specifically, for a sigmoidal utility, we may still be able to determine where the mode lies, but be less confident about the maximum utility or the slope of the function, which may be underestimated and overestimated, respectively. Collecting data with longer trial durations, e.g. 50 × the price, could provide more observed and less censored long leisure bouts; however, such long trials would mean longer running experiments and be less practicable from the perspective of the experimentalist.

In attempting to determine a subject's innate preference for leisure, irrespective of all else, we assumed that PRPs were independent and identically distributed. This may not be the case in presence of fatigue or satiation, which introduce runs of work bouts interspersed by short PRPs, followed by a long PRP that reduces fatigue/satiation. Apriori, physical fatigue is less likely to play an important role in generating this data due to the meticulous experimental design. Subjects were well rested on trailing trials, limiting physical fatigue. Satiation is also unlikely to play a major role since non-satiating BSR rewards were used. Further, subjects were observed to work continuously throughout a trial, with very little leisure, even for the very highest reward intensities, contrary to what would be the case if satiation was present. However, to test whether fatigue or satiation empirically still played a role, we conducted statistical tests of independence. Statistical tests of the data showed that for some trials on medium prices, PRPs were not independent and identically distributed as we had assumed. Specifically, there were runs in the PRP sequences on some trials. Whether the long leisure durations on these trials manifest a sigmoidal microscopic utility of leisure or are caused by fatigue is an interesting question. It must be noted that fatigue alone does not explain these results; any effect of fatigue is in addition to that of payoff since PRPs are shorter for higher payoffs. Since even for longer prices, only around 10% of the trials were found to have non-independent PRPs, it is empirically unlikely that fatigue plays a major role in these data. We must note, however, that statistical tests may only reject the null hypothesis of independence; when a test of runs or autocorrelation comes out negative, it does not automatically imply that the PRPs are independent, just that it cannot be concluded that they are non-independent. In the next phase of our programme, we shall quantitatively fit our normative, microscopic models of fatigue and satiation to these data in order to infer which of the set of models, with or without fatigue best accounts for the data we observe.

6.8 Additional data figures

Here we include additional data figures. We show relevant figures with all the RI,P conditions and display fewer conditions in the main-text.

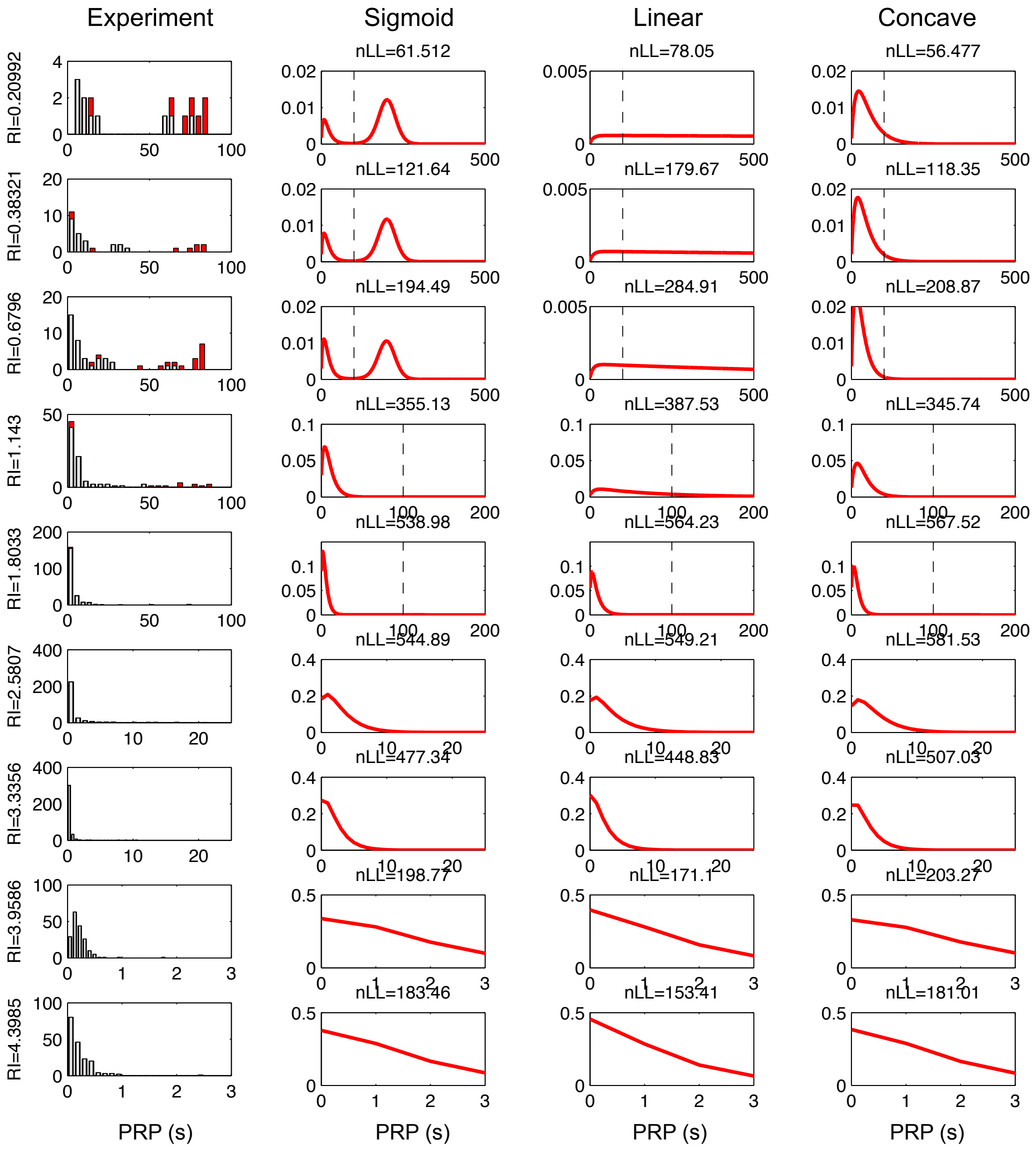


Figure 6.16: Model fits to PRPs on a RI 'sweep', Price= 4s for subject (F9). RI increases from top to bottom panels, as denoted in the labels on the left. Left to right: Experiment, distribution predicted by best fit sigmoid, linear, and concave microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison. Note the different x-axis scales for experimental data.

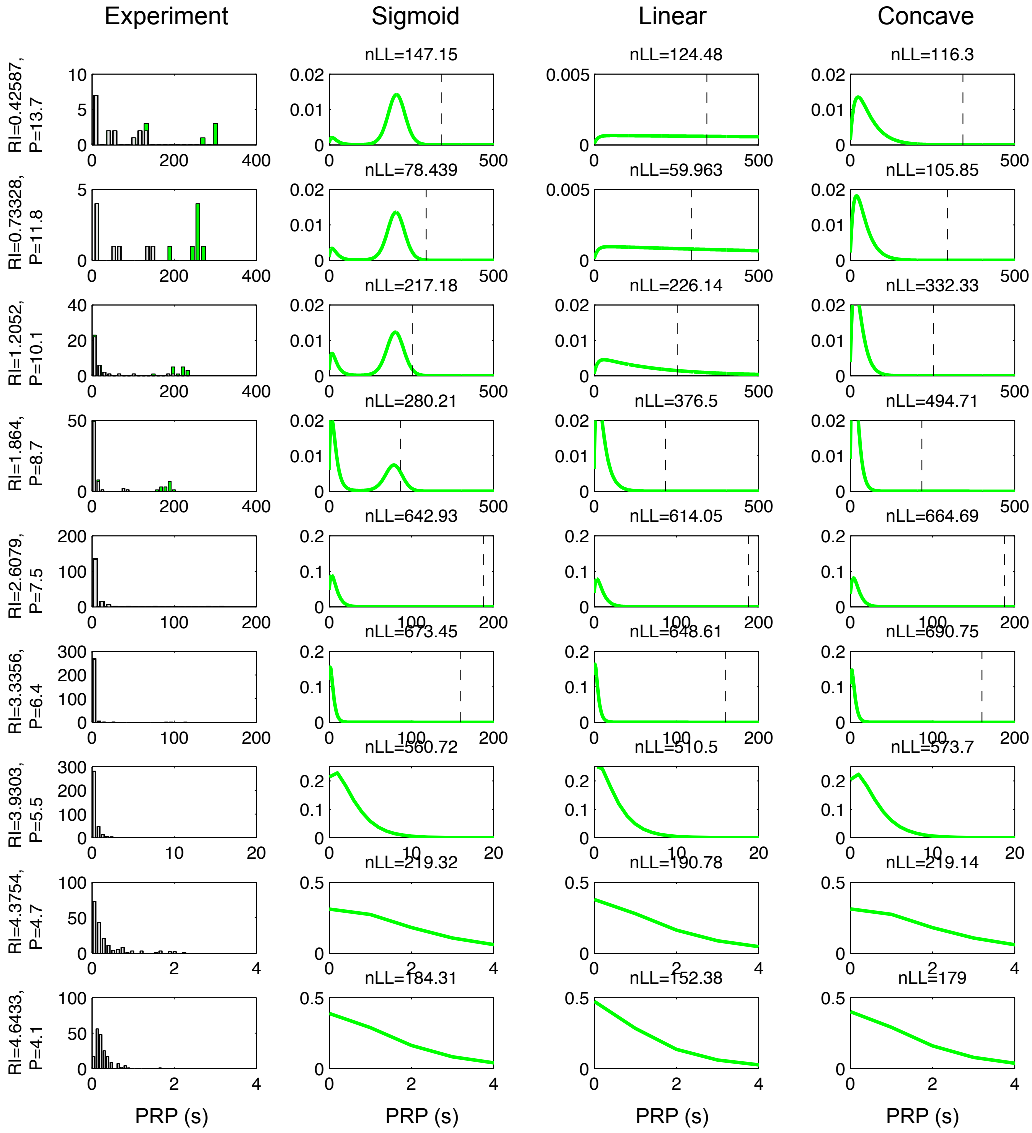


Figure 6.17: Model fits to PRPs on a radial 'sweep' for subject (F9). RI increases, Price decreases (payoff increases) from top to bottom panels, as denoted in the labels on the left. Left to right: Experiment, distributions predicted by best fit sigmoid, linear, and concave microscopic utilities of leisure. For experiment panels, coloured bars show censored data. PRP durations are at least as long as the duration on the x-axis. For model fits, numbers at the top give the negative log-likelihood (nLL) for that RI,P combination. Dashed lines show 25 × Price. Note that the axes scales change from condition to condition, but they are changed in pairs for the sake of comparison. Note the different x-axis scales for experimental data.

Chapter 7

Contributions and future work

This thesis makes six main contributions. The most important is a novel approach to characterising temporally relevant behaviour. Most previous research investigating temporal choices used macroscopic characterisations of behaviour Baum (2002, 2001, 2004, 1995); Baum and Rachlin (1969); Baum (1976), reporting average times and response rates. Following Niv et al. (2007), we characterise behaviour from a microscopic perspective, studying the detailed temporal topography of choice Ferster and Skinner (1957); Gilbert (1958); Shull et al. (2001); Williams et al. (2009b,a,c). Previous research using microscopic approaches Haccou and Meelis (1992); Breton (2013) have been descriptive, characterising what the animal does, rather than being normative: positing why it might do so. In this thesis, we developed a novel, normative, microscopic approach to characterising behaviour. This characterises all the choices that an individual makes over time, but from the perspective of its attempting to maximise its returns.

Secondly, we put forward a generic theoretical framework for studying the underlying computations, algorithmic mechanisms and neural implementations of real-time cost-benefit decision-making. By situating our framework within the broader theoretical framework of, and by employing the techniques from reinforcement learning, we ensured our formulations were theoretically sound and with relatively few ad-hoc assumptions. For instance, the linear opportunity cost of time (represented as the product of reward rate and time) critical to our models in Chapters 3-6 is not a simplifying assumption, but an automatic consequence of formulating our problems as average-reward semi-Markov decision processes (see Chapter 2). The formulation as a semi-Markov decision process was sensible since we proposed that subjects chose both actions and their durations simultaneously in order to approximately maximise their returns.

While we largely focussed at the computational level in this thesis, our framework provides a foundation for studying critical algorithmic and psychological processes and neural computations at appropriate timescales. Real-time or quasi-real-time recording methods in routine use in neuroscience such as electrophysiology, large-scale imaging, or fast-scan cyclic voltammetry allow us to correlate the activity of neural populations or concentrations of neuromodulators with the execution of behaviours. Likewise, fast causal manipulations via such methods as optogenetics allow the circuits governing these behaviours to be probed in a highly selective manner. There is an evident mismatch between the microscopic timescale over which these methods operate and the macroscopic timescales over which (a) behaviour has often been characterised; and (b) the quantities such as costs and benefits which underpin the pertinence of the behaviour have been defined. Our normative microscopic account may therefore provide an illuminating framework within which to build explanations that span multiple levels.

We applied our framework to the question of how to allocate limited time between work and leisure when both are attractive. The third contribution of this thesis is to provide a novel, normative, microscopic theory of this division. In order to explain the partial allocation of time between work and leisure, previous macroscopic research from labour supply theory had assumed an interaction between the marginal utilities of work and leisure–as if leisure is beneficial because of the recent history of work. We showed (see Chapter 4) that this assumption is not necessary when choices are microscopic and stochastic. We proposed that leisure is beneficial on its own accord and studied different functional forms which would reflect starkly different preferences for the duration of leisure. Similarly, accounts from behavioural psychology had proposed the need for stochasticity. We showed that, for certain microscopic utilities, deterministic choices were sufficient for achieving partial allocation. We showed that by integrating our microscopic choices we could build macroscopic characterisations that were not only equivalent to, but richer than those afforded by previous macroscopic characterisations. We therefore built a superset of traditional macroscopic quantifications.

Fourthly, we formulated normative, dynamic models of fatigue and satiation (Chapter 5). These extended our framework to the case where leisure was beneficial because of the recent history of work (in the case of fatigue) or rewards (in the case of satiation). We used the latter to generate a novel, normative microscopic cause for the backward bending labour supply curve that had been predicted or assumed by microeconomists and studied by behavioural economists and psychologists. Our models of fatigue and satiation provide new testable behavioural predictions at the microscopic level, which could be verified by future experimental research. In particular, we proposed two perspectives according to which decision-making could occur in the presence of fatigue or satiation: prospectively and retrospectively. Since both of these make similar behavioural predictions, we emphasised the need for careful experimental design, to avoid confounding one with the other, when attempting to understand which is playing a greater role.

Fifthly, we provided a microscopic characterisation of how rodent subjects distribute their work and leisure bouts. Along with Breton (2013), we showed that leisure bouts were bimodally distributed, with the mixture weight on the short mode increasing with payoff.

Finally, we provided, to our knowledge, the first empirical and quantitative study of the microscopic utility of leisure (Chapter 6)– and how it is related to those of rewards. We inferred this from rodent subjects, where we could control possible confounds such as fatigue, satiation and effort costs.

There are many possible directions for future work. The most imperative would be to fit the entire dataset with the full micro SMDP model. In fitting only PRPs with our simplified 2-state model we had discarded some amount of, possibly crucial, data. We also only considered two microscopic utility forms when we fitted the entire dataset in Chapter 6. It would be interesting to see how and why different subjects differed in their preference for durations of leisure when all the data are accounted for, and more utility functions are considered. We should also test these models against alternatives, such as the models of fatigue and satiation that we proposed.

We could then consider aspects of the neural implementation of the microscopic behaviour. Previous macroscopic analyses from pharmacological and drugs of addiction studies have revealed that an increase in the tonic release of the dopamine shifts the mountain towards longer prices Hernandez et al. (2010); Trujillo-Pisanty et al. (2011); Hernandez et al. (2012), as if, for instance, dopamine multiplies the intensity of the reward. Equally, models of instrumental vigour have posited that tonic dopamine computes or carries the average reward rate. This would realize the opportunity cost of time Niv et al. (2007); Cools et al. (2011); Dayan (2012). Finally, it has been suggested as being involved in overcoming the cost of effort Salamone and Correa (2002). However, the macroscopic mountain model cannot distinguish whether the effect of tonic dopamine is to make a reward more rewarding or a cost less costly or directly change increase the reward rate-they all yield the same time allocation. This further shows the limitation of macroscopic approaches to characterising behaviour in neuroscience, and suggests that microscopic approaches like ours, with their greater predictive power, are possibly more key to unpicking these cost-benefit computations. By analysing the microscopic data under dopamine agonists (or antagonists) when compared to control conditions, from the above pharmacological and drug studies, we could understand the computations performed or signals carried by tonic dopamine.

Our normative microscopic framework can be applied to a variety of tasks, which we did not consider in this thesis. For example, we could provide a novel method of studying the neural circuits and neuromodulation underlying effort costs Salamone and Correa (2002) and investigate whether these are different from those underlying fatigue. Similarly, we could attempt to provide an account of the computations performed by the neuromodulator serotonin in waiting through delays Miyazaki et al. (2011, 2012); Fletcher (1995); Jolly et al. (1999); Ho et al. (1998); Bizot et al. (1988, 1999).

Throughout the thesis, we modelled epochs in a trial after the reward intensity and price were known for sure. However, before subjects gain a minimal experience of the reward intensity and price on a trial, they face partial observability. They have to decide whether to explore (by depressing the lever to find out about the benefits of working) or exploit the option of leisure (albeit in ignorance of or greater uncertainty about the price). This leads to a form of optimal stopping problem. A particularly interesting case may arise when rewards are delivered probabilistically (according to a Bernoulli process, Breton (2013)). Given that the subject is extensively trained over many months and knows that the reward intensity and probability of reward delivery are fixed on a trial, it could explore till it inferred these more precisely. Subjects would have to infer whether a sequence of unrewarded work bouts is due to the reward intensity for the trial being low or because these were merely due to probabilistic reward (un)delivery. If the probability of reward delivery is high, then subjects may mistake a few unrewarded work bouts as evidence for the payoff on that trial being low. They would then unwittingly quit working on that trial after those few work bouts. This would be especially pernicious on high payoff trials since subjects would lose the opportunity to harvest lucrative rewards. Similarly, if the probability of reward delivery is low, then subjects should explore longer, even on high payoff trials. Our preliminary analyses (not shown here) revealed that rodent subjects are indeed capable of such sophisticated active inference. Macroscopic characterisations either simply ignore this exploration, and thus discard a significantly large proportion of data, or average across these. In either case, they mischaracterise highly informative data about the decision-making and inference process that the rat undergoes. Since our normative, microscopic approach, by definition attempts to decipher what an animal is trying to achieve in real-time, we would be able to account for the above behaviour. A crucial question then becomes what cost an animal would pay (temporally, eg. by manipulating the price, or in terms of effort or punishment) in order to gain an extra piece of information by working.

We previously mentioned that the experiments reported in this thesis are conducted in a sequence of triads–leading, test and trailing. The leading and trailing trials have the highest and lowest reward intensities, respectively. Preliminary analysis of data reveal that at the end of training, subjects rest for the entire duration of trailing trials (not shown here, but see Breton (2013)). That is, they do not even sample the lever once to infer the reward intensity and price parameters. Further analyses revealed that there are post-priming pauses (PPPs) that occur at the very beginning of each trial when the subject receives a priming train of stimulation to 'remind' or 'reset' the behaviour. In the case of the trailing trial, these PPPs last the entire trial. The PPPs are shortest for the leading trials and in-between for test trials. Except in the case of the trailing trial, PPPs are instrumentally deleterious. The subject should not waste any time engaging in leisure and should immediately attempt to receive at least one reward by working, so that it can infer the reward intensity and price parameters. Like a component of the PRP, PPPs also scale with expected payoff. We therefore consider PPPs to be Pavlovian. The PPPs reflect that the rat subjects know the random world triad structure of high-random-low payoff trials, despite the often long durations of test trials. This shows that the rat subjects are capable of possessing highly sophisticated representations about the structure of their world. When the PPPs were analysed over the entire training period, they were observed to start from the same duration for all three trial types. Over the course of training, the subjects learned the structure of the world. We aim to analyse this data further ourselves, and build a normative model of how a rodent builds such a sophisticated representation. For this we shall use an infinite hidden Markov model Beal (2002): a hidden Markov model whose number of hidden can states grow with the amount of data till they converge to the appropriate number of states. Thus, for this experiment, we should expect the number of states to grow from 1 to 3. A subject which has such a model of the world makes predictions of the payoff on the next trial based on the payoff experienced on the current one. Contrary to a model in which the subject simply counts trials, this could lead to confusion when, for instance, the payoff on the test trial is low. The subject could confuse this for a trailing trial and work on the subsequent trailing trial. It will be interesting to see how well our model predicts the data and how sophisticated a model of the world a rat can possess.

In conclusion, we are ordered with a wealth of data whose full psychological and neural implications can only be extracted by means of an account that provides a normative underpinning for its rich, real-time, details.

Bibliography

K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow. Capital-Labor Substitution and Economic Efficiency. *The Review of Economics and Statistics*, 43(3):225–250, 1961. (page 104)

A. Arvanitogiannis and P. Shizgal. The reinforcement mountain: allocation of behavior as a function of the rate and intensity of rewarding brain stimulation. *Behav Neurosci*, 122(5):1126–38, Oct. 2008. ISSN 0735-7044. doi: 10.1037/a0012679. (pages 40, 60, 62, 70, 95, and 116)

I. Barofsky and D. Hurwitz. Within ratio responding during fixed ratio performance. *Psychonomic Science*, 1968. URL `http://link.springer.com/article/10.3758/BF03327691`. (pages 129 and 139)

J. E. Barrett and J. A. Stanley. Effects of ethanol on multiple fixed-interval fixed-ratio schedule performances: dynamic interactions at different fixed-ratio values. *Journal of the experimental analysis of behavior*, 34(2):185–98, Sept. 1980. ISSN 0022-5002. doi: 10.1901/jeab.1980.34-185. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1332996&tool=pmcentrez&rendertype=abstract`. (page 53)

R. C. Battalio, L. Green, and J. H. Kagel. Income-Leisure Tradeoffs of Animal Workers. *The American Economic Review*, 71(4):621–632, 1981. (pages 40, 96, 129, and 139)

W. M. Baum. On two types of deviation from the matching law: bias and undermatching. *J Exp Anal Behav*, 22(1):231–42, July 1974. ISSN 0022-5002. (pages 40, 41, 57, 95, and 115)

W. M. Baum. Time-based and count-based measurement of preference. *J Exp Anal Behav*, 26(1):27–35, July 1976. ISSN 0022-5002. (pages 39 and 190)

W. M. Baum. Optimization and the matching law as accounts of instrumental behavior. *J Exp Anal Behav*, 36(3):387–403, Nov. 1981. ISSN 0022-5002. (page 40)

W. M. Baum. Choice, changeover, and travel. *Journal of the experimental analysis of behavior*, 38(1):35–49, July 1982. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347826&tool=pmcentrez&rendertype=abstract`. (page 59)

W. M. Baum. Performances on ratio and interval schedules of reinforcement: Data and theory. *J Exp Anal Behav*, 59(2):245–64, Mar. 1993. ISSN 0022-5002. (page 53)

W. M. Baum. Introduction to molar behavior analysis. *Mexican Journal of Behavior Analysis*, 21:7–25, 1995. (pages 39 and 190)

W. M. Baum. Molar versus as a paradigm clash. *J Exp Anal Behav*, 75(3):338–41; discussion 367–78, May 2001. ISSN 0022-5002. (pages 39 and 190)

W. M. Baum. From molecular to molar: a paradigm shift in behavior analysis. *J Exp Anal Behav*, 78(1):95–116, July 2002. ISSN 0022-5002. (pages 39 and 190)

W. M. Baum. Molar and molecular views of choice. *Behavioural Processes*, 66 (3):349–59, June 2004. ISSN 0376-6357. (pages 39 and 190)

W. M. Baum and H. C. Rachlin. Choice as time allocation. *J Exp Anal Behav*, 12(6):861–74, Nov. 1969. ISSN 0022-5002. (pages 39, 40, 60, 104, 115, and 190)

R. Baumeister. Ego depletion and self-control failure: An energy model of the self's executive function. *Self and Identity*, 2002. URL `http://www.tandfonline.com/doi/abs/10.1080/152988602317319302`. (pages 129 and 148)

R. Baumeister and E. Bratslavsky. Ego depletion: is the active self a limited resource? *Journal of personality . . .*, 1998. URL `http://psycnet.apa.org/journals/psp/74/5/1252/`. (pages 129 and 148)

R. Baumeister, K. Vohs, and D. Tice. The strength model of self-control. *Current directions in . . .*, 2007. URL `http://cdp.sagepub.com/content/16/6/351.short`. (page 148)

H. M. Bayer and P. W. Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1): 129–41, July 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.05.020. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1564381&tool=pmcentrez&rendertype=abstract`. (page 40)

M. Beal. The infinite hidden Markov model. *Advances in neural* ..., 2002. URL `https://papers.nips.cc/paper/1956-the-infinite-hidden-markov-model.pdf`. (page 194)

K. C. Berridge. The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191(3):391–431, Apr. 2007. ISSN 0033-3158. doi: 10.1007/s00213-006-0578-x. (page 98)

C. Bielajew and P. Shizgal. Behaviorally derived measures of conduction velocity in the substrate for rewarding medial forebrain bundle stimulation. *Brain research*, 237(1):107–19, Apr. 1982. ISSN 0006-8993. URL `http://www.ncbi.nlm.nih.gov/pubmed/7074353`. (page 69)

C. Bielajew and P. Shizgal. Evidence implicating descending fibers in self-stimulation of the medial forebrain bundle. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 6(4):919–29, Apr. 1986. ISSN 0270-6474. URL `http://www.ncbi.nlm.nih.gov/pubmed/3486258`. (page 69)

G. Bigelow and I. Liebson. Cost factors controlling alcoholic drinking. *The Psychological Record*, 1972. URL `http://psycnet.apa.org/psycinfo/1973-02797-001`. (pages 129 and 139)

L. Bizo, S. Bogdanov, and P. Killeen. Satiation causes within-session decreases in instrumental responding. *Journal of Experimental* ..., 1998. URL `http://psycnet.apa.org/journals/xan/24/4/439/`. (pages 130 and 151)

J. Bizot, C. Le Bihan, A. J. Puech, M. Hamon, and M. Thibot. Serotonin and tolerance to delay of reward in rats. *Psychopharmacology (Berl)*, 146(4):400–412, Oct 1999. (pages 94 and 193)

J. C. Bizot, M. H. Thibot, C. Le Bihan, P. Soubri, and P. Simon. Effects of imipramine-like drugs and serotonin uptake blockers on delay of reward in rats. possible implication in the behavioral mechanism of action of antidepressants. *J Pharmacol Exp Ther*, 246(3):1144–1151, Sep 1988. (pages 94 and 193)

R. Blundell and T. Macurdy. Labor supply: A review of alternative approaches. In *Handbook of Labor Economics*, volume 3, Part A, pages 1559–1695. Elsevier, 1999. URL `http://ideas.repec.org/h/eee/labchp/3-27.html`. (pages 51 and 121)

H. Boelens and P. F. Kop. Concurrent schedules: Spatial separation of response alternatives. *Journal of the experimental analysis of behavior*, 40(1):35–45, July 1983. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347842&tool=pmcentrez&rendertype=abstract`. (page 59)

M. M. Botvinick, S. Huffstetler, and J. T. McGuire. Effort discounting in human nucleus accumbens. *Cognitive, affective & behavioral neuroscience*, 9(1):16–27, Mar. 2009. ISSN 1530-7026. doi: 10.3758/CABN.9.1.16. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2744387&tool=pmcentrez&rendertype=abstract`. (pages 122 and 153)

C. M. Bradshaw, E. Szabadi, and P. Bevan. The effect of punishment on free-operant choice behavior in humans. *Journal of the experimental analysis of behavior*, 31(1):71–81, Jan. 1979. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1332790&tool=pmcentrez&rendertype=abstract`. (page 53)

C. M. Bradshaw, H. V. Ruddle, and E. Szabadi. Relationship between response rate and reinforcement frequency in variable-interval schedules: III. The effect of d-amphetamine. *Journal of the experimental analysis of behavior*, 36(1):29–39, July 1981a. ISSN 0022-5002. doi: 10.1901/jeab.1981.36-29. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1333050&tool=pmcentrez&rendertype=abstract`. (page 53)

C. M. Bradshaw, H. V. Ruddle, and E. Szabadi. Relationship between response rate and reinforcement frequency in variable-interval schedules: II. Effect of the volume of sucrose reinforcement. *Journal of the experimental analysis of behavior*, 35(3):263–9, May 1981b. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1333081&tool=pmcentrez&rendertype=abstract`. (page 53)

K. Breland and M. Breland. The misbehavior of organisms. *Am Psychol*, 16(11): 681–684, 1961. ISSN 0003-066X. doi: 10.1037/h0040090. (page 78)

Y. Breton, K. Conover, and P. Shizgal. Probability discounting of brain stimulation reward in the rat. Number 892.14, October 2009a. 39th Annual Meeting of the Society for Neuroscience (Neuroscience 2009). (pages 19, 73, and 74)

Y.-A. Breton. *Molar and Molecular Models of Performance for Rewarding Brain Stimulation.* Phd thesis, Concordia University, 2013. (pages 62, 64, 70, 73, 75, 104, 184, 190, 192, 193, and 194)

Y.-A. Breton, J. C. Marcus, and P. Shizgal. Rattus Psychologicus: construction of preferences by self-stimulating rats. *Behav Brain Res*, 202(1):77–91, Aug. 2009b. ISSN 1872-7549. doi: 10.1016/j.bbr.2009.03.019. (pages 40, 54, 56, and 72)

C. Camerer, L. Babcock, G. Loewenstein, and R. Thaler. Labor Supply of New York City Cabdrivers: One Day at a Time. *Q J Econ*, 112(2):407–441, May 1997. ISSN 0033-5533. doi: 10.1162/003355397555244. (pages 40, 96, 129, 139, 141, 142, 143, 150, and 151)

A. Caplin and M. Dean. Axiomatic methods, dopamine and reward prediction error. *Current opinion in neurobiology*, 18(2):197–202, Apr. 2008. ISSN 0959-4388. doi: 10.1016/j.conb.2008.07.007. URL `http://www.ncbi.nlm.nih.gov/pubmed/18678251`. (page 107)

E. Charnov. Optimal foraging, the marginal value theorem. *Theoretical population biology*, 9(2):129–136, 1976. URL `http://scholar.google.co.uk/scholar?hl=en&q=charnov+1976&btnG=&as_sdt=1%2C5&as_sdtp=#1`. (page 108)

Y. Chou. Testing alternative models of labour supply: Evidence from taxi drivers in Singapore. *The Singapore Economic Review*, 47(17):17–47, 2002. URL `http://www.worldscientific.com/doi/abs/10.1142/S0217590802000389`. (pages 141 and 142)

J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, and N. Uchida. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383):85–8, Feb. 2012. ISSN 1476-4687. doi: 10.1038/nature10754. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3271183&tool=pmcentrez&rendertype=abstract`. (page 40)

G. Collier and W. Jennings. Work as a determinant of instrumental performance. *Journal of Comparative and Physiological ...*, 1969. URL `http://psycnet.apa.org/journals/com/68/4/659/`. (pages 129 and 139)

G. Collier and D. A. Levitsky. Operant running as a function of deprivation and effort. *Journal of comparative and physiological psychology*, 66(2):522–3, Oct. 1968. ISSN 0021-9940. URL `http://www.ncbi.nlm.nih.gov/pubmed/5722068`. (page 152)

G. Collier, E. Hirsch, D. Levitsky, and A. I. Leshner. Effort as a dimension of spontaneous activity in rats. *Journal of comparative and physiological psychology*, 88(1):89–96, Jan. 1975. ISSN 0021-9940. URL `http://www.ncbi.nlm.nih.gov/pubmed/1120820`. (page 152)

K. L. Conover and P. Shizgal. Differential effects of postingestive feedback on the reward value of sucrose and lateral hypothalamic stimulation in rats. *Behav Neurosci*, 108(3):559–72, June 1994a. ISSN 0735-7044. (page 68)

K. L. Conover and P. Shizgal. Competition and summation between rewarding effects of sucrose and lateral hypothalamic stimulation in the rat. *Behav Neurosci*, 108(3):537–48, June 1994b. ISSN 0735-7044. (page 68)

K. L. Conover and P. Shizgal. Employing labor-supply theory to measure the reward value of electrical brain stimulation. *Games and Economic Behavior*, 52(2):283–304, Aug. 2005. ISSN 08998256. doi: 10.1016/j.geb.2004.08.003. (pages 40, 96, and 104)

K. L. Conover, B. Woodside, and P. Shizgal. Effects of sodium depletion on competition and summation between rewarding effects of salt and lateral hypothalamic stimulation in the rat. *Behav Neurosci*, 108(3):549–58, June 1994. ISSN 0735-7044. (page 68)

R. Cools, K. Nakamura, and N. D. Daw. Serotonin and dopamine: unifying affective, activational, and decision functions. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 36(1): 98–113, Jan. 2011. ISSN 1740-634X. doi: 10.1038/npp.2010.121. (pages 40, 41, 65, 66, 98, 99, and 192)

P. L. Croxson, M. E. Walton, J. X. O'Reilly, T. E. J. Behrens, and M. F. S. Rushworth. Effort-based cost-benefit valuation and the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(14):4531–41, Apr. 2009. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.4515-08.2009. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2954048&tool=pmcentrez&rendertype=abstract`. (pages 152 and 153)

J. K. Dagsvik, Z. Jia, T. Kornstad, and T. O. Thoresen. Theoretical and practical arguments for modeling labor supply as a choice among latent jobs. Technical report, 2012. (pages 51 and 105)

J. Dallery and P. L. Soto. Herrnstein's hyperbolic matching equation and behavioral pharmacology: review and critique. *Behavioural pharmacology*, 15(7): 443–59, Nov. 2004. ISSN 0955-8810. URL http://www.ncbi.nlm.nih.gov/pubmed/15472567. (page 59)

J. Dallery, J. J. McDowell, and J. S. Lancaster. Falsification of matching theory's account of single-alternative responding: Herrnstein's k varies with sucrose concentration. *J Exp Anal Behav*, 73(1):23–43, Jan. 2000. ISSN 0022-5002. doi: 10.1901/jeab.2000.73-23. (page 40)

N. D. Daw and D. S. Touretzky. Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, 14(11):2567–83, Nov. 2002. ISSN 0899-7667. doi: 10.1162/089976602760407973. (pages 66, 76, 99, and 108)

P. Dayan. Instrumental vigour in punishment and reward. *Eur J Neurosci*, 35(7): 1152–1168, Apr. 2012. ISSN 0953816X. doi: 10.1111/j.1460-9568.2012.08026.x. (pages 40, 41, 65, 66, 98, 99, 105, and 192)

P. Dayan, Y. Niv, B. Seymour, and N. D. Daw. The misbehavior of value and the discipline of the will. *Neural Net*, 19(8):1153–60, Oct. 2006. ISSN 0893-6080. doi: 10.1016/j.neunet.2006.03.002. (page 78)

E. Diener, D. Wirtz, and S. Oishi. End Effects of Rated Life Quality: The James Dean Effect. *Psychological Science*, 12(2):124–128, Mar. 2001. ISSN 0956-7976. doi: 10.1111/1467-9280.00321. (page 95)

H. Dienske and J. Metz. Mother-infant body contact in macaques: a time interval analysis. *Biology of behaviour*, 1977. URL http://scholar.google.co.uk/scholar?hl=en&q=Dienske+and+Metz+1977&btnG=&as_sdt=1%2C5&as_sdtp=#1. (page 64)

J. A. Dinsmoor. The effect of hunger on discriminated responding. *Journal of abnormal and social psychology*, 47(1):67–72, Jan. 1952. ISSN 0096-851X. URL http://www.ncbi.nlm.nih.gov/pubmed/14907249. (page 151)

M. Domjan. Stepping outside the box in considering the C/T ratio. *Behavioural processes*, 62(1-3):103–114, Apr. 2003. ISSN 1872-8308. URL http://www.ncbi.nlm.nih.gov/pubmed/12729972. (page 53)

P. Dupas and J. Robinson. Daily Needs, Income Targets and Labor Supply: Evidence from Kenya. Aug. 2013. URL http://www.nber.org/papers/w19264. (pages 96 and 151)

D. E. Edmonds and C. R. Gallistel. 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):876–83, Nov. 1974. ISSN 0021-9940. URL http://www.ncbi.nlm.nih.gov/pubmed/4430753. (page 71)

D. E. Edmonds, J. R. Stellar, and C. R. Gallistel. Parametric analysis of brain stimulation reward in the rat: II. Temporal summation in the reward system. *Journal of comparative and physiological psychology*, 87(5):860–9, Nov. 1974. ISSN 0021-9940. URL http://www.ncbi.nlm.nih.gov/pubmed/4430751. (page 71)

H. S. Farber. Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers. *Journal of Political Economy*, 113(1):46–82, 2005. ISSN 00223808. doi: 10.1086/426040. URL http://www.journals.uchicago.edu/cgi-bin/resolve?id=doi:10.1086/426040. (pages 141, 142, and 150)

E. Fehr and L. Goette. Do workers work more if wages are high? Evidence from a randomized field experiment. *The American Economic Review*, 2007. URL http://www.jstor.org/stable/30034396. (pages 142 and 150)

M. Felton and D. O. Lyon. The post-reinforcement pause. *Journal of the experimental analysis of behavior*, 9(2):131–4, Mar. 1966. ISSN 0022-5002. doi: 10.1901/jeab.1966.9-131. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338164&tool=pmcentrez&rendertype=abstract. (page 54)

C. Ferster and B. F. Skinner. *Schedules of reinforcement.* Appleton-Century-Crofts, New York, 1957. (pages 39, 53, 62, and 190)

K. Fischer and E. Fantino. The dissociation of discriminative and conditioned reinforcing functions of stimuli with changes in deprivation. *Journal of the experimental analysis of behavior*, 11(6):703–10, Nov. 1968. ISSN 0022-5002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338623&tool=pmcentrez&rendertype=abstract. (pages 129 and 151)

P. J. Fletcher. Effects of combined or separate 5,7-dihydroxytryptamine lesions of the dorsal and median raphe nuclei on responding maintained by a drl 20s schedule of food reinforcement. *Brain Res*, 675(1-2):45–54, Mar 1995. (pages 94 and 193)

S. B. Floresco and S. Ghods-Sharifi. Amygdala-prefrontal cortical circuitry regulates effort-based decision making. *Cerebral cortex (New York, N.Y. : 1991)*, 17(2):251–60, Feb. 2007. ISSN 1047-3211. URL `http://www.ncbi.nlm.nih.gov/pubmed/16495432`. (pages 152 and 153)

S. B. Floresco, J. R. St Onge, S. Ghods-Sharifi, and C. A. Winstanley. Cortico-limbic-striatal circuits subserving different forms of cost-benefit decision making. *Cognitive, affective & behavioral neuroscience*, 8(4):375–89, Dec. 2008a. ISSN 1530-7026. URL `http://www.ncbi.nlm.nih.gov/pubmed/19033236`. (page 152)

S. B. Floresco, M. T. L. Tse, and S. Ghods-Sharifi. Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 33(8):1966–79, July 2008b. ISSN 0893-133X. URL `http://dx.doi.org/10.1038/sj.npp.1301565`. (page 152)

M. Foster, K. Blackman, and W. Temple. Open versus closed economies: performance of domestic hens under fixed ratio schedules. *Journal of the experimental analysis of behavior*, 67(1):67–89, Jan. 1997. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1284588&tool=pmcentrez&rendertype=abstract`. (page 53)

G. Fouriezos, C. Bielajew, and W. Pagotto. Task difficulty increases thresholds of rewarding brain stimulation. *Behavioural brain research*, 37(1):1–7, Feb. 1990. ISSN 0166-4328. URL `http://www.ncbi.nlm.nih.gov/pubmed/2310490`. (page 60)

R. H. Frank. *Microeconomics and Behavior.* McGraw-Hill Higher Education, 2005. ISBN 0071115498. (pages 40, 41, 47, 96, and 104)

K. Franklin. Catecholamines and self-stimulation: Reward and performance effects dissociated. *Pharmacology Biochemistry and Behavior*, 9(6):813–820, Dec. 1978. ISSN 00913057. doi: 10.1016/0091-3057(78)90361-1. URL `http://www.sciencedirect.com/science/article/pii/0091305778903611`. (page 71)

M. Gailliot and R. Baumeister. Self-control relies on glucose as a limited energy source: willpower is more than a metaphor. *Journal of personality ...*, 2007. URL `http://psycnet.apa.org/journals/psp/92/2/325/`. (page 148)

C. R. Gallistel and J. Gibbon. Time, rate, and conditioning. *Psychological review*, 107(2):289–344, Apr. 2000. ISSN 0033-295X. URL http://www.ncbi.nlm.nih.gov/pubmed/10789198. (page 53)

C. R. Gallistel and M. Leon. Measuring the subjective magnitude of brain stimulation reward by titration with rate of reward. *Behav Neurosci*, 105(6):913–25, Dec. 1991. ISSN 0735-7044. (pages 60 and 70)

C. R. Gallistel, J. R. Stellar, and E. Bubis. 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):848–59, Nov. 1974. ISSN 0021-9940. URL http://www.ncbi.nlm.nih.gov/pubmed/4430750. (pages 70 and 71)

J. O. Gan, M. E. Walton, and P. E. M. Phillips. Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. *Nature neuroscience*, 13(1):25–7, Jan. 2010. ISSN 1546-1726. doi: 10.1038/nn.2460. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2800310&tool=pmcentrez&rendertype=abstract. (pages 152 and 153)

S. Ghods-Sharifi and S. B. Floresco. Differential effects on effort discounting induced by inactivations of the nucleus accumbens core or shell. *Behavioral neuroscience*, 124(2):179–91, Apr. 2010. ISSN 1939-0084. doi: 10.1037/a0018932. URL http://www.ncbi.nlm.nih.gov/pubmed/20364878. (pages 152 and 153)

J. Gibbon. Scalar expectancy theory and Weber's law in animal timing. *Psych Rev*, 84(3):279–325, 1977. (pages 53 and 93)

T. F. Gilbert. Fundamental dimensional properties of the operant. *Psych Rev*, 65(5):272–82, Sept. 1958. ISSN 0033-295X. (pages 39, 62, and 190)

L. Goette and D. Huffman. Incentives and the allocation of effort over time: the joint role of affective and cognitive decision making. 2006. URL http://www.econstor.eu/handle/10419/33697. (pages 142, 150, and 151)

L. Green and H. Rachlin. Economic substitutability of electrical brain stimulation, food, and water. *J Exp Anal Behav*, 55(2):133–43, Mar. 1991. ISSN 0022-5002. doi: 10.1901/jeab.1991.55-133. (pages 40 and 68)

L. Green, J. H. Kagel, and R. C. Battalio. Consumption-leisure tradeoffs in pigeons: Effects of changing marginal wage rates by varying amount of reinforcement. *J Exp Anal Behav*, 47(1):17–28, Jan. 1987. ISSN 0022-5002. (pages 40, 96, 129, and 139)

M. Guitart-Masip, L. Fuentemilla, D. R. Bach, Q. J. M. Huys, P. Dayan, R. J. Dolan, and E. Duzel. Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *J Neurosci*, 31(21):7867–75, May 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.6376-10.2011. (page 78)

P. Haccou and E. Meelis. *Statistical Analysis of Behavioural Data: An Approach Based on Time-structured Models*. Oxford University Press, USA, 1992. ISBN 0198546637. (pages 40, 41, 64, 97, and 190)

A. L. Hamilton, J. R. Stellar, and E. B. Hart. Reward, performance, and the response strength method in self-stimulating rats: validation and neuroleptics. *Phys & Behav*, 35(6):897–904, Dec. 1985. ISSN 0031-9384. (pages 60 and 70)

G. Hanoch. The" Backward-bending" Supply of Labor. *The Journal of Political Economy*, 1965. URL `http://www.jstor.org/stable/1829888`. (pages 129 and 139)

A. S. Hart, R. B. Rutledge, P. W. Glimcher, and P. E. M. Phillips. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(3):698–704, Jan. 2014. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2489-13.2014. URL `http://www.ncbi.nlm.nih.gov/pubmed/24431428`. (page 107)

G. Hernandez, S. Hamdani, H. Rajabi, K. Conover, J. Stewart, A. Arvanitogiannis, and P. Shizgal. Prolonged rewarding stimulation of the rat medial forebrain bundle: neurochemical and behavioral consequences. *Behavioral neuroscience*, 120(4):888–904, Aug. 2006. ISSN 0735-7044. doi: 10.1037/0735-7044.120.4.888. URL `http://www.ncbi.nlm.nih.gov/pubmed/16893295`. (page 69)

G. Hernandez, E. Haines, and P. Shizgal. Potentiation of intracranial self-stimulation during prolonged subcutaneous infusion of cocaine. *Journal of neuroscience methods*, 175(1):79–87, Oct. 2008. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2008.08.005. URL `http://www.ncbi.nlm.nih.gov/pubmed/18765253`. (pages 59 and 71)

G. Hernandez, Y.-A. Breton, K. Conover, and P. Shizgal. At what stage of neural processing does cocaine act to boost pursuit of rewards? *PloS one*, 5 (11):e15081, Jan. 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0015081. (pages 17, 18, 40, 61, 62, 63, 70, 71, 72, 95, 98, 116, 122, and 192)

G. Hernandez, I. Trujillo-Pisanty, M.-P. Cossette, K. Conover, and P. Shizgal. Role of Dopamine Tone in the Pursuit of Brain Stimulation Reward. *J Neurosci*, 32(32):11032–11041, Aug. 2012. ISSN 0270-6474. (pages 40, 62, 70, 71, 98, 122, and 192)

R. J. Herrnstein. Relative and absolute strength of response as a function of frequency of reinforcement. *J Exp Anal Behav*, 4:267–72, July 1961. ISSN 0022-5002. doi: 10.1901/jeab.1961.4-267. (pages 16, 40, 41, 57, 58, 59, 95, and 115)

R. J. Herrnstein. On the law of effect. *J Exp Anal Behav*, 13(2):243–66, Mar. 1970. ISSN 0022-5002. (pages 16, 53, 57, 58, and 59)

R. J. Herrnstein. Formal properties of the matching law. *J Exp Anal Behav*, 21 (1):159–64, Jan. 1974. ISSN 0022-5002. (pages 40, 41, and 95)

M. Y. Ho, S. S. Al-Zahrani, A. S. Al-Ruwaitea, C. M. Bradshaw, and E. Szabadi. 5-hydroxytryptamine and impulse control: prospects for a behavioural analysis. *J Psychopharmacol*, 12(1):68–78, 1998. (pages 94 and 193)

W. Hodos and E. S. Valenstein. An evaluation of response rate as a measure of rewarding intracranial stimulation. *Journal of comparative and physiological psychology*, 55:80–4, Feb. 1962. ISSN 0021-9940. URL `http://www.ncbi.nlm.nih.gov/pubmed/13907981`. (pages 68 and 69)

J. R. Hollerman, L. Tremblay, and W. Schultz. Involvement of basal ganglia and orbitofrontal cortex in goal-directed behavior. *Progress in brain research*, 126: 193–215, Jan. 2000. ISSN 0079-6123. doi: 10.1016/S0079-6123(00)26015-9. URL `http://www.ncbi.nlm.nih.gov/pubmed/11105648`. (page 40)

T. Hosokawa, S. W. Kennerley, J. Sloan, and J. D. Wallis. Single-neuron mechanisms underlying cost-benefit analysis in frontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(44):17385–97, Oct. 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2221-13.2013. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3812506&tool=pmcentrez&rendertype=abstract`. (pages 152 and 153)

D. C. Jolly, J. B. Richards, and L. S. Seiden. Serotonergic mediation of drl 72s behavior: receptor subtype involvement in a behavioral screen for antidepressant drugs. *Biol Psychiatry*, 45(9):1151–1162, May 1999. (pages 94 and 193)

A. Kacelnik and B. Marsh. Cost can increase preference in starlings. *Anim Behav*, 63(2):245–250, 2002. (pages 96, 121, and 158)

J. H. Kagel, R. C. Battalio, and L. Green. *Economic Choice Theory: An Experimental Analysis of Animal Behavior.* Cambridge University Press, 1995. (pages 40, 50, and 96)

R. B. Kanarek and G. Collier. Effort as a determinant of choice in rats. *Journal of comparative and physiological psychology*, 84(2):332–8, Aug. 1973. ISSN 0021-9940. URL http://www.ncbi.nlm.nih.gov/pubmed/4723929. (page 152)

S. W. Kennerley, M. E. Walton, T. E. J. Behrens, M. J. Buckley, and M. F. S. Rushworth. Optimal decision making and the anterior cingulate cortex. *Nature neuroscience*, 9(7):940–7, July 2006. ISSN 1097-6256. doi: 10.1038/nn1724. URL http://www.ncbi.nlm.nih.gov/pubmed/16783368. (pages 152 and 153)

S. W. Kennerley, A. F. Dahmubed, A. H. Lara, and J. D. Wallis. Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of cognitive neuroscience*, 21(6):1162–78, June 2009. ISSN 0898-929X. doi: 10.1162/jocn.2009.21100. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2715848&tool=pmcentrez&rendertype=abstract. (pages 152 and 153)

P. Killeen. The matching law. *J Exp Anal Behav*, 17(3):489–95, May 1972. ISSN 0022-5002. (pages 60 and 116)

P. R. Killeen. Economics, ecologics, and mechanics: The dynamics of responding under conditions of varying motivation. *Journal of the experimental analysis of behavior*, 64(3):405–31, Nov. 1995. ISSN 0022-5002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1350147&tool=pmcentrez&rendertype=abstract. (pages 53, 129, and 151)

W. Kool and M. Botvinick. A Labor/Leisure Tradeoff in Cognitive Control. *Journal of experimental psychology. General*, Dec. 2012. ISSN 1939-2222. doi: 10.1037/a0031048. URL http://www.ncbi.nlm.nih.gov/pubmed/23230991. (pages 15, 48, 121, 122, and 149)

I. T. Kurniawan, B. Seymour, D. Talmi, W. Yoshida, N. Chater, and R. J. Dolan. Choosing to make an effort: the role of striatum in signaling physical effort of a chosen action. *Journal of neurophysiology*, 104(1):313–21, July 2010. ISSN 1522-1598. doi: 10.1152/jn.00027.2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2904211&tool=pmcentrez&rendertype=abstract. (pages 152 and 153)

I. T. Kurniawan, M. Guitart-Masip, P. Dayan, and R. J. Dolan. Effort and valuation in the brain: the effects of anticipation and execution. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(14):6160–9, Apr. 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.4777-12.2013. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3639311&tool=pmcentrez&rendertype=abstract`. (pages 122, 152, and 153)

R. Kurzban. Does the brain consume additional glucose during self-control tasks? *Evolutionary psychology : an international journal of evolutionary approaches to psychology and behavior*, 8(2):244–59, Jan. 2010. ISSN 1474-7049. URL `http://www.ncbi.nlm.nih.gov/pubmed/22947794`. (page 148)

M. Leon and C. R. Gallistel. The function relating the subjective magnitude of brain stimulation reward to stimulation strength varies with site of stimulation. *Behav Brain Res*, 52(2):183–93, Dec. 1992. ISSN 0166-4328. (pages 60 and 70)

M. Lyon and T. Robbins. The action of central nervous system stimulant drugs: a general theory concerning amphetamine effects. *Current Developments in Psychopharmcology*, 2:79–163, 1975. (page 98)

T. A. Mark and C. R. Gallistel. Subjective reward magnitude of medial forebrain stimulation as a function of train duration and pulse frequency. *Behav Neurosci*, 107(2):389–401, Apr. 1993. ISSN 0735-7044. (pages 60 and 70)

D. Marr. *Vision*. W.H.Freeman & Co Ltd, 1982. ISBN 0716712849. (page 40)

J. E. Mazur. Steady-state performance on fixed-, mixed-, and random-ratio schedules. *Journal of the experimental analysis of behavior*, 39(2):293–307, Mar. 1983. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347922&tool=pmcentrez&rendertype=abstract`. (page 53)

S. M. McClure, N. D. Daw, and P. R. Montague. A computational substrate for incentive salience. *Trends in neurosciences*, 26(8):423–8, Aug. 2003. ISSN 0166-2236. URL `http://www.ncbi.nlm.nih.gov/pubmed/12900173`. (page 40)

J. J. McDowell. On the falsifiability of matching theory. *J Exp Anal Behav*, 45 (1):63–74, Jan. 1986. ISSN 0022-5002. doi: 10.1901/jeab.1986.45-63. (pages 40, 41, and 95)

J. J. McDowell. On the classic and modern theories of matching. *J Exp Anal Behav*, 84(1):111–27, July 2005. ISSN 0022-5002. doi: 10.1901/jeab.2005.59-04. (pages 40, 41, and 95)

D. L. McFadden. Econometric analysis of qualitative response models. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2 of *Handbook of Econometrics*, chapter 24, pages 1395–1457. Elsevier, December 1984. URL `http://ideas.repec.org/h/eee/ecochp/2-24.html`. (pages 50, 51, and 105)

F. McSweeney. Rate of reinforcement and session duration as determinants of within-session patterns of responding. *Animal Learning & Behavior*, 1992. URL `http://link.springer.com/article/10.3758/BF03200413`. (pages 129 and 151)

F. McSweeney and K. Johnson. The effect of time between sessions on within-session patterns of responding. *Behavioural Processes*, 1994. URL `http://www.sciencedirect.com/science/article/pii/0376635794900078`. (pages 129 and 151)

F. McSweeney and E. Murphy. Criticisms of the Satiety Hypothesis as an Explanation for WithinSession Decreases in Responding. *Journal of the Experimental ...*, 2000. URL `http://onlinelibrary.wiley.com/doi/10.1901/jeab.2000.74-347/abstract`. (page 151)

F. McSweeney, J. Hatfield, and T. Allen. Within-session responding as a function of post-session feedings. *Behavioural Processes*, 1991. URL `http://www.sciencedirect.com/science/article/pii/037663579190092E`. (pages 129 and 151)

F. K. McSweeney and J. M. Roll. Responding changes systematically within sessions during conditioning procedures. *Journal of the experimental analysis of behavior*, 60(3):621–40, Nov. 1993. ISSN 0022-5002. doi: 10.1901/jeab.1993.60-621. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1322170&tool=pmcentrez&rendertype=abstract`. (pages 129 and 151)

F. K. McSweeney, C. L. Melville, M. A. Buck, and J. E. Whipple. Local rates of responding and reinforcement during concurrent schedules. *Journal of the experimental analysis of behavior*, 40(1):79–98, July 1983. ISSN 0022-5002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347846&tool=pmcentrez&rendertype=abstract`. (pages 16 and 58)

R. Meisch and T. Thompson. Ethanol as a reinforcer: Effects of fixed-ratio size and food deprivation. *Psychopharmacologia*, 1973. URL `http://link.springer.com/article/10.1007/BF00421402`. (pages 129 and 139)

R. Meisch and T. Thompson. Rapid establishment of ethanol as a reinforcer for rats. *Psychopharmacologia*, 1974a. URL `http://link.springer.com/article/10.1007/BF00428917`. (pages 129 and 139)

R. Meisch and T. Thompson. Ethanol intake as a function of concentration during food deprivation and satiation. *Pharmacology Biochemistry and Behavior*, 1974b. URL `http://www.sciencedirect.com/science/article/pii/0091305774900252`. (pages 129 and 139)

F. Meyniel, C. Sergent, L. Rigoux, J. Daunizeau, and M. Pessiglione. Neurocomputational account of how the human brain decides when to have a break. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7):2641–6, Feb. 2013. ISSN 1091-6490. doi: 10.1073/pnas.1211925110. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3574948&tool=pmcentrez&rendertype=abstract`. (pages 122, 129, and 153)

F. Meyniel, L. Safra, and M. Pessiglione. How the brain decides when to work and when to rest: dissociation of implicit-reactive from explicit-predictive computational processes. *PLoS computational biology*, 10(4): e1003584, Apr. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003584. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3990494&tool=pmcentrez&rendertype=abstract`. (page 153)

E. Miliaressis, P. P. Rompre, P. Laviolette, L. Philippe, and D. Coulombe. The curve-shift paradigm in self-stimulation. *Physiology & behavior*, 37(1):85–91, Jan. 1986. ISSN 0031-9384. URL `http://www.ncbi.nlm.nih.gov/pubmed/3016774`. (page 59)

K. Miyazaki, K. W. Miyazaki, and K. Doya. Activation of dorsal raphe serotonin neurons underlies waiting for delayed rewards. *J Neurosci*, 31(2):469–79, Jan. 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.3714-10.2011. (pages 94 and 193)

K. W. Miyazaki, K. Miyazaki, and K. Doya. Activation of dorsal raphe serotonin neurons is necessary for waiting for delayed rewards. *J Neurosci*, 32(31):10451–7, Aug. 2012. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.0915-12.2012. (pages 94 and 193)

P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 16(5):1936–47, Mar. 1996. ISSN 0270-6474. URL `http://www.ncbi.nlm.nih.gov/pubmed/8774460`. (pages 40 and 69)

B. Murray and P. Shizgal. 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*, 19(2):135–146, June 1991. ISSN 0889-6313. doi: 10.3758/BF03327183. URL `http://link.springer.com/article/10.3758/BF03327183`. (page 71)

R. Nieuwenhuys, L. M. Geeraedts, and J. G. Veening. The medial forebrain bundle of the rat. I. General introduction. *The Journal of comparative neurology*, 206 (1):49–81, Mar. 1982. ISSN 0021-9967. doi: 10.1002/cne.902060106. URL `http://www.ncbi.nlm.nih.gov/pubmed/6124562`. (page 68)

Y. Niv, N. D. Daw, D. Joel, and P. Dayan. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–20, Apr. 2007. ISSN 0033-3158. doi: 10.1007/s00213-006-0502-4. (pages 40, 41, 64, 65, 66, 75, 76, 98, 99, 105, 108, 152, 190, and 192)

R. K. Niyogi, Y.-A. Breton, R. B. Solomon, K. Conover, P. Shizgal, and P. Dayan. Optimal indolence: a normative microscopic approach to work and leisure. *Journal of The Royal Society Interface*, 11(91):20130969–20130969, Nov. 2013. ISSN 1742-5689. doi: 10.1098/rsif.2013.0969. URL `http://www.ncbi.nlm.nih.gov/pubmed/24284898`. (page 105)

J. Olds. Satiation effects in self-stimulation of the brain. *Journal of comparative and physiological psychology*, 51(6):675–8, Dec. 1958. ISSN 0021-9940. URL `http://www.ncbi.nlm.nih.gov/pubmed/13620802`. (page 68)

J. Olds and P. Milner. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *J Comp and Phys Psych*, 47(6): 419–27, Dec. 1954. ISSN 0021-9940. (pages 42, 68, and 72)

J. Olds and M. E. Olds. Positive reinforcement produced by stimulating hypothalamus with iproniazid and other compounds. *Science (New York, N.Y.)*, 127 (3307):1175–6, May 1958. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/13555860. (page 68)

E. Peterson. The temporal pattern of mosquito flight activity. *Behaviour*, 1980. URL http://www.jstor.org/stable/4534012. (page 64)

P. E. M. Phillips, M. E. Walton, and T. C. Jhou. Calculating utility: preclinical evidence for cost-benefit analysis by mesolimbic dopamine. *Psychopharmacology*, 191(3):483–95, Apr. 2007. ISSN 0033-3158. doi: 10.1007/s00213-006-0626-6. URL http://www.ncbi.nlm.nih.gov/pubmed/17119929. (page 152)

S. S. Pliskoff and J. G. Fetterman. Undermatching and overmatching: The fixed-ratio changeover requirement. *Journal of the experimental analysis of behavior*, 36(1):21–7, July 1981. ISSN 0022-5002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1333049&tool=pmcentrez&rendertype=abstract. (page 59)

C. Prévost, M. Pessiglione, E. Météreau, M.-L. Cléry-Melin, and J.-C. Dreher. Separate valuation subsystems for delay and effort decision costs. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(42): 14080–90, Oct. 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2752-10.2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20962229. (pages 152 and 153)

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics)*. Wiley-Blackwell, 2005. ISBN 0471727822. (pages 41, 65, 66, 67, 75, 76, 81, 99, 100, 105, and 108)

F. Putters, J. Metz, and S. Kooijman. The identification of a simple function of a Markov chain in a behavioural context: Barbs do it (almost) randomly. *Nieuw Archief voor Wiskunde*, 1984. URL http://scholar.google.co.uk/scholar?q=putters+et+al+1984+barbs&btnG=&hl=en&as_sdt=0%2C5#0. (page 64)

P. H. Rudebeck, M. E. Walton, A. N. Smyth, D. M. Bannerman, and M. F. S. Rushworth. Separate neural pathways process different decision costs. *Nature neuroscience*, 9(9):1161–8, Sept. 2006. ISSN 1097-6256. doi: 10.1038/nn1756. URL http://www.ncbi.nlm.nih.gov/pubmed/16921368. (pages 152 and 153)

R. B. Rutledge, M. Dean, A. Caplin, and P. W. Glimcher. Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30 (40):13525–36, Oct. 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1747-10.2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2957369&tool=pmcentrez&rendertype=abstract. (page 107)

J. D. Salamone and M. Correa. Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behav Brain Res*, 137(1-2):3–25, Dec. 2002. ISSN 0166-4328. (pages 40, 65, 98, 122, 152, 192, and 193)

J. D. Salamone, M. S. Cousins, and S. Bucher. Anhedonia or anergia? Effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a T-maze cost/benefit procedure. *Behavioural brain research*, 65(2):221–9, Dec. 1994. ISSN 0166-4328. URL http://www.ncbi.nlm.nih.gov/pubmed/7718155. (page 152)

J. D. Salamone, M. Correa, A. Farrar, and S. M. Mingote. Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology*, 191(3):461–82, Apr. 2007. ISSN 0033-3158. doi: 10.1007/s00213-006-0668-9. URL http://www.ncbi.nlm.nih.gov/pubmed/17225164. (page 152)

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–9, Mar. 1997. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9054347. (pages 40 and 69)

M. Shidara, T. G. Aigner, and B. J. Richmond. Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J Neurosci*, 18(7):2613–25, Apr. 1998. ISSN 0270-6474. (page 78)

P. Shizgal and G. Mathews. Electrical stimulation of the rat diencephalon: differential effects of interrupted stimulation on on- and off-responding. *Brain research*, 129(2):319–33, July 1977. ISSN 0006-8993. URL http://www.ncbi.nlm.nih.gov/pubmed/884507. (page 70)

P. Shizgal, C. Bielajew, D. Corbett, R. Skelton, and J. Yeomans. Behavioral methods for inferring anatomical linkage between rewarding brain stimulation sites. *Journal of comparative and physiological psychology*, 94(2):227–37, Apr. 1980. ISSN 0021-9940. URL http://www.ncbi.nlm.nih.gov/pubmed/6965946. (page 69)

R. L. Shull and S. S. Pliskoff. Changeover delay and concurrent schedules: some effects on relative performance measures. *Journal of the experimental analysis of behavior*, 10(6):517–27, Nov. 1967. ISSN 0022-5002. doi: 10.1901/jeab.1967.10-517. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338422&tool=pmcentrez&rendertype=abstract. (page 59)

R. L. Shull, S. T. Gaynor, and J. A. Grimes. Response rate viewed as engagement bouts: effects of relative reinforcement and schedule type. *J Exp Anal Behav*, 75(3):247–74, May 2001. ISSN 0022-5002. doi: 10.1901/jeab.2001.75-247. (pages 39, 62, and 190)

J. M. Simmons and C. R. Gallistel. Saturation of subjective reward magnitude as a function of current and pulse frequency. *Behav Neurosci*, 108(1):151–60, Feb. 1994. ISSN 0735-7044. (pages 60 and 70)

S. Singh. Soft dynamic programming algorithms: Convergence proofs. *Proceedings of Workshop on Computational Learning*, 1993. (pages 67 and 100)

B. F. Skinner. *The behavior of organisms: an experimental analysis.* Appleton-Century-Crofts, New York, 1938. (page 40)

B. F. Skinner. Selection by consequences. *Science (New York, N.Y.)*, 213(4507):501–4, July 1981. ISSN 0036-8075. (page 40)

R. B. Solomon, K. Conover, and P. Shizgal. Measurement of subjective opportunity costs in rats working for rewarding brain stimulation. in preparation. (page 73)

B. Sonnenschein, K. Conover, and P. Shizgal. Growth of brain stimulation reward as a function of duration and stimulation strength. *Behav Neurosci*, 117(5):978–94, Oct. 2003. ISSN 0735-7044. doi: 10.1037/0735-7044.117.5.978. (pages 60 and 70)

P. L. Soto, J. J. McDowell, and J. Dallery. Effects of adding a second reinforcement alternative: implications for Herrnstein's interpretation of r(e). *Journal of the experimental analysis of behavior*, 84(2):185–225, Sept. 2005. ISSN 0022-5002. doi: 10.1901/jeab.2005.09-05. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1243979&tool=pmcentrez&rendertype=abstract. (page 59)

P. L. Soto, J. J. McDowell, and J. Dallery. Feedback functions, optimization, and the relation of response rate to reinforcer rate. *Journal of the experimental analysis of behavior*, 85(1):57–71, Jan. 2006. ISSN 0022-5002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1397792&tool=pmcentrez&rendertype=abstract. (page 59)

D. Stephens and J. Krebs. *Foraging theory: monographs in behavior and ecology.* Princeton University Press, Princeton, NJ, 1986. URL http://www.lavoisier.fr/livre/notice.asp?ouvrage=1491998. (page 108)

J. R. Stevens, A. G. Rosati, K. R. Ross, and M. D. Hauser. Will travel for food: spatial discounting in two new world monkeys. *Current biology : CB*, 15(20): 1855–60, Oct. 2005. ISSN 0960-9822. doi: 10.1016/j.cub.2005.09.016. URL http://www.ncbi.nlm.nih.gov/pubmed/16243033. (page 152)

L. P. Sugrue, G. S. Corrado, and W. T. Newsome. Matching behavior and the representation of value in the parietal cortex. *Science (New York, N.Y.)*, 304 (5678):1782–7, June 2004. ISSN 1095-9203. doi: 10.1126/science.1094765. URL http://www.ncbi.nlm.nih.gov/pubmed/15205529. (page 59)

R. Sutton and A. Barto. *Reinforcement learning: An introduction*, volume 28. Cambridge University Press, 1998. (pages 41, 53, 67, 75, 81, 100, and 105)

R. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999. (page 97)

Y. Takikawa, R. Kawagoe, H. Itoh, H. Nakahara, and O. Hikosaka. Modulation of saccadic eye movements by predicted reward outcome. *Exp Brain Res*, 142(2): 284–91, Jan. 2002. ISSN 0014-4819. doi: 10.1007/s00221-001-0928-1. (page 78)

I. Trujillo-Pisanty, G. Hernandez, I. Moreau-Debord, M.-P. Cossette, K. Conover, J. F. Cheer, and P. Shizgal. Cannabinoid receptor blockade reduces the opportunity cost at which rats maintain operant performance for rewarding brain stimulation. *J Neurosci*, 31(14):5426–35, Apr. 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.0079-11.2011. (pages 18, 40, 62, 63, 70, 71, 98, 122, and 192)

J. G. Veening, L. W. Swanson, W. M. Cowan, R. Nieuwenhuys, and L. M. Geeraedts. The medial forebrain bundle of the rat. II. An autoradiographic study of the topography of the major descending and ascending components. *The Journal of comparative neurology*, 206(1):82–108, Mar. 1982. ISSN 0021-9967. doi: 10.1002/cne.902060107. URL `http://www.ncbi.nlm.nih.gov/pubmed/6980232`. (page 68)

K. Vohs and R. Baumeister. Making choices impairs subsequent self-control: a limited-resource account of decision making, self-regulation, and active initiative. *Journal of personality . . .* , 2008. URL `http://psycnet.apa.org/journals/psp/94/5/883/`. (page 129)

P. Waelti, A. Dickinson, and W. Schultz. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412(6842):43–8, July 2001. ISSN 0028-0836. doi: 10.1038/35083500. URL `http://www.ncbi.nlm.nih.gov/pubmed/11452299`. (page 40)

M. E. Walton, D. M. Bannerman, and M. F. S. Rushworth. The role of rat medial frontal cortex in effort-based decision making. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(24):10996–1003, Dec. 2002. ISSN 1529-2401. URL `http://www.ncbi.nlm.nih.gov/pubmed/12486195`. (pages 152 and 153)

M. E. Walton, D. M. Bannerman, K. Alterescu, and M. F. S. Rushworth. Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(16):6475–9, July 2003. ISSN 1529-2401. URL `http://www.ncbi.nlm.nih.gov/pubmed/12878688`. (pages 152 and 153)

M. E. Walton, P. L. Croxson, M. F. S. Rushworth, and D. M. Bannerman. The mesocortical dopamine projection to anterior cingulate cortex plays no role in guiding effort-related decisions. *Behavioral neuroscience*, 119(1):323–8, Feb. 2005. ISSN 0735-7044. doi: 10.1037/0735-7044.119.1.323. URL `http://www.ncbi.nlm.nih.gov/pubmed/15727537`. (page 152)

M. E. Walton, S. W. Kennerley, D. M. Bannerman, P. E. M. Phillips, and M. F. S. Rushworth. Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural networks : the official journal of the International Neural Network Society*, 19(8):1302–14, Oct. 2006. ISSN 0893-6080. doi: 10.1016/j.neunet.2006.03.005. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2519033&tool=pmcentrez&rendertype=abstract`. (page 152)

M. E. Walton, P. H. Rudebeck, D. M. Bannerman, and M. F. S. Rushworth. Calculating the cost of acting in frontal cortex. *Annals of the New York Academy of Sciences*, 1104:340–56, May 2007. ISSN 0077-8923. doi: 10.1196/annals.1390.009. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2519032&tool=pmcentrez&rendertype=abstract`. (page 152)

M. E. Walton, J. Groves, K. A. Jennings, P. L. Croxson, T. Sharp, M. F. S. Rushworth, and D. M. Bannerman. Comparing the role of the anterior cingulate cortex and 6-hydroxydopamine nucleus accumbens lesions on operant effort-based decision making. *The European journal of neuroscience*, 29(8):1678–91, Apr. 2009. ISSN 1460-9568. doi: 10.1111/j.1460-9568.2009.06726.x. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2954046&tool=pmcentrez&rendertype=abstract`. (pages 152 and 153)

M. J. Wanat, C. M. Kuhnen, and P. E. M. Phillips. Delays conferred by escalating costs modulate dopamine release to rewards but not their predictors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(36):12020–7, Sept. 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2691-10.2010. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2946195&tool=pmcentrez&rendertype=abstract`. (pages 152 and 153)

M. A. Waraczynski. The central extended amygdala network as a proposed circuit underlying reward valuation. *Neuroscience and biobehavioral reviews*, 30(4):472–96, Jan. 2006. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2005.09.001. URL `http://www.ncbi.nlm.nih.gov/pubmed/16243397`. (page 71)

J. H. Wearden and I. S. Burgess. Matching since Baum (1979). *Journal of the experimental analysis of behavior*, 38(3):339–48, Nov. 1982. ISSN 0022-5002. doi: 10.1901/jeab.1982.38-339. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347873&tool=pmcentrez&rendertype=abstract`. (page 59)

R. H. Wiley and S. Hartnett. Mechanisms of spacing in groups of juncos: measurement of behavioural tendencies in social situations. *Animal Behaviour*, 1980. URL http://www.sciencedirect.com/science/article/pii/S0003347280800893. (page 64)

J. Williams, G. Sagvolden, E. Taylor, and T. Sagvolden. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 2. Control by novelty. *Behav Brain Res*, 198(2):283–90, Mar. 2009a. ISSN 1872-7549. doi: 10.1016/j.bbr.2008.08.045. (pages 39, 62, and 190)

J. Williams, G. Sagvolden, E. Taylor, and T. Sagvolden. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 1. Control by place, timing, and reinforcement rate. *Behav Brain Res*, 198(2):273–82, Mar. 2009b. ISSN 1872-7549. doi: 10.1016/j.bbr.2008.08.044. (pages 39, 62, and 190)

J. Williams, G. Sagvolden, E. Taylor, and T. Sagvolden. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 3. Control by reinforcer rate changes and predictability. *Behav Brain Res*, 198(2):291–7, Mar. 2009c. ISSN 1872-7549. doi: 10.1016/j.bbr.2008.08.046. (pages 39, 62, and 190)

R. Wise. Neuroleptics and operant behavior: the anhedonia hypothesis. *Behavioral and brain sciences*, 1982. URL http://journals.cambridge.org/abstract_S0140525X00010372. (page 69)

J. S. Yeomans. The absolute refractory periods of self-stimulation neurons. *Physiology & behavior*, 22(5):911–9, May 1979. ISSN 0031-9384. URL http://www.ncbi.nlm.nih.gov/pubmed/315569. (page 69)

J. S. Yeomans and J. K. Davis. Behavioral measurement of the post-stimulation excitability of neurons mediating self-stimulation by varying the voltage of paired pulses. *Behavioral biology*, 15(4):435–47, Dec. 1975. ISSN 0091-6773. URL http://www.ncbi.nlm.nih.gov/pubmed/1212153. (page 69)